Società di Bioinformatica Italiana
Bioinformatics Italian Society

BITS annual meeting 2014
February 26-28, Rome, Italy

# *ABSTRACT BOOK*

*Dear participant,*

*The Bioinformatics Italian Society, together with the Department of Biology of the University of Rome Tor Vergata, the Istituto Superiore di Sanità and the Department of Physics of Sapienza University of Rome welcomes you to the Eleventh Annual Conference of BITS.*

*The BITS Annual meeting, now at its eleventh edition, is the major event organized by the Italian research community in Bioinformatics.*

*We would like to warmly thank all the many people who worked hard for making this conference possible.*

*We hope you will enjoy the keynote talks, the scientific program, the poster displays, the exhibitions and the social events.*

*Have a nice, fruitful and enjoyable meeting,*

*Manuela, Anna and Elisabetta*

# Index

# Useful information

## Meeting location

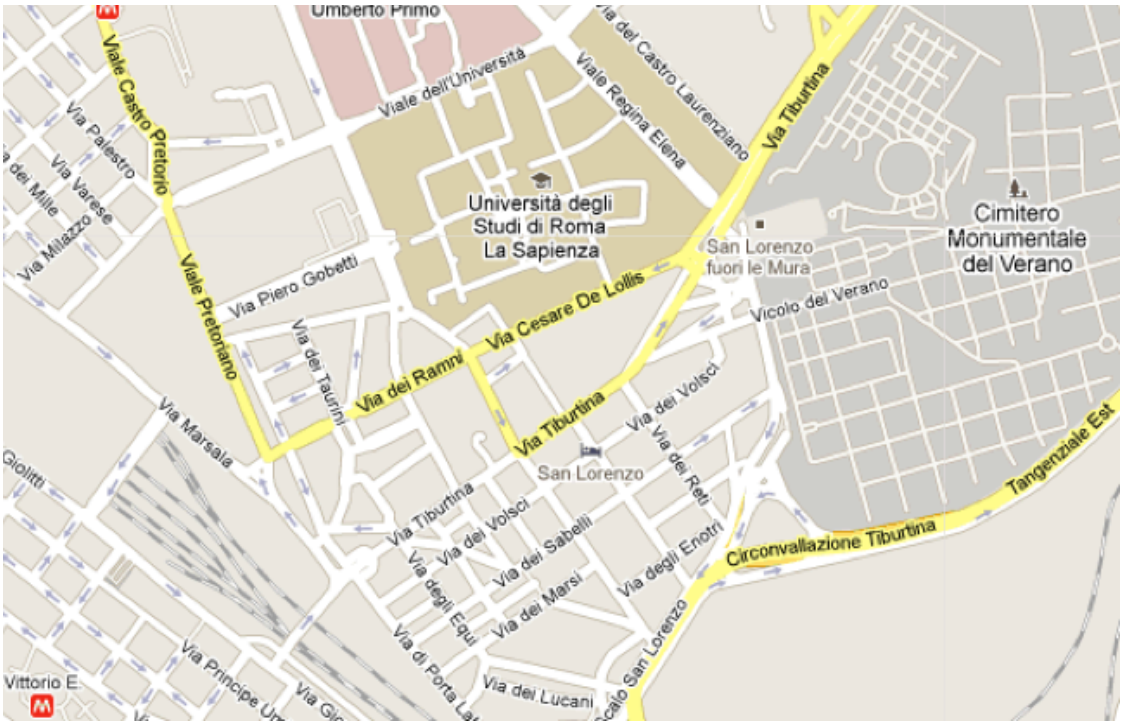
The Amaldi room is located in the Department of Physics "Marconi" on the first floor.
The main entrance of the University is in P.le Aldo Moro, 5.
The Physics Department is the second at your left
You can also take the University secondary entrance. In that case, follow the road until you get to the main University building (Rettorato). After you pass the Rettorato, the Physics Department is in front of you.

The University is located near San Lorenzo, It occupies roughly the two sides of the early stretch of Via Tiburtina starting from Termini railway station and ending at the Verano area. The latter includes the ancient basilica of San Lorenzo fuori le Mura, which the district takes its name from.
Originally a working-class neighborhood (its inhabitants were mostly workers of the Peroni Brewery and the freight yard), it has been a popular area.

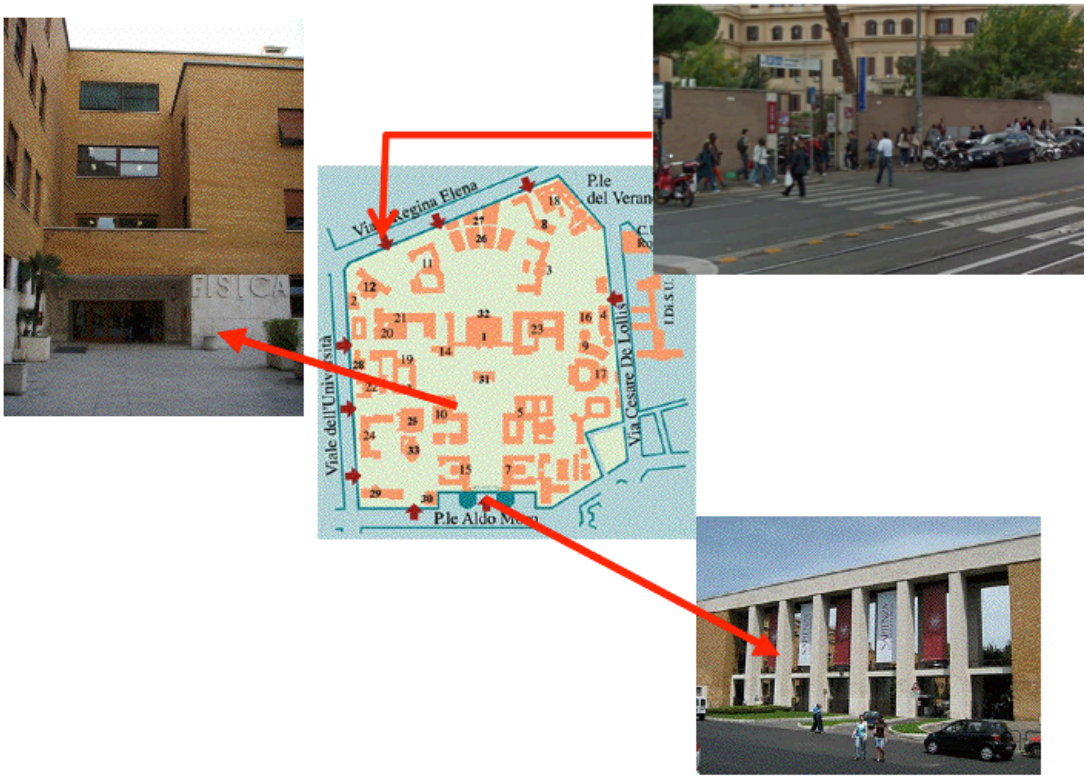
Today San Lorenzo is increasingly assuming the character of a student and young artist district. Its rustic environment is mostly trafficked by university students, artists, and creative types inspired by long nights and lounge-worthy mornings.

Every evening, hundreds of young people descend on the area. Pizzerias, trattorie, boutiques and other modern places are subsequently replacing the old popular workshops and small markets.

## Area around the University:




## The entrance to the University and to the Physics Department:

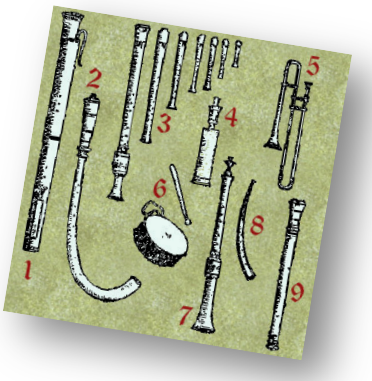


## Social events: Concert "Amor sacro, amor profano e ..... bestiario"

On February the 27th, in aula Amaldi after the regular program, there will be a concert "Sacred and profane love and ... bestiary" by the **Gruppo da Camera of the Schola Cantorum ARAMUS**, directed by **M° Osvaldo Guidotti**, and the **Ludus Zephyri** Group, playing Renaissence musical instruments.

The **Schola Cantorum** Santa Maria degli Angeli is the official choir of the Basilica with the same name. The choir is directed by Maestro Osvaldo Guidotti, official organist of the same Basilica, since 1989.
The choir has an intense concert activity in Italy and Europe. The group has executed as an absolute first performance works by Nicola Bellandi, Flavio Di Silvio, Osvaldo Guidotti, Alberto Meoli, Ennio Morricone and Vittorio Furgeri. Ennio Morricone has composed for the choir and has directed, in its first execution, the *Ave Regina Coelorum* for four voices, orchestra and pipe organ. The Choir also recorded several CDs, among which Mozart Requiem (1999), the Johannes Passion by JS Bach (2001), Petite Messe Solennelle by Rossini (2005), Messa in B minor by JS Bach (2011) and Stabat Mater bu O Guidotti (2011).
La Schola Cantorum S.Maria degli Angeli participated to many concert fairs both a cappella and with prestigious orchestras. It sang with artists such as José Carreras, Montserrat Caballé, Luciana Serra, Cecilia Gasdia, José Cura, Pietro Ballo, William Matteuzzi, Ines Salazar, and many more.


Maestro **Osvaldo Guidotti**, organist, director, composer was educated at the schools of Luigi Celeghin, Ivan Vandor, Giacomo Baroffio and Nicola Hansalik Samale. He graduated in Organ and organ composition, Choral music and choir direction at the Conservatory. He represented Italy at the International organ Festival "Domkirche St. Marien – Hamburg" and at the Festival *Organos Historicos del la Region de Mursia.* He is President and art director of ARAMUS, Associazione Romana Arte Musica, of the International organ Festival *"Romae Organum Monumentale*" , of the *"Organo in Festa"* fair and many other initiatives.



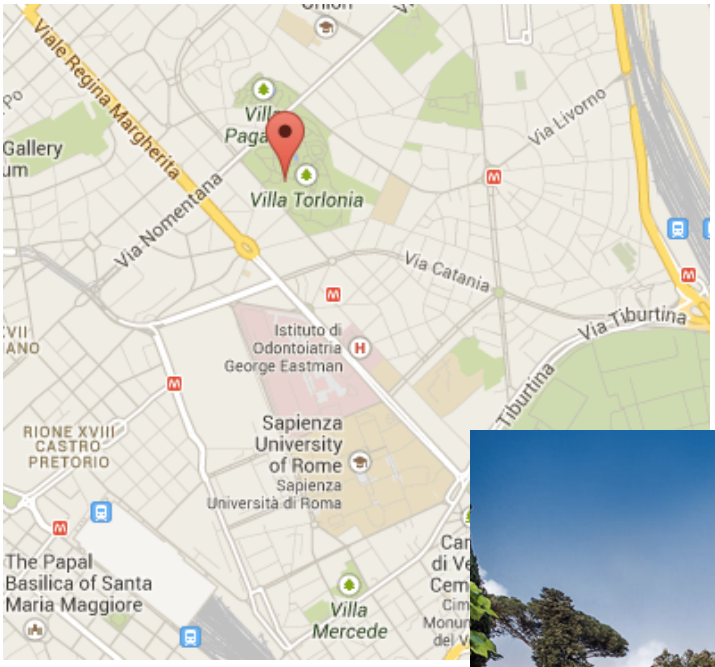

The **Ludus Zephyri** group is an ensemble of musicians, playing ancient instruments. For the BITS2014 Annual meeting, Sabine Cassola, Stefania Grillo, Eric Cassola, Dario Salerno and Ulrike Voss will play bowed string and wind renaissance instruments.

## Social events: the social dinner, co-sponsored by NOVARTIS

The social dinner will be at the Limonaia, within walking distance from the University. Please remember to take your ticket with you.

The restaurant is located in a splendid lodge with exposed beams and windows looking at the garden.

*The building was commissioned in the 9th century from Alessandro Torlonia to the architect Quintiliano Raimondi, it was meant to host the citrus trees during the winter season.*




## Transportation

**From Termini Main station**
Both the Department and the hotel are about 1.5 Km from the main station. You can either take a taxi or walk. Please only use official taxis.

**From Fiumicino Airport**
Take the fast train Leonardo Express to the main train station (Termini). It leaves every 30 mins (at :07 and :37) starting from 6:37 until 23:37. The ticket costs 14.00 Euro.
You can also take a taxi. During the day (depending on traffic) the cost is about 60-65 Euro. Please only use official taxis.

**From Ciampino Airport**
There is a bus service to Termini. It is a dedicated coach for RyanAir, Air Berlin and some other airlines. Tickets cost 8 Eur one way and Eur 12.50 return. The bus leaves in connection with incoming flights and the ticket boot is in front of the Arrival gate. Depending on traffic, it might take between 20 mins and 40 mins.
You can also take a taxi. During the day the cost should be around 40-50 Euro again depending on traffic. Please only use official taxis.

## BITS Steering Committee

- **Manuela HELMER-CITTERICH** (University of Rome "Tor Vergata") President
- **Francesca CICCARELLI** (European Institute of Oncology, Milan)
- **Angelo FACCHIANO** (CNR Istituto di Scienze dell'Alimentazione, Avellino) Secretary
- **Carmela GISSI** (University of Milan)
- **Sabino LIUNI** (CNR Institute of Biomedical Technologies, Bari) Treasurer
- **Roberto MARANGONI** (University of Pisa)
- **Paolo ROMANO** (IRCCS San Martino IST, Genoa)

## Organizers

- **Manuela Helmer-Citterich** (University of Rome "Tor Vergata")
- **Elisabetta Pizzi** (Istituto Superiore di Sanità, Rome, Italy)
- **Anna Tramontano** (Sapienza University of Rome)

## Program Committee

- **Claudia Angelini** (CNR - IAC, Napoli, Italy)
- **Marco Antoniotti** (University of Milano Bicocca, Italy)
- **Marcella Attimonelli** (University of Bari, Italy)
- **Gabriele Ausiello** (University of Rome "Tor Vergata", Italy)
- **Riccardo Bellazzi** (University of Pavia, Italy)
- **Roberta Bosotti** (Nerviano Medical Sciences, Nerviano (MI), Italy)
- **Raffaele Calogero** (University of Torino, Italy)
- **Nicola Cannata** (Next Generation Bioinformatics srl)
- **Rita Casadio** (University of Bologna, Italy)
- **Maria Luisa Chiusano** (University of Naples Federico II, Napoli, Italy)
- **Francesca Ciccarelli** (King's College London, UK)
- **Francesca Cordero** (University of Torino, Italy)
- **Susan Costantini** (INT Pascale)
- **Domenica D'Elia** (CNR, Institute for Biomedical Technologies, Bari, Italy)
- **Angelo Facchiano** (CNR, Institute of Food Science, Avellino, Italy)
- **Alfredo Ferro** (University of Catania, Italy)
- **Federico Fogolari** (University of Udine, Italy)
- **Carmela Gissi** (University of Milan, Italy)
- **Rosalba Giugno** (University of Catania, Italy)
- **Giorgio Grillo** (CNR - Institute for Biomedical Technologies, Bari, Italy)
- **Alessandro Guffanti** (Genomnia srl, Milano, Italy)
- **Manuela Helmer-Citterich** (University of Rome "Tor Vergata", Italy)
- **Giovanni Lavorgna** (Ospedale San Raffaele, Milan, Italy)
- **Vito Flavio Licciulli** (CNR - Institute for Biomedical Technologies, Bari, Italy)
- **Sabino Liuni** (CNR - Institute for Biomedical Technologies, Bari, Italy)

- **Alberto Magi** (University of Firenze, Italy)
- **Paolo Magni** (University of Pavia, Italy)
- **Anna Marabotti** (University of Salerno, Italy)
- **Roberto Marangoni** (University of Pisa, Italy)
- **Marco Masseroli** (Politecnico di Milano, Italy)
- **Giancarlo Mauri** (University of Milano Bicocca, Italy)
- **Cristian Micheletti** (SISSA, Scuola Internazionale Superiore di Studi Avanzati, Trieste, Italy)
- **Veronica Morea** (CNR - Institute of Molecular Biology and Pathology, Rome, Italy)
- **Marco Muselli** (CNR)
- **Alessandro Pandini** (King's College London, UK)
- **Stefano Pascarella** (Sapienza University of Rome, Italy)
- **Marco Pellegrini** (CNR - Istituto di Informatica e Telematica, Pisa, Italy)
- **Graziano Pesole** (CNR - Institute of Biomembranes and Bioenergetics, Bari, Italy)
- **Ernesto Picardi** (University of Bari, Italy)
- **Elisabetta Pizzi** (Istituto Superiore di Sanità, Rome, Italy)
- **Alberto Policriti** (University of Udine, Italy)
- **Uberto Pozzoli** (IRCCS 'E.Medea', Bosisio Parini (LC), Italy)
- **Alfredo Pulvirenti** (University of Catania, Italy)
- **Paolo Romano** (IRCCS San Martino IST, Genoa, Italy)
- **Remo Sanges** (Stazione Zoologica "Anton Dorhn", Napoli, Italy)
- **Roberto Tagliaferri** (University of Salerno, Italy)
- **Silvio C.E. Tosatto** (University of Padova, Italy)
- **Anna Tramontano** (Sapienza University of Rome, Italy)
- **Luigi Varesio** (Giannina Gaslini Institute, Genoa , Italy)
- **Saverio Vicario** (CNR - Institute for Biomedical Technologies, Bari, Italy)
- **Nicola Vitulo** (University of Padova, Italy)
- **Stefano Volinia** (University of Ferrara, Italy)
- **Andreas Zanzoni** (TAGC U1090, Inserm, Marseille, France)

## Local organizing committee

- **Salvatore Cirillo** (University of Rome Torvergata)
- **Daniel D'Andrea** (Sapienza University of Rome)
- **Alessio Colantoni** (University of Rome Torvergata)
- **Rosalba Lepore** (Sapienza University of Rome)
- **Pier Paolo Olimpieri** (Sapienza University of Rome)
- **Valentina Pica** (Sapienza University of Rome)
- **Marco Pietrosanto** (University of Rome Torvergata)
- **Domenico Raimondo** (Sapienza University of Rome)
- **Gabriella Sferra** (Istituto Superiore di Sanità, Rome, Italy)
- **Allegra Via** (Sapienza University of Rome)

## Our sponsors

We would like to warmly thank our sponsors. Without them the meeting would have been a different one.

# Program

## February 26th

| | | | |
|---|---|---|---|
| *12:00* | | Lunch and welcome | |
| *13:50* | | Set up even numbered posters | |
| *14:00* | | **Ewan Birney** | **PREPARATA LECTURE** |
| *14:50* | Omics 1 | D'Aurizio R | Exploiting whole-exome sequencing data to identify copy number variants |
| *15:10* | Omics 2 | Sanavia T | FunPat: a function-based pattern analysis pipeline for RNA-seq time-series data |
| *15:30* | Omics 3 | D'Elia D | mirNET: a web-based system for the analysis of miRNA:mRNA regulatory network |
| *15:50* | | Coffee and posters | |
| *16:35* | Omics 4 | Martignetti L | Detection of miRNA regulatory effect on triple negative breast cancer transcriptome |
| *16:55* | Omics 5 | Palluzzi F | euNet: a data-driven interactome building tool for very large interaction networks |
| *17:15* | Omics 6 | Ceol A. | From genome to molecular interactions, a plug-in to map next generation sequencing data to molecular interactions and their structures with the Integrated Genome Browser. |
| *17:35* | | BITS General assembly | |

## February 27th

| | | | |
|---|---|---|---|
| *8:30* | | ***Gunnar von Heijne*** | **EMBO LECTURE** |
| *9:20* | Omics 7 | Gissi C | Ascidian mitogenomics: comparison of evolutionary rates in closely related taxa provides evidence of ongoing speciation events |
| *9:40* | Omics 8 | Petrosino G | Retroelements and Long Noncoding RNAs Content Suggests Convergent Molecular Evolution Between Octopus and Mammals. |
| *10:00* | Omics 9 | Vicario S | Mean phylogenetic surprise: a unit of measure for describing changes across biological communities |
| *10:20* | Omics 10 | Palmeri A | KINhibit: Predicting resistance to kinase inhibitors from cancer genomics data |
| *10:40* | | Coffee and posters | |
| *11:25* | | Remove even numbered posters | |
| *11:35* | SSB HS | Stano P | Advancements and open questions in synthetic cells: new challenges for systems biology? |
| *12:05* | SSB 1 | Pasotti L | Bottom-up design of genetic circuits: characterization and re-use of biological building blocks to engineer predictable systems |
| *12:25* | SSB 2 | Calderone A | SIGNOR: a new resource to support experimental investigation of signaling networks. |
| *12:45* | SSB 3 | Graudenzi A | Stochastic differentiation and homeostasis in a multiscale model of intestinal crypt dynamics |

| | | | |
|---|---|---|---|
| *13:05* | SSB 4 | Sferra G. | Co-evolving proteins to infer protein-protein interactions: a re-assessment of the phylogenetic profiling |
| *13:25* | | Lunch | |
| *14:55* | | Setup odd numbered posters | |
| *15:05* | ALG 1 | Pisanti N | A dynamic programming algorithm for Haplotype Assembly of Future-Generation Sequencing Reads |
| *15:25* | ALG 2 | Brasca S | Vector Integration Site Identification through Density-based methods |
| *15:45* | ALG 3 | Prezza N | Fast randomized approximate string matching with succinct hash data structures |
| *16:05* | | Coffee and posters | |
| *16:50* | ALG 4 | Parodi S | Differential diagnosis of pleural mesothelioma using Logic Learning Machine |
| *17:10* | ALG 5 | Pellegrini M | Protein families comparisons using repeatome-based profiling |
| *17:30* | ALG 6 | Limongelli I | PaPI: the Pseudo Amino acid variant Predictor |
| *17:50* | National initiatives | Pino Macino | EPIGEN |
| *18:00* | National initiatives | Luciano Milanesi | INTEROMICS |
| *18:10* | National initiatives | Graziano Pesole | ELIXIR-ITA |
| *18:20* | | Concert | Amor sacro, amor profano e.... bestiario |
| | | Social dinner | co-sponsored by **Novartis** |
| | | **February 28th** | |
| *8:30* | | ***Franca Fraternali*** | **EPIGEN LECTURE** |
| *9:20* | STR 1 | Giollo M | Comprehensive large scale assessment of intrinsic protein disorder predictors |
| *9:40* | STR 2 | Punta M | An estimated 5% of new protein structures solved today represent a new Pfam family |
| *10:00* | STR 3 | Olimpieri PP | Tabhu: Tools for AntiBody Humanization. |
| *10:20* | STR 4 | Mattei E | A new alphabet and substitution matrix to represent and compare RNA secondary structure |
| *10:40* | STR 5 | Di Palma F | Structural and thermodynamics characterization of the ligand-aptamer interactions in the adenine riboswitch |
| *11:00* | | Coffee and posters | |
| *11:45* | **STO HS** | Fanelli D | Spatial and temporal order beyond the deterministic limit: the role of stochastic fluctuations in polulation |
| *12:15* | STO 1 | Riba A | k.pA combination of transcriptional and microRNA regulation improves the stability of the relative concentrations of target genes |
| *12:35* | STO 2 | Valleriani A | Single-molecule modeling of mRNA decay |
| *12:55* | STO 3 | Zambrano S | High-Throughput Quantification and Stochastic Modeling of NF- κB Dynamics. |
| *13:15* | STO 4 | Caravagna G | Enhancing simulation of chemical reactions at mesoscales |

## Poster instructions

Posters will be displayed in two rounds.

**Even numbered** posters must be set up upon arrival or at the coffee break of February **26th** at the latest, they MUST be taken out at 11:25 of the **27th**. There is a time slot in the program to do so

**Odd numbered** posters must be set up at 14:55 of the **27$^{th}$**. There is time in the program to do so. They MUST be taken out after the coffee break of the **28th**.

The program reports the session and number of your poster. Please locate the panel with the appropriate number. There will be two labels on the boards, one for the even and one for the odd numbered posters.

## Keynote speaker: Ewan Birney – Preparata Lecture

Dr. Ewan Birney originally trained in biochemistry but moved quickly into bioinformatics, publishing his first set of programs (Pairwise and Searchwise) as an undergraduate at Balliol College, Oxford. He did his PhD with Richard Durbin at the Wellcome Trust Sanger Institute and St John's College Cambridge. In 2000 Ihejoined EMBL-EBI as a Team Leader, became a Senior Scientist in EMBL in 2003 and was appointed Associate Director in 2012.

He is one of the founders of the Ensembl genome browser and other databases, and has played a key role in many large-scale genomics projects, notably the sequencing of the Human Genome in 2000 and the analysis of genome function in the ENCODE project. He has been Lead Analysis Coordinator for ENCODE since 2007; he also coordinated data analysis in the "1% Pilot". Dr Birney has played a vital role in annotating the genome sequences of the human, mouse, chicken and several other organisms; this work has had a profound impact on our understanding of genomic biology. His research group currently focuses on genomic algorithms and inter-individual differences in human and other species.
As Associate Director of EMBL-EBI, Dr Birney shares strategic oversight of EBI services with Rolf Apweiler (co-Associate Director). As well as these strategic aspects, Ewan Birney still does research, working on aspects such as DNA Compression, Functional genomics analysis (eg, a recent paper with the Furlong lab) and using inter individual differences to understand basic biology (eg, a paper of CTCF binding in two families)

He was awarded the Francis Crick Award from the Royal Society, the 2005 Overton Prize from the International Society for Computational Biology, the 2005 Benjamin Franklin Award for contributions in Open Source Bioinformatics and, in 2012, he was elected fellow of the European Molecular Biology Organisation (EMBO).

## Keynote speaker: Gunnar von Heijne – EMBO Lecture

Prof. von Heijne holds a PhD in Theoretical Physics and is presently Professor of Theoretical Chemistry at the Stockholm University, Director of the Center for Biomembrane Research, 2006 and Vice Director, Science for Life Laboratory Stockholm.

He has been awarded a number of prestigious prizes among which The T. Svedberg Award, The Swedish Biochemical Society, 1990 the Göran Gustafsson Prize, the Swedish Academy of Sciences, 1995, The Arrhenius Medal, The Swedish Chemical Society, 1997, the Björkén Prize, Uppsala University, 1998, the van Deenen Medal, Utrecht University, 2009, the Senior Scientist Award of the International Society for Computational Biology, 2012.
He is an elected member of the Royal Swedish Academy of Engineering Sciences, of EMBO, of the Royal Swedish Academy of Sciences, 1997 and of the Academia Europaea, 1998.

He has worked mainly on problems related to protein sorting and membrane protein biogenesis and structure. The work includes both bioinformatics methods development (e.g. methods for prediction of signal peptides and other sorting signals as well as prediction of membrane protein topology) and experimental studies in *E. coli* and eukaryotic systems. Among his important achievements, there is the discovery and experimental validation of the so-called "(-1,-3)-rule" (describes signal peptide cleavage sites) and the "positive inside" rule (describes membrane protein topology), the development of widely used prediction methods (e.g., TopPred, SignalP, TargetP, TMHMM), the first proteome-wide theoretical and experimental studies of membrane protein topology in *E. coli* and *S. cerevisiae*, the first quantitative analysis of the energetics of membrane protein assembly *in vivo*, and theoretical and experimental studies of so-called dual-topology membrane proteins and their role in the evolution of membrane protein structure.

## Keynote speaker: Franca Fraternali – EPIGEN Lecture

Prof. Franca Fraternali received her PhD from the University of Naples in the Department of Physical Chemistry under the supervision of Prof. Vincenzo Barone. She spent the last year of her PhD at the ETH in Zurich co-supervised by Prof. Wilfred van Gunsteren, and received an EMBO fellowship for a post-doctoral position to work in the same group. After post-doctoral work at the University of Strasbourg, and at the EMBL Heidelberg, she obtained in 1999 a position as permanent staff scientist at the MRC National Institute for Medical Research in the Mathematical Biology division directed by Prof. Willie Taylor. In 2005 she was appointed Lecturer in Bioinformatics at King's College London, and subsequently promoted to Reader (2009). She is presently Professor of Bioinformatics and Computational Biology (2013) in the Randall Division of Cell and Molecular Biophysics at King's College in London.

Her research interests are in the study of the physical nature of the interactions between protein-protein, protein-solvent, protein-lipid and protein-nucleic acid. Her group develops and applies bioinformatics methods to analyze available data on such interactions and molecular simulations methods to characterize and determine their stability.
In the recent years her group has focused on protein-protein interaction networks, their characterization in terms of 3D structures and conserved domains, and the analysis of the relative complexes interfaces. With the availability of large-scale data on genetic variations in health and disease the group is focusing in mapping this information to protein-protein and protein-nucleic acids interactions to uncover possible molecular mechanisms underlying genetic and somatic diseases.

## Oral presentations

Oral presentation: Omics 1

**Exploiting whole-exome sequencing data to identify copy number variants**

D'Aurizio R(1), Tattini L(2,4), Pippucci T(3), Pellegrini M(1), Magi A(4)

*(1) Laboratory of Integrative Systems Medicine (LISM), Institute of Informatics and Telematics and Institute of Clinical Physiology, National Research Council, Pisa (2) Laboratory of Molecular Genetics, G. Gaslini Institute, Genoa (3) Medical Genetics Unit, Department of Medical and Surgical Sciences, University of Bologna, Bologna (4) Department of Clinical and ExperimentalMedicine, University of Florence, Florence*

**Motivation:** Whole Exome Sequencing (WES) is a cost-effective and extensively used method involving the capture of enriched protein-coding regions by hybridizing genomic DNA to oligonucleotide probes (baits) and then sequencing using high-throughput technology. WES is a standard approach for detecting single nucleotide variants (SNVs) and small insertion-deletion (indel) variants. Recently new tools have been developed to identify major genomic rearrangements like Copy Number Variants. CNVs are structural variants larger than 50bp, demonstrated to be one of the main sources of genomic variation in humans [1000 Genomes Project Consortium, Abecasis, Nature 2010] causing various diseases including cancer and cardiovascular disease. Few computational tools have been developed to identify CNVs in targeted regions by WES using Read Count (RC) approach [Sathirapongsasuti, Bioinformatics 2011; Krumm, Genome Res 2012; Fromer, AmJHumGenet 2012; Li, Bioinformatics 2011; Magi, Genome Biol 2013]. None of them allows for the identification of CNVs in intergenic regions. We present a new tool able to infer CNVs in both coding and non coding regions from WES data.
**Methods:** Currently, there are several commercial target capture kits for human exome, three of them are the most widely used: NimbleGen's SeqCap , Agilent's SureSelect and Illumina's TruSeq. They share the ability to enrich targeted regions of genomes, mainly Consensus Coding Sequences (CCDS), but differ in probes design, target enrichment efficiency and biases [Clark, Nature Biot 2011; three papers of Sulonen, Parla and Asan from Special Issue of Genome Biology 2011]. We studied the distribution of reads generated by WES approach in coding and non coding regions of the genome. To this end, we used a survey dataset made of 30 WES samples that includes data generated by Clark and by our labs using all three enrichment kits. By exploiting this dataset, we evaluated the properties of Read Count measure (the number of reads that maps to specific region of the genome) in two distinct classes of genomic features: (i) all the CCDS and (ii) non-overlapping genomic windows of different sizes that belong to intergenic regions of the genome. We studied the relationship between RC data and classical genomic systematic biases, such as GC-content and mappability. As a further step, we evaluated the capability of RC data to predict the number of DNA copy of coding and non coding regions. Finally we developed a computational framework for exploiting WES data to identify and predict the genomic regions involved in copy number variation.
**Results:** We found that the three enrichment kits obtain different results in terms of percentage of reads that unambiguously map to out target regions: 38% of reads for TruSeq, 20-50% for SeqCap and 28-50% for SureSelect. As expected, we found that both in-target and out-target RC are affected by traditional biases (GC-content and mappability) and that this biases can be well-mitigated by traditional RC normalization methods [Yoon, Genome

Res 2009; Magi, Bioinformatics 2012]. By using WES data of individuals sequenced by the 1000 Genome project consortium and previously genotyped by Conrand et al. [Nature 2010] we studied the correlation between RC data and the copy number states. We found that RC data are well correlated with CNV state for both coding and non-coding regions, thus demonstrating the capability of our hybrid approach (RC data on coding and non-coding regions) to identify and predict the correct number of DNA copy of the genome. Finally we extended our previously developed pipeline (EXCAVATOR, Magi, Genome Bio 2013 ) to this novel framework. We applied our novel computational pipeline to the analysis of a cancer and a population datasets and we show its capability to recognise exomic as well as genomic regions involved in CNV.

**Contact email:** romina.daurizio@gmail.com

Oral presentation: Omics 2

**FunPat: a function-based pattern analysis pipeline for RNA-seq time-series data**

Sanavia T(1), Finotello F(1), Di Camillo B(1)

*(1) Department of Information Engineering, University of Padova, Padova*

**Motivation:** Next-Generation Sequencing technologies have been extensively applied to quantitative transcriptomics, making RNA-seq a valuable alternative to microarrays for measuring and comparing gene transcription levels. In this context, time-series expression experiments are essential to monitor transient gene response to stimuli. The conventional computational analysis pipeline usually includes: 1) selection of differentially expressed (DE) genes; 2) clustering of expression profiles; 3) functional analysis. This pipeline, however, suffers of some well-known drawbacks. DE genes are usually selected controlling the type-I error rate and correcting for multiple testing at the expense of stringent thresholds with a consequent loss of significant genes. With regard to the second step, the most used data-driven clustering approaches, such as k-means or hierarchical clustering, do not account for technical and biological noise. Moreover, the user has to set the number of clusters either a priori or a posteriori. Finally, the functional analysis is usually performed independently from the previous steps and suffers the redundancy of annotation and lack of specificity.
**Methods:** We introduce FunPat, an R package for time-series RNA-seq data, which searches for groups of DE genes sharing both temporal expression pattern and biological function. FunPat workflow integrates two computational modules, aimed at ranking DE genes and identifying classes of functionally-related genes, respectively. The Gene Ranking module takes expression data as input and calculates, for each gene, the area A of the region bounded by the corresponding time-series expression profile (Di Camillo et al. 2007). The Monte Carlo procedure used to derive the null distribution of A requires only two replicates for a single time point, thus addressing data-poor experimental design, which is currently a common situation in RNA-seq experiments. Genes that pass a desired False Discovery Rate (FDR) threshold are then called seeds, whereas the remaining, or a subset of them, selected by applying a soft-thresholding approach on not-adjusted p-values, are called candidates. Starting from expression data and functional annotations organized according to Gene-Sets, e.g. Gene Ontology (GO) terms (Ashburner et al., 2000), the Temporal Pattern Analysis module recovers among the candidate genes those that are associated to a Gene-Set containing at least one seed gene and share the same temporal expression pattern. Temporal patterns are defined by iteratively applying a linear model-based clustering to genes belonging to the same Gene-Set (Di Camillo et al., 2012). Model parameters are identified using least squares; thus, the procedure accounts for the error measurement, does not require the user to fix the number of clusters and is computationally efficient. If the Gene-Sets are organized according to a hierarchical structure, as in the case of GO, the most specific terms are searched first and genes significantly associated to a pattern are removed from the ancestor nodes. This provides the most specific annotation and prevents redundancy in the annotation. FunPat outputs the temporal patterns associated to each cluster of DE genes annotated to the selected functional Gene-Sets.
**Results:** FunPat was tested on 100 simulated datasets consisting of the expression data of 1000 genes monitored over 13 time points. 120 DE genes were simulated, belonging to 6 different temporal patterns, whereas 880 were random noise with characteristics similar to those observed in RNA-seq data. FunPat ability to select DE genes was assessed in terms of precision and recall. With respect to the selection performed by the Gene Ranking module using a FDR correction equal to 5%, FunPat is able to significantly increase the recall from 0.4

to 0.53 (p-value<1e-17, Wilcoxon signed-rank test) without decreasing the precision ,which is equal to 0.95 in correspondence to a requested false discovery rate of 0.05 (p-value=0.81, Wilcoxon test). The correct identification of the simulated patterns was tested against the hierarchical (HC) and k-means (KC) clustering, fixing the same number of clusters obtained by FunPat. The precision in cluster detection is significantly higher for FunPat (mean 0.67) than for HC (mean 0.5, p-value<1e-17) and comparable to KC (p-value=0.52). However, FunPat outperforms the other methods in terms of recall, equal to 0.88 and significantly greater than both HC (mean recall 0.76) and KC (mean recall 0.66), with p-value<1e-16. FunPat was also applied to a real dataset of time varying B cell vaccine responses in 5 different subjects (Henn et al., 2013). FunPat showed a higher reproducibility of the DE genes detected in different subjects than in the original paper, still confirming a higher similarity between previously vaccinated individuals, in accordance with the results of the study. This indirectly confirms the ability of FunPat to select DE genes with high precision and recall.

**Contact email:** barbara.dicamillo@dei.unipd.it

Oral presentation: Omics 3

**mirNET: a web-based system for the analysis of miRNA:mRNA regulatory networks**

Pio G(1), Ceci M(1), D'Elia D(2), Malerba D(1)

*(1) Department of Computer Science, University of Bari Aldo Moro, Via Orabona 4, 70125, Bari, Italy(2) CNR, Institute for Biomedical Technologies, Via Amendola 122/D, 70126, Bari, Italy*

**Motivation:** Understanding mechanisms and functions of microRNAs (miRNAs) is fundamental for the elucidation of many biological processes and of etiopathology of some diseases, such as tumors and neurodegenerative syndromes. We have developed a new biclustering algorithm, i.e. HOCCLUS2 [1], which is able to significantly correlate multiple miRNAs and their target genes to identify potential miRNA:mRNA regulatory networks. More recently, we developed a new probabilistic classifier [2] working in the semi-supervised ensemble learning setting, which allowed us to apply HOCCLUS2 on large-scale prediction data. In order to allow the researchers to exploit the obtained results, we have started to develop a web-based system, called mirNET, for the efficient query, retrieval, export, visualization and analysis of the discovered regulatory networks.
**Methods:** In [2], we presented a method which learns to combine the score of several prediction algorithms, in order to improve the reliability of the predicted interactions. The approach works in the semi-supervised ensemble learning setting which exploits information conveyed by both labeled (validated interactions, from miRTarBase) and unlabeled (predicted interactions, from mirDIP) instances. The algorithm HOCCLUS2 exploits the large set of produced predictions, with the associated probability, to extract a set of hierarchically organized biclusters. The construction of the hierarchy is performed by an iterative merging, considering both distance and density-based criteria.Extracted biclusters are also ranked on the basis of the p-values obtained by the Student's T-Test which compares intra- and inter-functional similarity of miRNA targets, computed on the basis of the gene classification provided in Gene Ontology (GO). mirNET database relies on PostgreSQL DBMS, while the web-based platform is built through the Play 2.2 Java framework and the Cytoscape library.
**Results:** The mirNET database stores the set of interactions identified in [2] and the biclusters extracted by HOCCLUS2 from such set of interactions, with different parameters. In particular, mirNET stores approximately 5 million predicted interactions between 934 human miRNAs and 30,875 mRNAs, which are exploited in the construction of the hierarchies of biclusters representing potential miRNA regulatory networks. The mirNET web interface allows users to perform extraction and visualization of single interactions (with the score/probability assigned by the learning algorithm) and of biclusters of interest, as well as to easily browse whole biclusters hierarchies. Biclusters hierarchy browsing (i.e., navigation among parents and children biclusters) helps to identify intrinsic hierarchical organization of miRNAs in each specific context. The interface for the analysis of biclusters also provides a graph-based visualization of the predicted miRNA-gene interaction network. The database query system provides a series of filters to facilitate and refine the retrieval of data on the basis of different criteria, such as the biclusters compactness and the p-values computed on the basis of GO hierarchies, that is pBP and pMF. In particular, the compactness measures the (score-) weighted percentage of interactions in the bicluster, normalized by the maximum number of possible interactions, and represents the average strength of the intra-bicluster connections. pBP and pMF values represent the p-values obtained by the Student's T-Test computed on the basis of Biological Process (BP) and Molecular Function (MF) Gene Ontology hierarchies, respectively. mirNET represents an important contribution to the

study of the regulatory role and function of miRNAs. Indeed, as shown in [1] and [2], in addition to the possibility to extract multiple and significant unknown co-targeting of miRNAs, HOCCLUS2 is able to give new clues for the identification of still unknown miRNA functional targeting which could be worth to be experimentally validated. This possibility is due to its ability to associate objects that are apparently not related. This paves the way to the systematic use of mirNET for a comprehensive analysis of all the possible multiple interactions established by miRNAs of interest. Moreover, since mirNET works on computational predictions, it offers the possibility to analyze single interactions and regulatory modules that would be otherwise impossible to reconstruct by considering only experimentally validated interactions, which are strictly dependent on the cell type and experimental conditions used.References:[1] G. Pio, M. Ceci, D. D'Elia, C. Loglisci, D. Malerba, A novel biclustering algorithm for the discovery of meaningful biological correlations between miRNAs and mRNAs, BMC Bioinformatics 14 (Suppl 7), S8, 2013[2] G. Pio, M. Ceci, D. D'Elia, D. Malerba, Integrating microRNA target predictions for the discovery of gene regulatory networks: a semi-supervised ensemble learning approach, BMC Bioinformatics (in press)
**Contact email:** gianvito.pio@uniba.it

Oral presentation: Omics 4

**Detection of miRNA regulatory effect on triple negative breast cancer transcriptome**

Bruno Tesson(1,2,3,4), Thierry Dubois(1,4), Anna Almeida(1,4), Gordon Tucker(5), Emmanuel Barillot(1,2,3), Loredana Martignetti(1,2,3)

*(1) Institut Curie, Paris, France (2) INSERM, U900, Paris, France (3) Mines ParisTech, Fontainebleau, France (4) Department of Translational Research, Paris, France (5) Servier Research Center, Croissy-sur-Seine, France*

**Motivation:** MicroRNAs (miRNAs) are endogenous ∼22 nucleotides RNA molecules discovered as fundamental repressors of gene expression in many biological system. Their role in diverse biological processes, from differentiation and proliferation to apoptosis, strongly support that they could be involved in cancer and indeed their deregulation has been linked to many types of cancer. Identifying key miRNAs contributing to the genesis and development of a particular disease is a challenging task. A major source of information to infer the actual regulatory activity of miRNAs derives from high-throughput experimental data such as transcriptome profiles. The basic assumption is that regulatory activity by miRNAs could be reflected by the expression changes of their target transcripts due to degradation. Existing tools mainly rely on case-control mRNA profile experiments involving strong perturbations such as the knockout/knockdown/overexpression of individual miRNAs. When dealing with less controlled conditions like normal or pathological tissue mRNA profiles, detecting miRNA-mediated target destabilization is more challenging due to inhomogeneous cell type, additional regulatory factors and complex RNA cross-regulation such as ''sponge'' effect.
**Methods:** We introduce here a rank-based method to detect miRNA regulatory activity in cancer derived tissue samples which combines measurements of gene and miRNA expression levels and sequence based target predictions. This approach is designed to detect modest but coordinate changes in the expression of sequence based predicted target genes. This is particularly suitable to infer miRNA regulatory effects from tissue expression profiles, in which these effects are subtle at the level of individual genes but impact large number of genes. The proposed method requires as input genome-wide miRNA and gene expression data from the same biological samples and sequence based predicted miRNA target sets. The algorithm ranks all genes based on the correlation between their expression with a given miRNA and it computes an enrichment score for a given miRNA based on both the correlation between gene and miRNA expression levels and the sequence based target predictions. As final result we obtain miRNAs showing statistically significant enrichment of their targets at either the top or bottom of the ranked list, which we consider as potential regulators in the analyzed conditions.
**Results:** We applied our algorithm to a breast cancer dataset including a cohort of 129 patient samples. A total of 136 miRNAs have been identified as potential regulators with FDR < 0.1. These observations have been validated on a second publicly available breast cancer dataset from The Cancer Genome Atlas (TCGA) project. The overlap between results obtained in the two datasets contains 33% of significant miRNAs (P V = 10−2 ). Among the top significant miRNAs, we found several members of the miR 17-92 cluster and of the miR-106b paralog cluster. Expression of these miRNAs is upregulated in several types of cancer, and they are considered oncomirs [1]. Finally, we concentrated our study on triple negative breast cancer to select miRNAs potentially relevant in this particular subgroup. [1] Hamilton MP, Nat Commun 2013 Nov 13;4:2730
**Contact email:** loredana.martignetti@curie.fr

Oral presentation: Omics 5

**euNet: a data-driven interactome building tool for very large interaction networks**

Palluzzi F (1) Jalili V (1) Pepe D (2) Pinoli P (1)

*(1) Department of Electronic, Information and Bioengineering, Politecnico di Milano, Milano*
*(2) Department of Brain and Behavioural Sciences, University of Pavia, Pavia*

**Motivation:** In the last few years, the rapid growth of publicly available interaction data, has been facilitating the extensive analysis of complex interactions among genomic regulatory elements. Yet, big genomic data repositories, such as NCBI GEO, ENCODE or TCGA, and the simultaneous drop in data production costs, encourages efficient data-driven analysis. The first step in analysing interaction networks is to build a reliable and exhaustive interactome. Traditional approaches make use of interaction data annotated on PPI (protein-protein-interaction) databases (e.g., MINT, IntAct), metabolic pathways databases (e.g., KEGG, Reactome), and bio-ontologies (e.g., Gene Ontology (GO)). Despite the advantages of such annotations, they may not be properly representative of our sample characteristics, or biased towards well-known master regulators or metabolic pathways, while the phenomenon underlying our experiments may be subtle, not well studied, or even unknown. We present euNet, a tool for building huge data-driven interaction networks, integrating heterogeneous Next Generation Sequencing (NGS) data, overcoming all the aforementioned limitations.
**Methods:** Our method is based on the concept of Genomic Vector Space (GVS). Having a set S of samples (or annotations) regarding a particular cell line or tissue of interest, and the set G of genes being studied in that experiment. The GVS is defined as the matrix (S x G). We use both RNA-seq data, from different cellular components and for different RNA types, and histone mark ChIP-seq data, to account also for the transcriptional activity of each gene (expressed, repressed and poised). Each gene is represented in GVS as the cluster of its transcriptional units, plus the promoter region (by default 1kb upstream the gene start coordinate). Each sample, in Wiggle format, provides an intensity value for each gene in GVS, calculated as the average signal value. Cosine similarity among genes (i.e., vectors) is calculated pairwise. In this context, angles between vectors are proved to be an effective parameter to assess expression level variation in different conditions. We utilize GVS for: (I) weight edges of interaction network by cosine similarity; (II) cluster vectors according to their expression values and regulation status, (III) enrich each cluster for functional GO terms (Biological Process and Molecular Function). The interaction network is built using chromatin accessibility NGS data (DHS and FAIRE) and transcription factor (TF) ChIP-seq samples. Gene coordinates, including the promoter region, are mapped inside each TF sample file (BED format) on both strands. The TF binding site is then assigned to the nearest Transcription Start Site (TSS). The binding motif for a given TF is then confirmed using MEME ChIP. The network is built connecting the TF to the nearest TSS, and so the corresponding gene, with a directed edge from TF to gene. ChIP-seq samples and annotation files require transformations and mappings to be processed. For example, multiple BED and Wiggle files are intersected, merged and ranked. In order to overcome the computational complexity of these operations, we perform an in-memory indexing of genes list and a linear off-memory (i.e. information are kept on hard disk and a single sequential scan is performed on them) processing of semi-continues Wiggle files. euNet can be applied on multiple BED/Wiggle and annotation files, outperforming in computational speed other commonly used tools with same functionalities, such as BEDTools or GROK. The euNet tool can also build GVS from annotation sources, giving the possibility of obtaining different measures of similarity. In this

case, the GVS is a matrix (T x G), where T is the set of terms (i.e. functional annotations). Each gene is represented by a binary vector, where the i-th entry of the vector is equal to 1 if and only if the gene is annotated to the i-th term (or any of its descendant in the Directed Acyclic Graph, when the term is part of an ontological structure).

**Results:** The euNet environment, offers a user-friendly, fully customizable, command line interface, written in Python, Java and R. It offers a flexible and optimal environment for: (1) similarity computation among any kind of genomic region, like for instance genes, enhancers, insulators, and unannotated regions; (2) handle various types of NGS data, in BED/Wiggle format, including annotation files; (3) building interaction networks, from experimental data and/or annotations; and (4) allowing their validation through enrichment analysis, both for GVS and interaction network itself. The first euNet release has been scheduled for Jan. 2014, in the SourceForge repository, under GNU/GPL3 license.

**Contact email:** fernando.palluzzi@polimi.it

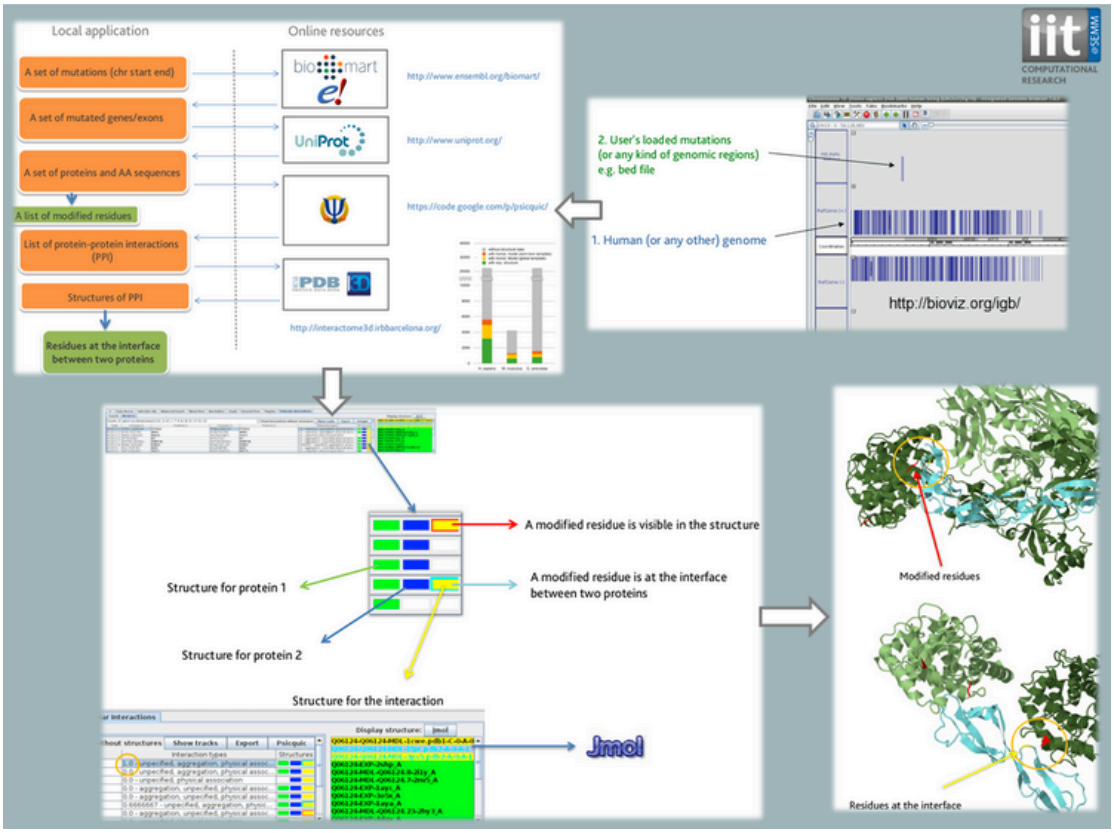
Oral presentation: Omics 6

**From genome to molecular interactions, a plug-in to map next generation sequencing data to molecular interactions and their structures with the Integrated Genome Browser.**

Ceol A., Muller H.

*Center for Genomic Science of IIT@SEMM, Istituto Italiano di Tecnologia (IIT), Via Adamello 16, 20139 Milan, Italy,*

**Motivation:** Current genome browsing still follows the paradigm that was developed more than a decade ago with the advent of the UCSC Genome Browser. The researcher can view annotations located in a tiny piece of the genome. The information is assembled at the server-side and sent to the client as a clickable image using the html map tag that is visualized in a web browser. This model puts all computational loads on the server side and precludes analytics. Modern genome browsers such as the Integrated Genome Browser (IGB) or the Integrative Genomics Viewer (IGV) are stand-alone applications executed on the client computer and connect to remote sources. IGB, in particular, implements a plugin interface enabling customized analytics. Next-generation sequencing has been applied to identify genomic variations possibly associated to many diseases. The next challenge consists in identifying the role of those mutations. Mutations located in protein interaction interfaces are for instance often associated with loss-of-function or gain-of-function. The identification of such residues requires the integration of disseminated biological data and bioinformatics tool. Fortunately, the possibility to extend the existing genome browsers allows us to integrate complex frameworks and make them available through already widely adopted tools.
**Methods:** We developed a framework to automatically map genomic regions to protein sequences and molecular interactions. In order to always work on updated information, all data are collected online through the use of standard web services: the frameworks associate to a set of selected genomic regions a list of genes (Ensembl Biomart, http://www.ensembl.org/biomart/martview/), the proteins they encode (Uniprot, http://www.uniprot.org/), the interactions in which those proteins are involved (Psicquic web services, http://psicquic.googlecode.com/), and the available structures (PDB and interactome3D, http://pdb.org/ and http://interactome3d.irbbarcelona.org) on which the selected residues are mapped . In order to make our framework available to any group independently of advanced computational skills, we implemented it as a new module to extend the Integrated Genome Browser.
**Results:** We present an extension for the Integrated Genome Browser, a widely used tool for the visualization and analyses of genomic data, which maps and visualizes genomic regions in protein-interaction structures and allows the researcher to make educated guesses about the functional impact of somatic mutations. The plugin automatically retrieves all the known molecular interactions from public repositories that involve the product of a selection of genes. The structures available for those interactions can be retrieved and displayed and the mutated residues be highlighted, providing the opportunity to identify those which are at the binding interface and may be the cause of a loss of function. This plugin will be particularly helpful in the analysis of somatic mutations in cancer research, where many passenger mutations must be distinguished from a few driver mutations.
**Contact email:** arnaud.ceol@iit.it
**URL:** http://cru.genomics.iit.it/igbmibundle


Local application
Online resources
A set of mutations (chr start end)
A set of mutated genes/exons
A set of proteins and AA sequences
A list of modified residues
List of protein-protein interactions (PPI)
Structures of PPI
Residues at the interface between two proteins
http://www.ensembl.org/biomart/
http://www.uniprot.org/
https://code.google.com/p/psicquic/
http://interactome3d.irbbarcelona.org/
2. User's loaded mutations (or any kind of genomic regions) e.g. bed file
1. Human (or any other) genome
http://bioviz.org/igb/
A modified residue is visible in the structure
A modified residue is at the interface between two proteins
Structure for protein 1
Structure for protein 2
Structure for the interaction
Jmol
Modified residues
Residues at the interface

Oral presentation: Omics 7

**Ascidian mitogenomics: comparison of evolutionary rates in closely related taxa provides evidence of ongoing speciation events**

Francesca Griggio(1), Ayelet Voskoboynik(2,3), Fabio Iannelli(1), Fabienne Justy(4), Marie-ka Tilak(4), Turon Xavier(5), Graziano Pesole6,7, Emmanuel J.P. Douzery(4), Francesco Mastrototaro(8), Carmela Gissi(1)

*1. Dip. Bioscienze, Università degli Studi di Milano, Milano, Italy 2. Institute for Stem Cell Biology and Regenerative Medicine, Stanford Univ. School of Medicine, Stanford, USA 3. Department of Developmental Biology, Stanford University, Hopkins Marine Station, Pacific Grove, CA 93950, USA 4. Institut des Sciences de l'Evolution de Montpellier (ISEM), UMR5554 CNRS-UM2-IRD, Université Montpellier II, Montpellier, France 5. Center for Advanced Studies of Blanes (CEAB-CSIC), CSIC, Blanes, Spain 6. Istituto di Biomembrane e Bioenergetica, CNR, Bari, Italy 7. Dip. Bioscienze, Biotecnologie e Biofarmaceutica, Università di Bari, Bari, Italy 8. Dip. Biologia, Università degli Studi di Bari, Bari, Italy*

**Motivation:** Ascidians are a fascinating group of filter-feeding marine chordates characterized by rapid evolution of both sequences and structure of their nuclear and mitochondrial genomes. Moreover, they include several model organisms used to investigate complex biological processes in chordates.
**Methods:** To study the evolutionary dynamics of ascidians at short phylogenetic distances, we sequenced 13 new mitogenomes and analyzed them, together with 15 other available mitogenomes, using a novel approach involving detailed whole-mitogenome comparisons of conspecific and congeneric pairs.
**Results:** The evolutionary rate was quite homogeneous at both intra-specific and congeneric level, with the lowest congeneric rates found in cryptic (morphologically undistinguishable) and affinis (distinct but morphologically very similar) species pairs. Moreover, congeneric nonsynonymous rates (dN) were up to two orders of magnitude higher than in intra-species pairs. Overall, a clear-cut gap sets apart conspecific from congeneric pairs. These evolutionary peculiarities allowed easily identifying an extraordinary intra-specific variability in the model ascidian Botryllus schlosseri, where most pairs show a dN value between those observed at intra-species and congeneric level, yet consistently lower than that of cryptic/affinis species pairs. These data suggest ongoing speciation events producing distinct B. schlosseri entities not yet corresponding to the emergence of cryptic species. Remarkably, these ongoing speciation events were undetectable by the cox1 barcode fragment, demonstrating that, at low phylogenetic distances, the whole mitogenome has a higher resolving power than cox1. Our study shows how whole-mitogenome comparative analyses performed on a suitable sample of congeneric and intra-species pairs may allow detecting not only cryptic species but also ongoing speciation events.
**Contact email:** carmela.gissi@unimi.it

Oral presentation: Omics 8

**Retroelements and Long Noncoding RNAs Content Suggests Convergent Molecular Evolution Between Octopus and Mammals.**

Petrosino G, Zarrella I, Ponte G, Musacchia F, Fiorito G, Sanges R.

*Laboratory of Animal Physiology and Evolution, Stazione Zoologica Anton Dohrn, Villa Comunale, 80121 Naples, Italy.*

**Motivation:** Cephalopoda, a class of the Phylum Mollusca, has undergone dramatic evolutionary changes in the body plan and in particular in the morphology of the nervous system [1]. The complexity of the nervous system can be recognized from the brain size, the number of neuronal cells and the neuroanatomical organization and it is comparable to that of vertebrates [2]. In the Octopus vulgaris, the nervous system reaches a great level of complexity both in the Central nervous system (CNS) and Peripheral nervous system (PNS). In the CNS, the supra- (SEM) and the sub-esophageal masses (SUB), and a pair of optic lobes (OL) are clearly defined; those are respectively involved in integrative sensory-motor, motor-control and visio-motor coordination. The PNS, mostly made-up by neural tissue present in the arms (ARM), is involved in the perception and as effector of movements [3]. These observations motivated us to analyze the molecular toolkit of the Octopus nervous system. The final goal of the project is to generate the reference transcriptome of the nervous system to understand, at the molecular level, its organization, development and evolution, allowing comparative analysis across both the different areas as well as multiple animal phyla.
**Methods:** We collected RNA samples from the main areas of the central and peripheral nervous systems (SEM, SUB, OL and ARM) and performed RNAseq on the Illumina HiSeq2000 platform. The transcriptome was assembled using Trinity, the expression levels obtained with Bowtie and the differential expression analysis performed using edgeR [4-6]. Annotations were collected trough a custom pipeline able to perform blastx, rpsblast, blastn against Uniref, CDD, and RFAM databases and identify putative long noncoding RNAs (lncRNAs). In order to perform comparative analysis, we downloaded the transcriptomes of nineteen Mollusks, one Annelida, two Arthropods, one Brachiopod, thirteen Chordates, two Echinoderms and one Nematode from several public databases. Repetitive elements composition of the transcriptomes was evaluated by RepeatMasker and the Repbase database [7,8] searching for bilateria repeats. Custom Perl and R scripts were used to analyze all the collected data.
**Results:** The RNA-seq generated about 850 million paired-end reads accounting for ~85 GB of sequence data. The assembly was processed and filtered producing a set of 64,477 unique transcript isoforms. We functionally annotated 21,030 (32,6%) transcripts and predicted an extremely high number of lncRNAs (7,806, 12,1%). Interestingly, about 41,000 (~64%) of transcripts contain repeats; in particular retroelements appear to be pervasively embedded into the transcriptome. Correlations between lncRNAs and repetitive elements have been reported in the human and mouse transcriptomes, and it has also been demonstrated that lncRNAs contaning SINEs can regulate genes involved in brain functioning and neurodegenerative diseases [9,10]. Thus, lncRNAs and retroelements have already been associated to cognitive abilities in mammals. Comparative analysis of the Octopus transcriptome highlighted a high frequency of transcript-embedded retroelements, similar to that of mammals and much higher than other analyzed species. We found that SINEs in the Octopus are significantly more abundant in lncRNAs than in protein coding genes, and identified an enrichment for both lncRNAs and SINEs in transcripts expressed in the CNS in

respect to the PNS. We speculate that a convergent evolutionary process has led to the evolution of mammalian-like molecular traits in the octopus nervous system, contributing to improve its cognitive abilities, unique among invertebrates.

**Supplementary information:**References 1. Borrelli, L. & Fiorito, G. Behavioral Analysis of Learning and Memory in: Cephalopods. In: Learning and Memory: A Comprehensive Reference (Menzel, R. Volume Editor; Byrne, J. Chief Editor). Elsevier, UK. February 2008. 2. Hochner, B.; An embodied view of octopus neurobiology. Curr Biol. 2012 Oct 23;22(20). 3. Young, J.Z. The Anatomy of the Nervous System of Octopus vulgaris.(1971)-(Oxford: Clarendon Press). 4. Haas, B.J. et al; De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat Protoc. 2013 Aug;8(8):1494-512. 5. Langmead, B. et al; Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 2009 10:R25. 6. Robinson, M.D. et al; edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010 January 1; 26(1): 139-140. 7. Saha, S. et al.; Empirical comparison of ab initio repeat finding programs. Nucleic Acids Res. 2008 Apr;36(7):2284-94. 8. Jurka, J. et al.; (2005) Repbase Update, a database of eukaryotic repetitive elements. Cytogenet Genome Res. 2005;110(1-4):462-7. 9. Kelleyc D. et al.; Transposable elements reveal a stem cell-specific class of long noncoding RNAs. Genome Biol. 2012 Nov 26;13(11). 10. Carrieri, C. et al.; Long non-coding antisense RNA controls Uchl1 translation through an embedded SINEB2 repeat. Nature. 2012 Nov 15;491(7424).

Oral presentation: Omics 9

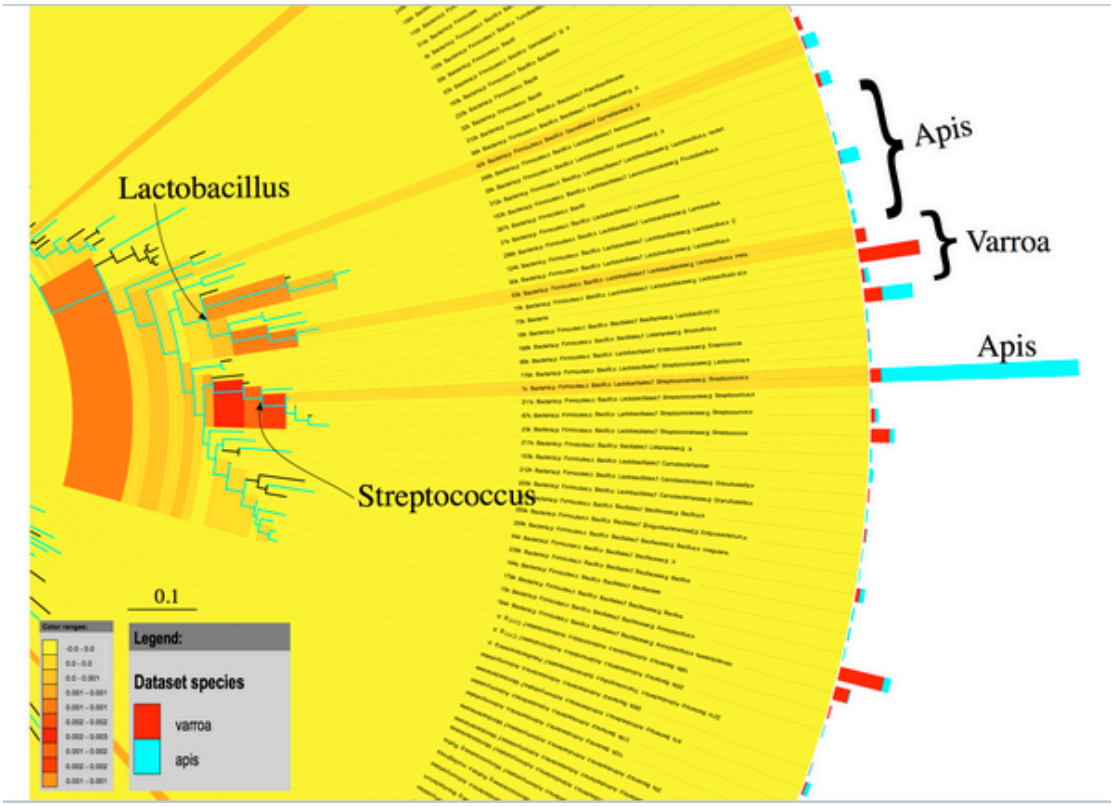
**Mean phylogenetic surprise: a unit of measure for describing changes across biological communities**

Balech B (1), Vicario S (2)

*1) Institute of Biomembranes and Bioenergetics (IBBE), National Research Council, Bari (Italy), 2) Consiglio Nazionale delle Ricerche – Istituto di Tecnologie Biomediche – Sede di Bari, via Amendola 122/D, 70126 Bari, Italy. E-mail: saverio.vicario@ba.itb.cnr.it.*

**Motivation:** The availability of NGS technology increased the amount of communities described by environmental sequencing, while the large effort of the biodiversity scientific community mobilized and exposed very large data sets on the presence of organisms in public repository as the Global Biodiversity information facility or SistemaAmbiente2010 in Italy. These two sources of check lists and species abundance data could be used to build knowledge on how communities are organized and how they correlate with abiotic factors. But, no real consensus exists on how to compare communities both with scalar (Jost, 2006) or phylogenetic indices (McCoy and Matsen, 2013). Generally, phylogenetic indices are expected to be more insightful. In fact, the presence of a given organism within a community depends on its ecological attributes built through the history of evolutionary adaptations and innovations of its lineage. So related organisms are likely to stay in a given community for shared reasons. This consideration makes clear that the correct level of taxonomic classification to study the correlation between abiotic factor and communities is not always species level. This strategy of including the contribution of higher level classification decreases the effect of low counts (McCoy and Matsen, 2013). Until now, all phylogenetic approaches are within the frame of index based approach where no clear statistical framework links the index to the phenomenon of interest, making difficult to include replicates within the experimental design and in general to interpret the results.
**Methods:** In this work, we show the phylogenetic entropy of degree one, a family member of diversity indices, proposed as by Chao et al. (2010), is indeed a true measure of information, that describes the mean surprise of an observer, sampling organisms and classifying them using a given rooted phylogenetic tree. This measure is used to build a formal measure of beta diversity that takes into account phylogenetic structure of communities that is measuring the decrease of surprise from the marginal to the conditional distribution of observation. This phylogenetic beta diversity measure is used to produce an ANOVA-like procedure which uses a phylogenetic tree of all organisms found in all sampled communities and highlight the contribution of each branch of the phylogeny and each group of communities present in the experimental design to the diversification across groups. This framework allows to include replicates (technical or biological) within the pipeline of analysis given that each groups could be represented by several sample communities.
**Results:** The proposed metric is numerically validated as information measure by testing different required properties of information metrics. Precision, accuracy and power of the procedure was tested using simulation of community with log normal distribution of abundance in which each group of community differs in the abundance of a given clade. Results are compared with Shannon based beta diversity estimates. An example use case is used to illustrate the web implementation of the procedure. The figure shows how the method correlates abundances of the leaf in the different groups with branch contribution to beta diversity in the phylogeny.
**Contact email:** saverio.vicario@ba.itb.cnr.it

**Supplementary information:**References Jost L. Entropy and diversity. Oikos. 2006;113(2):363?375. Chao A, Chiu C-H, Jost L. Phylogenetic diversity measures based on Hill numbers. Philosophical transactions of the Royal Society of London. Series B, Biological sciences. 2010;365(1558):3599?609. DP Faith. 1992. Conservation evaluation and phylogenetic diversity. Biological Conservation 61: 1-10. McCoy CO, Matsen FA. Abundance-weighted phylogenetic diversity measures distinguish microbial community states and are robust to sampling depth. arXiv: Quantitative Biology - Population and Evolution. 2013;1305.0306v.



Oral presentation: Omics 10

**KINhibit: Predicting resistance to kinase inhibitors from cancer genomics data**

Palmeri A(1), Creixell P(2), Longden J(2), Helmer-Citterich M(1), Ferkinghoff-Borg J(2), Linding R(2)

*(1) Centre for Molecular Bioinformatics, Department of Biology, University of Rome Tor Vergata (2) Cellular Signal Integration Group, Department of Systems Biology, Center for Biological Sequence Analysis, Technical University of Denmark*

**Motivation:** The pharmacological targeting of constitutively active signaling networks by kinase inhibitors, one of the most common cancer therapies, typically achieves high short-term success rates. However, in most cases, tumors develop resistance to these treatments in the long term by accumulating specific mutations. The aim of KINhibit is to predict genomic lesions that lead to resistance by perturbing protein kinases and their signaling networks.
**Methods:** Using data from a recently published and comprehensive study on kinase binding profiles (1) and sequence information from a kinome-wide multiple sequence alignment, KINhibit trains and optimizes different binding masks following a Gaussian Process coupled with a Genetic Algorithm-dependent optimization.
**Results:** KINhibit performances in predicting the kinase-inhibitor dissociation constant have been assessed at different redundancy levels between training and test datasets. All sequence similarities have been computed on the kinase domains only. The kendall correlation coefficient between predicted and observed dissociations constants across all inhibitors is still 0.2 at 70% redundancy. The error associated with the predictions resulted close to the experimental one. More than one third of the models generated for all the known inhibitors perform well also in the classification of binding and non-binding kinases, like for instance in the case of Imatinib (AUC = 0.78 +/- 0.1). KINhibit is in the process of being validated with an in vitro proliferation assay in 6 cancer cell lines.
**Supplementary information:**(1) Davis, M.I., Hunt, J.P., Herrgard, S., Ciceri, P., Wodicka, L.M., Pallares, G., Hocker, M., Treiber, D.K., and Zarrinkar, P.P. (2011). Comprehensive analysis of kinase inhibitor selectivity. Nat. Biotechnol. 29, 1046–1051.

## Special session: Synthetic and Systems biology

**Advancements and open questions in synthetic cells: new challenges for systems biology?**

Stano P

*Department of Sciences, University of Roma TRE, Rome*

**Background**

Born within the community of origin of life research, current synthetic cell technology derives from the convergence of liposome technology and cell-free systems. Such an approach has been recognized as one of the emerging trends in synthetic biology, and it complements the more usual bioengineering "top-down" design, based on manipulation of extant biological cells.

Minimal synthetic (or semi-synthetic) cells are built by encapsulating inside liposomes the minimal number of molecular components required for reconstructing defined cellular functions. Depending on the way in which they are designed and constructed, synthetic cells can be used as model of primitive cells, or as systems for biochemical investigations and biotechnological applications.

In origin of life studies, great emphasis is given to spontaneous and self-organizing processes; in other cases, instead, the control of assembly and behavior is central.

**Results**

In this contribution we will first introduce the concept of synthetic minimal cell, briefly commenting on the main historical developments and on the state-of-the-art of current research.

Synthetic cell research has often focused on basic cell mechanisms like solute entrapment, protein synthesis, DNA and RNA replication, lipid synthesis – generally operating within the theoretical framework of autopoiesis. The implantation of minimal transcription-translation systems for protein synthesis inside liposomes holds a key role for synthetic cells designed to perform specific functions.

With respect to these themes, we will comment on the open questions about the construction and the behavior of synthetic cells, giving emphasis to processes that are still poorly understood or that need to be integrated with computational approaches. Future developments, for instance in the field of biochemical-IT (Information Technology), will be also shortly commented.

**Conclusions**

The ultimate, ambitious goal of synthetic cells research is the assembly of a minimal living cell from inanimate matter. This goal should be approached stepwisely, by building cell-like systems of increasing complexity. However, since such a systems are built from minimal and well-characterized parts, their behavior can be in principle understood and fully predicted. As judging from the most recent issues, synthetic cell research will benefit from the integration with bioinformatics and systems biology approaches, as it happens in synthetic biology.

In fact, recent experimental advancements pave the way to future realistic scenarios where automatic design and numerical modeling will be integrated in this emerging field.

**Bottom-up design of genetic circuits: characterization and re-use of biological building blocks to engineer predictable systems**

Pasotti L(1,2), Zucca S(1,2), Politi N(1,2), Casanova M(1,2), Mazzini G(3), Cusella De Angelis MG(2), Magni P(1,2)

*(1) Dipartimento di Ingegneria Industriale e dell'Informazione, Università degli Studi di Pavia, Pavia (2) Centro di Ingegneria Tissutale, Pavia (3) Istituto di Genetica Molecolare, Consiglio Nazionale delle Ricerche, Pavia*

**Motivation:** In the future, Synthetic Biology may contribute solutions to a plethora of practical problems, from renewable energy production to disease treatment, by constructing complex genetic circuits which implement the pathways of interest. Microbes can be engineered to incorporate such functions or, otherwise, cells could be designed and built up from the scratch to yield a fully synthetic customized organism. In order to exploit the whole potential of Synthetic Biology, engineers must be able to realize the desired functions in the form of biological systems with predictable behaviour. Only in a modular framework this task can be accomplished and biological parts (e.g., promoters, ribosome binding sites - RBSs, etc.) or higher-order circuits (e.g., genetic logic inverters, amplifiers, etc.) can be effectively re-used. However, modularity boundaries of biological parts still need to be elucidated and the study of model systems, aided by the use of mathematical models, is carried out towards this goal. This abstract describes the recent efforts undertaken by the Synthetic Biology working group at the University of Pavia to investigate if biological parts can behave predictably when re-used in different contexts.
**Methods:** A collection of synthetic circuits have been designed, built up, modelled and experimentally characterized. They include a number of design motifs commonly found in nature and also in other circuits, such as: constitutive/inducible production of a target protein, transcriptional activators/repressors cascades and feedback-controlled regulation of a target signalling molecule. A rigorous bottom-up approach was adopted during the design process: first, basic parts and sub-circuits composing the final systems were individually characterized with the support of mathematical models that describe their steady-state and dynamic behaviour; then, the quantitative behaviour of the final composite systems was predicted from the knowledge of their individual components and, finally, predictions were compared with experimental data. Experimental characterization was carried out in vivo using Escherichia coli as chassis, which was engineered with ad-hoc constructed plasmids containing the DNA encoding the circuit of interest. Fluorescent reporter proteins (e.g., GFP) were used to indirectly measure the transcriptional output of a circuit. Pre-characterized inducible promoters were used as input blocks to measure the input-output transfer function of circuits that require a transcriptional signal as input. Engineered bacteria were grown in test tubes, in 96-well microplates or in a chemostat implemented with the Lambda Minifor bioreactor. Fluorescence assays were performed with a microplate reader (Tecan Infinite F200) or a flow cytometer (Partec PAS II) and absorbance at 600 nm was always used to estimate bacterial density. A model-based approach was adopted to analyze fluorescence and absorbance data. Empirical ordinary differential equation models based on Hill functions were used to describe the dynamic behaviour of the designed circuits and the least squares method was used to fit model parameters. A crucial issue is the propagation of uncertainty in parameter estimates through the network of interconnected modules. To consider this,

Monte Carlo methods were used and parameter distributions were obtained to support the prediction of the final circuit output.

**Results:** Three main classes of circuits were tested. The first class includes a set of constitutive and inducible promoters of graded strengths. Their output was measured in several contexts: different reporter genes, RBSs, strains and plasmids to investigate the conditions in which promoter activity changes unpredictably. The second class of circuits includes a set of transcriptional regulator cascades. They are composed by interconnected genetic logic inverters (i.e., simple circuits including a repressor gene and a specific repressible promoter) regulated by an inducible promoter. Cascades of two or more units were tested to evaluate the predictability of the output as a function of circuit complexity. The third class of circuits includes feedback-controlled networks for the close-loop regulation of a target molecule, constructed by re-engineering a natural quorum sensing network. The synthetic circuits include 1) a production module for the target molecule, under the control of a promoter triggered by an exogenous signal, and 2) a degradation module, regulated by a specific target molecule sensor, thus completing the negative feedback and yielding a higher-order complexity compared to the other classes of circuits. All the tested circuits showed the expected qualitative behaviour and mathematical model predictions with uncertainty propagation enabled the evaluation of their quantitative behaviour consistency. This work has contributed the characterization of a large amount of circuits and steps towards the learning of predictability boundaries of biological components.

**SIGNOR: a new resource to support experimental investigation of signaling networks.**

Gianni Cesareni(1,2), Leonardo Briganti(1), Alberto Calderone(1), Luisa Castagnoli(1), Francesca Langone(1), Milica Marinkovic(1), Anna Mattioni(1), Theodora Pavlidou(1), Daniele Peluso(2), Livia Perfetto(1), Daniela Posca(1), Elena Santonico(1), Alessandra Silvestri(1), Filomena Spada(1).

*(1) Department of Biology, University of Rome Tor Vergata, Rome. (2) IRCCS Fondazione Santa Lucia, Rome.*

**Motivation:** Given a sufficiently large experimental dataset, reverse engineering techniques allow de novo inference of gene networks, in the absence of any a priori knowledge. By this approach the whole interaction space is searched for connections explaining the experimental data. This computationally intensive strategy is applicable, on a large scale, to gene regulatory networks because of the vast amount of data accumulated with gene array and RNAseq experiments. However, no such high-content high-throughput experimental technology exists for signaling networks. Thus, de novo network reconstruction is not feasible for large biochemical networks and model assembly, or better optimization, is achieved by confronting the experimental data with an interaction subspace constrained by available literature evidence.
**Methods:** SIGNOR, the SIGnaling Network Open Resource, was developed to support experimental approaches based on multi-parametric analysis of cell systems. Typically, in such approaches, the state of a cell and its dynamic changes upon system perturbation are revealed by antibodies monitoring the activation of key sentinel proteins. The experimental results are then compared with a literature derived logic network thereby allowing for network optimization. Although several databases representing pathway relationship according to different data model exists (Reactome, KEGG, …) we are not aware of any project aimed at the coverage of logic relationships between signaling proteins . SIGNOR was developed to fill this gap by offering a large network of experimentally validated logic relationships between signaling proteins that can be used as an a priori model for optimization strategies.
**Results:** The core of SIGNOR is a collection of approximately 5000 manually-annotated logic relationships between proteins that participate in signal transduction. Each relationship is linked to the literature reporting the experimental evidence and is assigned a score based on experimental support. In addition each node is annotated with the chemical inhibitors that modulate its activity and, when available, the antibodies that can be used to monitor the activity of the nodes of the network are also annotated.
**Contact email:** cesareni@uniroma2.it
**URL:** http://160.80.34.9:8084/SIGNOR/

**Stochastic differentiation and homeostasis in a multiscale model of intestinal crypt dynamics**

Graudenzi A(1), Caravagna G(1), De Matteis G(2), Antoniotti M(1)

*(1) Dept. of Informatics, Systems and Communication, University of Milan-Bicocca, Milan (2) Dept. of Mathematics and Information Sciences, Northumbria University, Newcastle, UK.*

**Motivation:** Intestinal crypts are invaginations in the intestine connective tissue and are the loci in which colorectal tumors are supposed to originate and develop. Some of the key structural and dynamical properties of crypts have been characterized as, for instance, the coordinate migration of cells dividing and progressively differentiating from the stem cell niche at the bottom of the crypt toward the intestinal lumen, divided in distinct layers of cell populations of different types. However, an overall picture of the complex interplay between gene regulation and the general dynamical behavior of the cells populating the crypt is still missing. In particular, we here investigate the relation between the phenomenon of stochastic differentiation and the overall homeostasis of the system.
**Methods:** We here introduce the results of the analyses on a multiscale model of intestinal crypt dynamics, which describes (i) a gene regulatory network-based model of cell differentiation, as that provided by Noisy Random Boolean Networks and (ii) a morphological model of crypt dynamics, as that given by the Cellular Potts Model. The GRN model is focused on the emergent properties of the networks, i.e. the dynamical gene activation patterns standing for the modes of functioning of the cells, the resistance of which against noise is related to the degree of differentiation. On the other hand, the spatial model is ruled by an energy minimization criterion, according to which cells tend toward optimal size and position settings. Both the models rely on a few a priori assumptions only and the focus is on the properties emerging from the dynamics: in particular, the GRN dynamics drives the processes of cell growth and cell differentiation at the spatial level. Several key processes related to the functioning of the crypt are represented in the model, such as: gene activation patterns, stochastic differentiation, signaling pathways, cell movement, cell growth, mitosis and apoptosis. Particular attention is also devoted to the investigation of the influence of biological noise on the overall dynamics.
**Results:** We here show that stochastic differentiation might be sufficient to drive crypts toward homeostasis, according to some key quantitative and qualitative measures, which were validated against experimental data and the current biological knowledge, when possible. In particular, the correct stratification of cell populations belonging to different types is shown to be maintained under distinct initial spatial configurations of the system. The stem cell niche is also preserved. Nevertheless, an increase of the level of initial spatial disorder can lead to the appearance of systems in which the correct stratification is not ensured. Also, the proportion of cell populations is kept within a range that is in accordance with experimental evidences in distinct configurations of the crypt, as well as the rate of newborn and dead cells, hinting at the intrinsic capability of the system to ensure a correct dynamical turnover, i.e. the renewal of the tissue. Furthermore, from experimental results it is known that cells at the bottom of the crypt move slower toward the top than cells in the upper part and this was reproduced in our model, by highlighting a correlation between the distance from the bottom of the crypt and the average vertical velocity of the cells. Besides, the average vertical velocity of cells was observed to range from around 0 micron/hour at the bottom of the crypt to 2,5 micron/hour at the top. By estimating the average time needed for a random descendent of a stem cell to complete the progressively faster

migration toward the lumen, it turns out that around 65 hours (~3 days) are needed, a result that is in perfect agreement with experimental data. Experimental evidences suggest also that epithelial cells migrate in coordination as sheets in culture and this was reproduced in our model, as proven by different measures of spatial correlation, such as Moran index, Pearson correlation and other correlation measures. Besides, our model implementation allows to track the descendants of each stem cell in the crypt, permitting to investigate the process of clonal expansion. Regardless of the initial configuration, in most cases all proliferative cells have a small number of descendants, whereas only in certain cases a few cells actually show a relatively high number of descendants. However, even the most proliferative cells fails in colonizing the whole crypt, pointing once more to the homeostasis of the system, in which a correct proportion of the cell populations is kept along the course of the simulation. Finally, we remark that analyses underway are aimed at investigating how the progressive accumulation of mutations and alterations at the GRN level may induce the appearance of aberrant structures and behaviors at the spatial scale, eventually leading to the emergence and development of colorectal tumors.

**Contact email:** alex.graudenzi@unimib.it

**Co-evolving proteins to infer protein-protein interactions: a re-assessment of the phylogenetic profiling**

Sferra G.(1), Santoni D.(2)., Ponzi M.(1), Pizzi E.(1)

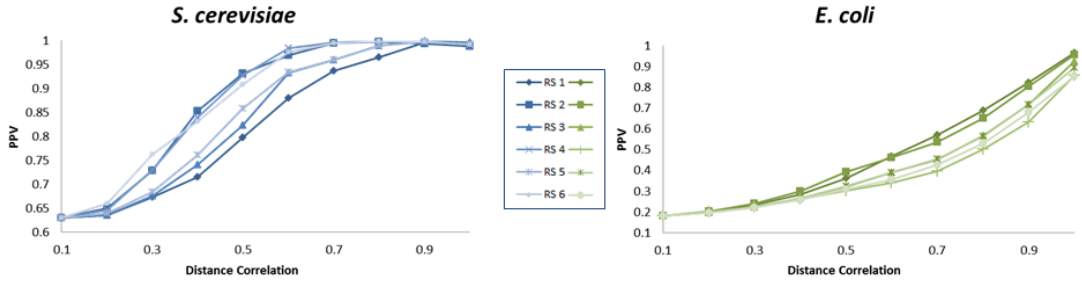
*(1) Istituto Superiore di Sanità, Roma (2) Istituto di Analisi dei Sistemi ed Informatica - CNR, Roma*

**Motivation:** Phylogenetic profiling is one of the mostly used methods to infer protein-protein interactions from genomic data, the basic idea being that co-evolving proteins are also functionally related. Phylogenetic profiles are vectors containing probability values calculated according to P = -1/log2(E), where E are the E-values obtained aligning proteins of a target organism and proteins from genomes in a reference set. Co-evolution is established by a similarity measure between profiles. It is known that both size and composition of the reference set affect the method outcome, however a workflow to decide the number and kind of organisms to be considered has never been proposed. Nowadays, more than seven thousands sequenced genomes are available to address this point. Mutual Information is the election-method for profile comparison due to its capability to capture linear and non-linear correlations. Despite of this, it has been recently shown that correlation-based measures are more sensitive than the non-metric ones. In this study we applied a recently proposed statistical measure (Distance correlation, DC), successfully exploited in other fields, to carry out a re-assessment of the phylogenetic profiling. In particular, we constructed reference sets different in size and genome composition and tested the predictive performance of DC with three target organisms.
**Methods:** On the basis of the main Tree of Life, we proposed novel criteria to construct the reference sets. First, from the 1133 species available in the eggNOG database (version 3.0, http://eggnog.embl.de/), we defined each species as “core” or “peripheral” according to the distance from the common ancestor. The reference sets were constructed subtracting progressively “peripheral” species. BLAST database searches were performed between proteins of the target organism and proteins in the RSs. Alignments with an E-value < 10-1 are retained and used to construct the phylogenetic profiles. To perform all vs all phylogenetic profile comparison of the target organism, we developed a parallel version of the software in the R “energy” package. The software is adaptable to a wide range of machines making possible the calculation of DC to large sets of data on a standard Work Station (Windows/Linux/MacOS). To evaluate the prediction performance, the results were compared with Gold Standards inferred from the KEGG pathway database (http://www.kegg.jp/). Proteins belonging to the same pathway were considered as True Positives. A graph-based algorithm was developed to select True Negatives as protein pairs belonging to non-overlapped pathways.
**Results:** An assessment was performed on the six reference sets using phylogenetic profiles derived for Saccharomyces cerevisiae and Escherichia coli proteins. Results are shown in the figure. Best predictive performances were obtained for phylogenetic profiles derived from RS2, RS4 and RS6 in the case of S. cerevisiae, and from RS1 and RS2 in the case of E. coli. According to this result, RS2 was utilized for next analyses. We used Plasmodium falciparum proteins as a benchmark. This protozoan cell contains a characteristic organelle (apicoplast) originated from two events of endosymbiosis, each of which corresponds to a gene transfer from the engulfed organism to the nuclear DNA of the Plasmodium. Clustering analysis of phylogenetic profiles reflects the evolutionary origins of the parasite proteins, as proteins

derived from the cyanobacteria and the green alga, the two engulfed organisms, form distant clusters.

**Contact email:** gabriella.sferra@hotmail.it

**Supplementary information:**Name, composition and size of the reference sets: RS1 (120 eukaryotes; 594 prokaryotes; 61 archea; a total of 774 organisms); RS2 (45 eukaryotes; 594 prokaryotes; 61 archea; a total of 699 organisms); RS3 (45 eukaryotes; 230 prokaryotes; 61 archea; a total of 336 organisms); RS4 (18 eukaryotes; 230 prokaryotes; 61 archea; a total of 309 organisms); RS5 (22 eukaryotes; 115 prokaryotes; 30 archea; a total of 167 organisms); RS6 (9 eukaryotes; 115 prokaryotes; 30 archea; a total of 154 organisms);



## Oral presentations

Oral presentation: Algorithms for Bioinformatics 1

**A dynamic programming algorithm for Haplotype Assembly of Future-Generation Sequencing Reads**

Murray Patterson (1), Tobias Marschall (1), Nadia Pisanti (2), Leo van Iersel (1), Leen Stougie (1,3), Gunnar W. Klau (1,3), Alexander Schonhuth, (1).

*(1) Life Sciences, CWI , Amsterdam, The Netherlands (2) Department of Computer Science, University of Pisa, Italy (3) VU University Amsterdam, The Netherlands*

**Motivation:** The human genome is diploid, that is each of its chromosomes comes in two copies. This requires to phase single nucleotide polymorphisms (SNPs), that is, to assign them to the two copies, beyond just detecting them. The resulting haplotypes, lists of SNPs belonging to one copy, are crucial for downstream analyses in population genetics. Currently, statistical approaches, which avoid making use of direct read information constitute the state-of-the-art. Haplotype assembly, which addresses phasing directly from sequencing reads, suffers from the fact that sequencing reads of the current generation are too short to serve the purposes of genome-wide phasing. Future sequencing technologies, however, bear the promise to generate reads of lengths and error rates that allow to bridge all SNP positions in the genome at sufficient amounts of SNPs per read. Existing haplotype assembly approaches, however, precisely profit, in terms of computational complexity, from the limited length of current generation reads, because their runtime is usually exponential in the number of SNPs per sequencing read. This implies that such approaches will not be able to exploit the decisive benefits of long enough, future-generation reads.
**Methods:** Here, we suggest WhatsHap, a novel dynamic programming approach to haplotype assembly. It is the first approach that yields provably optimal solutions for the minimum error correction (MEC) problem in runtime linear in the number of SNPs per sequencing read, making it suitable for future-generation reads. Our approach is FPT with coverage as the parameter.
**Results:** We demonstrate that WhatsHap can handle datasets of coverage up to 20x, processing chromosomes on standard workstations in only 1-2 hours. Our simulation study shows that the quality of haplotypes assembled by WhatsHap significantly improves with increasing read length, both in terms of genome coverage as well as in terms of switch errors. The switch error rates we achieve in our simulations are superior to those obtained by state-of-the-art statistical phasers.
**Contact email:** pisanti@di.unipi.it
**Supplementary information:***This work will be presented at RECOMB 2014* This abstract can fit several sessions: - Genomics - Next Generation Sequencing - Algorithms for Bioinformatics - Genome sequencing, resequencing and annotation. How to survive the jungle in the “Hot Omics Era"

Oral presentation: Algorithms for Bioinformatics 2

**Vector Integration Site Identification through Density-based methods**

Andrea Calabria(1), Stefano Brasca(1), Giulio Spinozzi(1), Eugenio Montini(1)

*(1)HSR-TIGET - The San Raffaele Telethon Institute for Gene Therapy, Safety of gene therapy and insertional mutagenesis Unit; San Raffaele Scientific Institute, Division of Regenerative medicine, Stem cells, and Gene therapy; Milan; Italy*

**Motivation:** Vector integration sites (IS) analysis is being increasingly used to monitor safety and efficacy of gene therapy (GT) treatments, from preclinical studies to clinical trials, as well as to get insights on matter of basic biological research since IS are exploited as unique genetic markers, specific for cells harboring such proviral insertions and stably transmitted to progeny. State of the art PCR-based methods for retrieval of vector-host genomic junctions, joined with high-throughput sequencing analyses and bioinformatics pipelines, allow IS detection and cell tracking. Each biochemical process, such as PCRs or NGS sequencing, may introduce some artifacts among read sequences (to different extent according to underlying processes) leading to slight alterations in exact IS mapping location. Recent GT lentiviral vector clinical trials take advantage of a rigid sliding-window approach to identify IS: under this method, for each interval of 3 nucleotides covered by reads in the targeted genome, ISs are located in the first base of the window, regardless of internal reads arrangement and without taking account of reads' distribution fashion, peaks location or any statistical consideration about artifacts. This way, misleading mapping of ISs' might occur along with under/over-estimation of the overall number of IS. Here we propose a density-based method to identify ISs, thereby overcoming such limitations, which considers the peak of piled-up reads as the real mark of a single vector integration event despite sequence/mapping errors; similar approaches have been already successfully implemented in other NGS fields, such as ChIP-seq or Methylation quantification. Furthermore, a density-based method is more robust and reliable since it reduces false positives, especially when applied to highly targeted regions, seen as archetype of malignant events.
**Methods:** The density-based approach we designed starts acquiring all mapped reads, accounting starting bases with their reads pileup count, then splits targeted regions into sub-regions (ensemble) of neighboring reads, defining for such ensembles an histogram of covered bases with heights corresponding to the count of piled-up reads. Once all ensembles have been detected, our procedure identifies ISs using a 3 steps model upon each one: (1) Exploration, detecting all peaks of the ensemble, (2) Evaluation, scoring all bases surrounding each peak, and (3) Decision, identifying ISs among local peaks and their surrounding bases. The Exploration step incrementally detects local peaks and, for each one, it considers the first N nearest-neighbors bases as related to the same biological process underlying the peak: the Evaluation process then assigns a score to such bases with respect of the peak; N parameter is given as input, typically is 6 (up to 3 base far from each side of the peak as noted in literature). The score is derived from a comparison between a theoretical statistical curve (e.g. Gaussian process) and the curve derived from the histogram of the peak and its surroundings, thus generating during the incremental process consecutives empirical distributions with potential overlapping ("conflict area" whose covered bases seems to belong to different peaks). Decision step will determine the most likely attribution, assigning all the covered bases of the ensemble univocally to specific peaks, using score comparisons: all final peaks, together with surrounding and assigned covered bases will be converted into unique ISs, located according to peaks placement and characterized by an overall reads count.

**Results:** We implemented this density-based IS identification approach in Python, leveraging on a MySQL database of mapped reads. We exploited the datasets of the two published hematopoietic GT clinical trials with self-inactivating lentiviral vectors (metachromatic leukodystrophy and Wiskott–Aldrich syndrome) ongoing at TIGET, supplied with a follow up of 18 months after treatment resulting in >10 million proviral-host genomic junctions from samples of 6 patients. We compare our new approach versus the sliding-window one then validate it in silico through a simulated test dataset, yielded by a mathematical process that mimics PCR procedures, including a wider range of potential mapping scenarios to estimate the precision and recall of both methods. Exploiting these data and comparisons, we provided a quantitative measure of improvements in IS identification adopting this density-based method. On the computational side, both implementations perform similarly with respect to the number of retrieved mapped sequences under the same conditions. Our results show that this density-based method for IS identification and mapping is more precise and reliable than the sliding-window strategy.
**Contact email:** brasca.stefano@hsr.it

**Fast randomized approximate string matching with succinct hash data structures**

Prezza N(1), Policriti A(1,2)

*(1) Department of Mathematics and Computer Science, University of Udine, Udine, Italy (2) Istituto di Genomica Applicata, Udine, Italy*

**Motivation:** The advent of New Generation Sequencing (NGS) technologies opened a new era in the field of DNA sequencing providing the researchers with powerful instruments able to produce millions of short (and, lately, long) reads per single run. This technology breakthrough poses considerable computational challenges since the sequenced fragments need to be quickly aligned - usually admitting errors - against genomes whose size is often of the order of the gigabases. From the algorithmic point of view, the problem of indexing texts to support pattern matching in the big-data domain is receiving significant attention, also due to the recent computational breakthroughs in the fields of succinct and compressed text indexes (a typical example is the FM self-index used nowadays by many aligners, for example Bowtie and BWA). Even though these important results solved most of the problems related to exact pattern matching, the problem of indexing a text with a succinct - or compressed - data structure that supports approximate pattern matching still represents a considerable challenge. In particular, the most significant results to date are able to guarantee only one between space and speed efficiency: not both at the same time. Indexes that are fast (i.e. are able to solve the problem in time that is linear with respect to the query length) require much of space to be stored, often too much for practical uses. On the other hand, light indexes (requiring only linear or compressed space with respect to the text size) require too much time to answer a query, especially if the number of allowed errors is particularly high.
**Methods:** We approach the problem with a hash-based randomized algorithm that is able to reach expected fast performances and linear space requirements at the same time. The two goals are obtained with the use of two particular classes of hash functions: Hamming-aware and de Bruijn hash functions, respectively. The former permit to "squeeze" the Hamming sphere of radius k centered at the query pattern P, to a Hamming sphere of radius O(k) centered at the hash of the query. Thanks to these functions the algorithm's expected complexity reaches performances comparable in complexity with the fastest algorithms available in the literature for the same problem. The second class of hash functions (de Bruijn) characterizes the hash functions which are homomorphisms on de Bruijn graphs. We show that, using this particular class of hash functions, the corresponding hash index can be represented in linear space introducing only a small slowdown of O(log|P|) for the lookup operation. We call dB-hash our succinct hash data structure. We study the time complexity of our algorithm in the framework of smoothed analysis, a tool introduced by D. Spielman and S.H. Teng which interpolates continuously between worst and average case assuming that the input instance is perturbed by random noise. This analysis tool is particularly suitable in the biological context of DNA sequence alignment since both genomes and sequenced reads are subject to random sequencing errors, random point mutations and further kinds of noise coming from other sources. The theoretical results that we obtain are general, and can be used as a basis for the smoothed analysis of hash-based algorithms using any kind of XOR-based hash function (i.e. functions computed using the bitwise XOR operator).
**Results:** We show that the presence of random noise perturbing the problem instance (i.e. the text and the query) improves the expected complexity of our algorithm in that it has the

effect of distributing uniformly the load in the hash. This is a common result in smoothed analysis, where often the conclusion is that the presence of noise somehow makes the problem easier to solve. In particular, assuming that each bit of the instance is perturbed (i.e. flipped) with probability p, we show that for any p>0 there exists a minimum pattern's length m such that if the structure indexes text substrings of length at least m, then the hash load is uniformly distribuited. This is a strong result since it does not make restrictive assumptions on the amplitude p of the noise, which is only constrained to be greater than zero. Our algorithm has been implemented in the short-reads aligner BW-ERNE, the natural adaptation of the hash-based ERNE aligner to the dB-hash data structure. Experimental results on the vitis vinifera genome show that the dB-hash requires 8 times less space than the standard hash used by ERNE. Tests on reads coming from a NGS experiment show moreover that BW-ERNE maintains the same alignment efficiency of ERNE with only a small penality in the query times (due to the fact that the lookup operation is more expensive on the succinct version of the data structure). The comparison with other aligners such as Bowtie and BWA shows, finally, that BW-ERNE is able to attain both the advantages from succinct data structures (small space) and hash indexes (alignment efficiency).
**Contact email:** prezza.nicola@spes.uniud.it

**Differential diagnosis of pleural mesothelioma using Logic Learning Machine**

Parodi S(1), Filiberti R(2), Marroni P(3), Montani E(1), Muselli M(1)

*(1) Institute of Electronics, Computer and Telecommunication Engineering, National Research Council of Italy, Genoa. (2) Epidemiology, Biostatistics and Clinical Trials, IRCCS AOU San Martino-IST, Genoa, Italy. (3) Department of Laboratory Medicine, IRCCS AOU, San Martino-IST, Genoa, Italy*

**Motivation:** Tumour markers (TMs) are standard tools for the differential diagnosis of cancer, but in some cases the occurrence of non-specific symptoms and different malignancies involving the same cancer site may cause a high proportion of misdiagnoses. Classification accuracy can be improved by combining information from different TMs using standard data mining techniques, like decisional tree (DT), artificial neural network (ANN) and k-nearest neighbour (KNN) classifiers. Unfortunately, each of these methods suffers from some unavoidable limitations. In particular, DT tends to show a low classification performance, whereas ANN and KNN produce a black-box type classification that does not provide biological information useful for clinical purposes. Among malignancies whose correct diagnosis is often difficult to be achieved, malignant pleural mesothelioma (MPM) is one of the most highly fatal, and its incidence is rapidly increasing due to the widespread past exposure to asbestos in both environmental and occupational settings. The correct diagnosis of MPM is often hampered by the presence of atypical symptoms that may cause misdiagnosis with either metastasis from other malignancies (MTX) or benign pleurisies (BP). Cytological examination (CE) following thoracoscopic-guided biopsy may allow to identify malignant cells, but the high prevalence of non-neoplastic cells may cause a severe underestimate of specificity. Moreover, a positive results from CE examination does not allow to distinguish MPM from MTX. Many TMs have been demonstrated to be useful complementary tools for the diagnosis of MPM, in particular the soluble mesothelin-related peptide, SMRP, CYFRA 21-1 and CEA1. However, a method to combine information from such TM and CE for the differential diagnosis of MPM does not exist yet.
**Methods:** Logic Learning Machine (LLM) is an innovative method of supervised data mining that represents an efficient implementation of the switching neural network model. LLM allows to solve classification problems producing sets of intelligible rules capable of achieving an accuracy comparable or superior to that of best machine learning methods. In contrast with algorithms for decision trees that employ a divide-and-conquer approach, methods based on Boolean function synthesis adopt an aggregative policy: at any iteration some patterns belonging to the same output class are clustered to produce an intelligible rule. Suitable heuristic algorithms are employed to generate rules exhibiting the highest covering and the lowest error. The present investigation is aimed at assessing the performance of LLM for the classification of MPM, MTX and BP by combining information from TMs concentrations and CE status. A consecutive cohort of 169 patients (52 MPM, 62 MTX and 55 BP) admitted for diagnosis to two pulmonary departments in Northern Italy from 2009 to 2011was included in the analyses. Concentration of SMRP, CYFRA21-1 and CEA1 was measured in pleural effusion and all patients underwent CE. Accuracy of LLM was compared to standard methods of supervised learning machine, namely: DT, ANN, and KNN. To improve the efficiency of LLM algorithm data were previously discretized using Attribute Driven Incremental Discretization algorithm. Finally, to obtain an unbiased estimate of the LLM performance, the entire dataset was randomly split into a training (70%, n=118) and a test set (30%, n= 51).

**Results:** LLM outperformed all competing methods. Accuracy evaluated on the test set was 74.5% for LLM, 68.6% for DT, 64.7% for ANN and 51.0% for KNN. In more details, the proportion of correctly classified samples by LLM was: 81.8% for MPM, 53.3% for MTX and 85.7% for BP. The corresponding estimates for the competing methods were: DT = 77.3% for MPM , 60.0% for MTX and 64.3% for BP; KNN = 45.5% for MPM, 53.3% for MTX and 57.1% for BP; ANN = 68.2% for MPM, 33.3% for MTX and 92.9% for BP. LLM classification was based on a set of six highly discriminant rules. Interestingly, the rules with the highest covering for MPM, MTX and BP were in a very good agreement with the most recent knowledge about the role of the considered TMs in the biology of mesothelioma. In fact, MPM patients were identified by high values of SMRP, a specific marker for mesothelioma, a high value of CYFRA 21-1, a non-specific TM found elevated in several malignancies, but low levels of CEA1, which was reported to be specific for pleural MTX. The corresponding covering was 86.7%. Accordingly, the best rule for MTX identification was based on high values of CEA1 (covering 55.3%), and the best rule for BP was based on low values of both SMRP and CYFRA 21-1 and a negative test result for CE (covering 92.7%). The present investigation indicates that LLM is a flexible and powerful tool for the differential classification of malignant mesothelioma. Further studies on larger cohorts are needed in order to obtain stable and reproducible rules for MPM classification.
**Contact email:** stefano.parodi@ieiit.cnr.it

**Protein families comparisons using repeatome-based profiling**

Genovese L.M.(1) Geraci F.(1) Pellegrini M.(1)

*(1) Istituto di Informatica e Telematica del CNR - Pisa*

**Motivation:** Protein architectures form a complex multilayered hierarchy. The primary linear sequence of amino acids residues arranges itself in 3-dimensional space so to form local structures (secondary and super-secondary structures, and extends up to fully functional folded proteins (tertiary and quaternary structures) with their functional characterization. For a majority of proteins only the primary AA sequence is known reliably, while the most valuable characterization in structural ond/or functional terms is routinely attained with the use of prediction tools that try to find matching homologous proteins within databases of validated structural/functional hierarchies (e.g. SCOP, CATH). As remarked in [Simossis and Heringa 2006], at the moment no systematic analysis has been done on how incorporating repetitive features of the primary sequence might help in improving alignment quality of homologous proteins (and protein families) matching. Here we report initial findings in the direction of repeatome-based profiling of protein families with the aim of improving current alignment/matching technologies and classification methods.
**Methods:** PTRStalker [Pellegrini et al. 2012] is an algorithm designed to detect Fuzzy TR (FTR) in protein sequences (20AA alphabet). Using PTRStalker as a black-box we compute a FTR-profile for a protein P by (a) detect the set FTR(P) of FTR in P (b) compute mean of the FTR over ten random shuffling of P (c) remove from FTR(P) all TR of length smaller than the mean computed at (b). The statistically filtered FTR are then turned into a vector of features that include the length of P, the length of all FTR after the statistical filtering in the order of appearance along the protein, and the features of the background random shuffling FTR distribution (mean and max values). This FTR-descriptor for the protein P can be used in different ways. In the next section we report good performance of this descriptor in a direct characterization of structured and unstructured proteins. Also we have used this descriptor together with the Euclidean metric to perform unsupervised learning (clustering) of SCOP protein families obtaining highly homogeneous clusters. As next step we plan to apply this new protein descriptor in conjunction with other descriptors (primary sequence, secondary structure, etc.) in the framework of Chung and Yona 2004 in order to improve the prediction of distant homologies among protein families by augmenting family profiles with FTR descriptors.
**Results:** We have tested three validated data sets. The first data set (DS1) is a collection of 92 sequences covering 54037 bps from [Walsh et al. 2012] corresponding to 18725 bps validated secondary structures (mostly solenoids). This benchmark is intended to measure the capability of PTRStalker in detecting existing secondary structures. After statistical filtering PTRStalker returns 95 Fuzzy Tandem Repeats of which 67 overlap known SS in DS1. In terms of base counts the reported FTR cover 17544 bases of which 11594 cover known SS in DS1 (recall: 0.62, precision: 0.66). The second data set (DS2) is a collection of 105 proteins from the database DisProt classified as 100% disordered. The rationale of the experiment is that disordered protein should be relatively free of long tandem repeats. We split the data in three groups of 35 protein each, of length range [45-110][111-208], and [>209] and in each class we tested the hypothesis that the disordered proteins of that length class have FTR statistically equivalent from that of randomly shuffled proteins. The Wilcoxon signed rank test on the length of the longest FTR found in each of the three classes are respectively: 0.199, 0.135 and 0.008. This result implies that such unstructured proteins are indeed free

of significant FTR at least up to length 200. This measure is in line with the findings of experiment on DS1. The third data set (DS3) is composed of 507 non redundant proteins in 6 SCOP superfamilies from [Paccanaro et al. 2006] selected as a challenge for clustering algorithm. Within any superfamily protein pairs have high sequence divergence, but high structural similarity. For each protein we build (see section methods) a descriptor or its FTR profile, including also the background as measured by random shuffling the proteins sequences. Clustering made with the tool Amica [Geraci et al. 2008] using Euclidean distance and a target of 30 clusters has produced 26 highly homogeneous clusters at the superfamily level (with hypergeometric test p-value < 0.004, with BHY FDR adjustment for multiple testing.) covering 90% of the input set. This experiment implies that FTR characterization of proteins is a promising new feature that can be used in novel clustering and classification tasks

**Contact email:** marco.pellegrini@iit.cnr.it

**Supplementary information:**Simossis, V.A. and Heringa, J. (2006). Local structure prediction of proteins. In: Computational Methods for Protein Structure Prediction and Modeling (Xu, Y., Xu, D., Liang J, Eds.), Springer-Verlag, GmbH. Chung R, Yona G. (2004) Protein family comparison using statistical models and predicted structural information. BMC Bioinformatics. Nov 25;5:183. Walsh, Ian and Sirocco, Francesco G. and Minervini, Giovanni and Di Domenico, Tomás and Ferrari, Carlo and Tosatto, Silvio C.E. (2012). RAPHAEL: Recognition, periodicity and insertion assignment of solenoid protein structures. Bioinformatics. 10.1093/bioinformatics/bts550 M. Pellegrini, and M. Elena Renda and A. Vecchio. Ab Initio Detection of Fuzzy Amino Acid Tandem Repeats in Protein Sequences. BMC Bioinformatics 2012, Vol. 13(Suppl 3):S8, doi:10.1186/1471-2105-13-S3-S8. March 2012. Sickmeier M, Hamilton JA, LeGall T, Vacic V, Cortese MS, Tantos A, Szabo B, Tompa P, Chen J, Uversky VN, Obradovic Z, Dunker AK. DisProt: the Database of Disordered Proteins. Nucleic Acids Res. 2007 Jan;35(Database issue):D786-93. A. Paccanaro, J.A. Casbon, M.A.S. Saqi. (2006). Spectral clustering of protein sequences Nucleic acids research 34 (5), 1571-1580 F. Geraci, M. Pellegrini, E. Renda. AMIC@: All MIcroarray Clusterings @ once. Nucleic Acids Research , Vol. 36, Web Server Issue W315~W319, 2008.

**PaPI: the Pseudo Amino acid variant Predictor**

Marini S (1), Limongelli I (2), Bellazzi R (1)

*(1) Department of Industrial and Information Engineering, University of Pavia (2)National Institute of Neurology, C. Mondino Foundation, Pavia*

**Motivation:** Sequenced based disease association studies are becoming a reality. Rare variants with mild and large impact on protein functions will play an important role in this scenario. Therefore, accurate prediction algorithms are needed in order to score genome variants and classify them into benign or being protein-damaging mutations. Existing variant prediction tools like Polyphen2, Sift, Provean and Carol are based on features such as structural and comparative evolutionary considerations and sequence protein homolog similarity. We developed a new coding variant prediction algorithm (PaPI) based on a pseudo amino acid (PseAA) protein sequence discrete model representation. This allowed us to apply an empirical machine learning approach on amino acid sequence pattern alterations. PaPI shows a better accuracy than the above mentioned tools. Polyphen2 and Sift combination was demonstrated to be the best solution for missense variant. Our results suggests that PaPI is able to further increase the accuracy in predict the impact of single nucleotide variations (SNVs) leading to missense mutations and all the other type of genomic coding variants such as insertions, deletions and indels (DIVs).Furthermore, PaPI is also able to predict a sensible percentage of variants that other tools fail the correct prediction, and (theoretically) capable of dealing with INDELS of any length.
**Methods:** PaPI is an ensemble method combining a Random Forest (RF) model with Polyphen2 and Sift. RF applications are widespread in Genomics, from GWAS studies to RNA-protein binding prediction [ref6]. For our RF approach we trained the algorithm on three feature sets: (a) PseAAs, (b) quantitative attributes at whole-protein level, and (c) evolutionary conservation scores such as Gerp++, Siphy and PhyloP. Given a peptide of any length, PseAAs allow to embed patterns of its physico-chemical properties in a fixed number of features, reflecting both compositional and sequential order. PseAAs are characterized by two user-selected parameters: λ, related to minimum sequence length, and w, a weight discriminating composition and positional features. Being our instances mutations, we measured wild and mutated PseAAsand coded their differences as features. We utilized type II PseAA based on hydrophobicity and hydrophilicitycalculated with Pseb, and the RF model was based on Weka.PaPI score is the result of a voting scheme between RF, Polyphen2 and Sift scores that are normalized in order to be comparable. A majority voting scheme is applied when each of the three models provides a prediction. The normalized score of the most confident tool (distance from decisionthreshold) is taken as the final score. The same criteria are applied in case of only two predictions, otherwise only RF score is showed.Training and test set were obtained combining a disease and benign variant set. Disease variant set was extracted from HGMD professional (May 2013 version). Variants were retained for disease annotation (“DM” flagged) and frequency in 1000 genome project (1TGP, MAF < 0.05). Benign variant set was extracted from 1TGP and the Exome Sequencing Project (ESP) retaining for common variants (MAF > 0.06). Resulting variants were annotated by PaPI framework (Java) and only SNV-missense mutation and DIV-in-frame alterations were included in the final set. Because of the difference in size between cases and controls (about 97.000 disease and 27.903 benign variants) training and test set were built on three randomly quasi-balanced sets splitting each of them in 70% and 30%, respectively.The best

combination of λ, w and RF parameters was found via 10-fold cross validation on each training set.

**Results:** PaPI was tested on the three variation sets discussed above. Considering solely the RF model, the resulting area under the curve (AUC) on test sets was 0.899 and the accuracy 0.832 in average, similar to the performances of Sift (AUC=0.899, acc=0.827) and Polyphen2 (AUC=0.915, acc=0.845)on the very same test sets. The final PaPI score obtained by the majority voting scheme increased AUC to 0.926 and accuracy to 0.866.PaPI outperforms Carol as well.Being Carol a combination of Polyphen2 and Sift scores on only SNVs-missense, this suggests that the RF model is complementary to both Polyphen2 and Sift. PaPI is able to deal with RefSeq, Ensembl and GENECODE gene models further increasing the sensitivity of predictions. Interestingly, for SNVs-missense and DIVs-in-frame (not introducing a stop-codon), PaPI prediction failure rate is zero, on the contrary of the other variant prediction tools.

**Contact email:** ivan.limongelli@unipv.it

**Comprehensive large scale assessment of intrinsic protein disorder predictors**

Walsh I (1)*, Giollo M (1,2)*, Di Domenico T (1), Ferrari C (2), Tosatto S (1)

*(1) Department of Biomedical Sciences, University of Padua, Viale G. Colombo 3, 35131 Padova, Italy. (2) Department of Information Engineering, University of Padova, Via Gradenigo 6, 35121 Padova, Italy *Contributed equally*

**Motivation:** High throughput experiments are producing a large amount of data that needs careful annotation. Over the last decade a number of fast tools claiming accurate prediction were developed, but an objective and third party evaluation is often missing. In this work, we analysed the performance of 11 state-of-the-art protein disorder predictors, and compared their results to detect their strengths and biases. Interestingly, these predictors are part of public database like MobiDB, suggesting that a validation of these widely used methods is necessary.
**Methods:** The following fast disorder predictors have been selected in our study: FoldIndex, GlobPlot, IUPred (short and long), Espritz (X-ray, NMR, DisProt), RONN, DisEMBL (456 and HL) and VSL2b. The former four use amino acid poteintials to predict regions with high folding chance, while the latter seven are based on machine learning methods trained on missing residues in the PDB or DisProt databases. All of these tools have been validated on ca. 25.000 UniProt sequences with experimental annotation in MobiDB. Our evaluation is therefore considering a test set with at least one order of magnitude more targets than the predictor's original papers.
**Results:** Our assessment shows that disorder detection is a hard task, and there is room for improvement in the overall prediction quality. In fact, there are classes of proteins, like those related to Biological Adehesion or Signaling, where predictor performance is clearly different than the average accuracy. This is probably due to a bias in the design of the tested methods, which is also clear from the tendency to predict N and C terminal regions as unstructured. All methods typically use PDB structures for training, which are generally disordered at the N and C termini. On the other hand, the typical user is interested in UniProt sequences, which contain the PDB chains and additional residues. In addition, different methods tend to disagree when they predict a protein region as disordered, suggesting that they use very diverse definitions of the problem. For this reason, it is important to develop new guidelines for this problem, and design a continuous assessment of predictors for intrinsical protein disorder.

**An estimated 5% of new protein structures solved today represent a new Pfam family**

Mistry J (1), Kloppmann E (2,3), Rost N (2,3), Punta M (1,4)

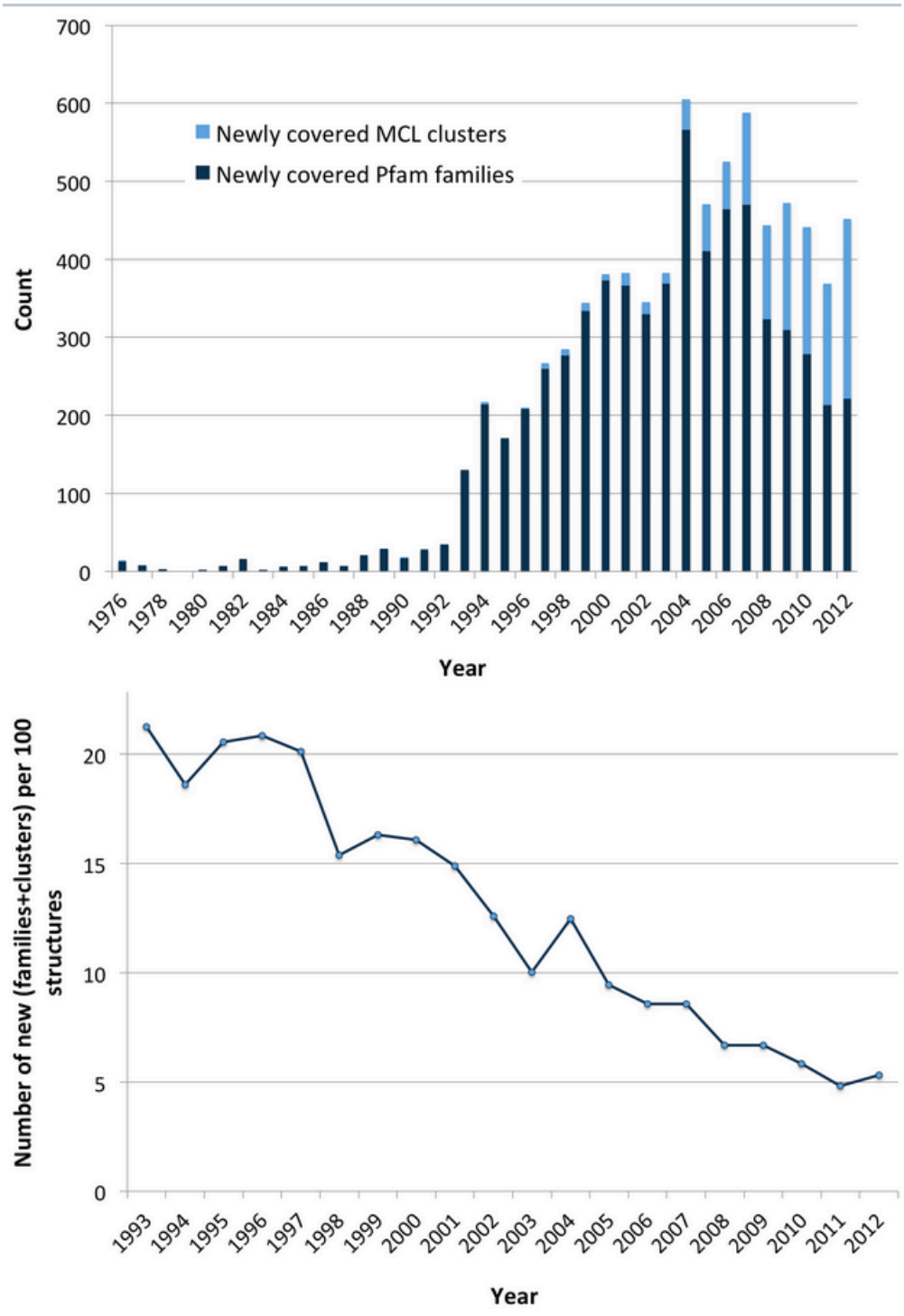
*(1) European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL–EBI), Hinxton, Cambridge, England (2) Department of Bioinformatics and Computational Biology, Technical University Munich, Garching, Germany (3) New York Consortium on Membrane Protein Structure, New York Structural Biology Center, New York, NY, USA (4) Sanger Institute, Hinxton, Cambridge, England*

**Motivation:** High-resolution structural knowledge is key to understanding how proteins function at the molecular level. Both the number of structures and the number of chains released in the Protein Data Bank (PDB) have been steadily increasing over time. In 2012, 211% more protein structures and 242% more protein chains were released in the PDB than 10 years before. Given this wealth of new structures, we wanted to investigate how structural coverage of the protein-sequence space has changed over the years.
**Methods:** We based our analysis on a list of PDB protein chains provided by the CATH database team that satisfied the following criteria: i) the experimental method was either NMR, or a different experimental method with resolution <= 4.0 Å; ii) the fraction of non-alpha carbon was >= 0.7; iii) sequence length was >= 40 residues. Additionally, we only considered structures that were released in the period 1976-2012. In total, 196,469 distinct PDB protein chains satisfied the above constraints. We ran these sequences with pfam_scan against the 14,831 profile-HMMs in Pfam 27.0, and used PDB release dates to calculate the number of newly covered Pfam families per year. We clustered the 4,815 PDB sequences for which pfam_scan did not return a significant hit by running phmmer all-vs-all and applying the clustering algorithm MCL using the phmmer E-values as input. Again, we used PDB release dates to calculate the number of newly covered clusters per year.
**Results:** In Fig. 1(a), we show sequence novelty of proteins released in the PDB between 1976 and 2012, where novelty was measured by counting the number of Pfam families that acquired their first structural representative in that year (dark blue). Stacked on top of this is the number of MCL clusters that acquired their first structural representative in the same year (light blue). We see that the number of newly covered Pfam families and MCL clusters has remained relatively stable over the last 5 years, at around 450. This is significantly fewer than achieved in 2004 and 2007 (605 and 588, respectively), and fewer, on average, than observed in the 4 years between 2004 and 2007. In Fig. 1(b), we plot the number of newly covered Pfam families and clusters per 100 released structures. We see that, over the last 20 years, we have gone from about 20 to about 5 newly covered families per 100 released structures. Thus, at the current pace, achieving complete structural coverage for families that are in today's Pfam (release 27.0) would take an estimated 15 to 20 years. Why has the novelty of proteins released in the PDB declined over the years? One important reason is that structurally uncharacterized families occupy a relatively small portion of the sequence space (17% of protein regions in Pfam). Also, among the 'structurally covered' families are some that are very large and functionally diverse. Effective structural coverage of these families often requires solving structures for more than one representative. Further, families that lack a structural representative are enriched in domains of unknown function (or DUFs; 36% compared with only 10% among families with a structural representative). Finally, these families are predicted to be enriched in coiled-coil, disordered and transmembrane residues which constitute regions that often make experimental structural characterization of proteins all the more challenging. While there are many good reasons for structural

biologists to focus their efforts on families that are already structurally covered, we identified about 1,000 Pfam families with at least one human member that have no structural representative, are not DUFs, are not multispan membrane proteins and are not particularly enriched in disorder. This large set of structurally uncharacterized human proteins suggests that it may not be time yet to give up the pursuit of a targeted but more comprehensive structural coverage of the protein sequence space.

**Contact email:** mpunta@ebi.ac.uk

**Supplementary information:**This work has been published: PMID:24189229



**Tabhu: Tools for AntiBody Humanization.**

Olimpieri PP(1), Chailyan A(2), Marcatili P(3), Tramontano A(1,4)

*(1) Department of Physics, Sapienza University of Rome, Roma (2) Department of Biochemistry and Molecular Biology (BMB), University of Southern Denmark, Odense M (3) Center for Biological Sequence Analysis, Department of Systems Biology, Technical University of Denmark, Kongens Lyngby*

**Motivation:** Monoclonal antibodies (mAbs) are a prominent class of therapeutic molecules. The high specificity and affinity toward their cognate antigens makes them excellent drug candidates. Historically, mouse mAbs were the first to be adopted in clinical practice. Despite their efficacy they suffer from several deficiencies, such as a high chance of eliciting an immunogenic reaction in humans. Antibody humanisation methods are designed to produce molecules that retain their ability to recognize the antigen, while showing a better safety profile. This can be accomplished by grafting the non-human regions determining the antigen specificity into a suitable human template. Not always the "donor" parts accommodate well in the human environment, leading to changes in the overall antibody three-dimensional structure and loss in antigen binding affinity. To retain the correct conformation and therefore the affinity for the antigen, is usually necessary mutating back to the parental critical residues inside the "acceptor" molecule. Finding the appropriate set of back-mutations for a specific antibody is commonly an hard task involving expensive and time consuming experiments.
**Methods:** proABC is based on a random forest algorithm trained from sequence and sequence-derived features. It takes as input an antibody sequence, processes it to gain information on critical antibody features (Hypervariable loop canonical structures and lengths, heavy and light chain variable domain germline families), and estimates for each residue the probability that it interacts with the cognate antigen. We integrated proABC in Tabhu, a web server specifically designed to help researchers humanizing a non-human antibody. Tabhu provides a set of tools covering any aspects of the humanisation rational cycle. The template selection has been accounted by blasting the input sequence against a database of more than 2500 antibody sequences. Moreover sequence-derived information about the non-human antibody are calculated on the fly to allow the user selecting the most suitable human template. After that, the user can select the region to be grafted and Tabhu can predict the antigen contacting residues and evaluate all the possible back-mutations. After selecting the desired set of back-mutations, the humanised variant is evaluated and its three dimensional model built and tested. Users can perform several cycles of back-mutations and test different humanised variants to select the most satisfying one. Eventually, Tabhu returns as output the nucleotide sequence and the three-dimensional model of the final humanised variant.
**Results:** We developed proABC (prediction of antibody contacts) to identify, on the basis of the antibody sequence alone, which residues of an antibody directly interact with its cognate antigen. The method reaches a recall and specificity as high as 80% and is available as a free and easy to use web server (the proABC web server). proABC has been integrated in the new web server for antibody humanisation Tabhu (Tools for AntiBody Humanisation). TAbHu is a unique system that given an input mouse sequence can: • Propose a list of suitable human templates by blasting the input sequence against a database of antibody sequences. At the same time it calculates sequence-derived features (like canonical structures, germline genes), important for the template selection. • Choose and graft

regions from the input antibody to the selected template. • Predict the paratopes for the input antibody and the humanised variant using a random forest algorithm. • Score and rank all the possible back-mutations by mean of a new in house built scoring algorithm. • Allow the user to select one or more back mutations and evaluate the new humanised variant. • Predict and model the three-dimensional structure humanised variant and perform a structural check to find cavities and bumps generated while back-mutating. • Allow the user to select the expression cell line and output the nucleotide sequence of the final humanised variant adopting the appropriate codon usage.
**Contact email:** anna.tramontano@uniroma1.it

**A new alphabet and substitution matrix to represent and compare RNA secondary structures**

Mattei E(1), Ferrè F(1), Helmer-Citterich M(1)

*(1)Centre for Molecular Bioinformatics, Department of Biology, University of Rome "Tor Vergata", Via della Ricerca Scientifica, 00133, Rome, ITALY*

**Motivation:** Structural information has been demonstrated to be crucial in RNA analysis and functional annotation[1, 2]. Nevertheless, how to include such structural information is still a debated problem[3–8]. Dot-bracket notation is the most common and simple representation for RNA secondary structures, consisting in a three-character alphabet representing paired and unpaired positions. This simple representation stores no information about the structural context of the nucleotide, leading to ambiguity requiring further processing steps to dissolve. Indeed, all unpaired nucleotides are described with dots even if they can be part of a loop, an internal loop or an unstructured region. Furthermore, few changes in this simple representation could lead to a complete different structure that could not be simply identified. Here we present a new encoding for RNA secondary structure called BEAR (Brand nEw Alphabet for RNA) and define a substitution matrix for RNA secondary structure elements (MBR - Matrix of BEAR encoded RNA ). We prove that BEAR and MBR can be used to improve the performances and results of existing procedure for RNA alignment and classification.
**Methods:** Within a string of characters, the BEAR encoding allows to store information about RNA secondary structure. BEAR unambiguously associates information about secondary structure to each nucleotide in a RNA sequence. In the BEAR encoding, different sets of characters are associated to the different RNA basic structures (loop, internal loop, stem, bulge), that is to say every letter carries the information about the length and type of structure it belongs to (e.g. 'a' identify a stem of length one). Differently from the dot-bracket notation described above, the assignment of each nucleotide to the secondary structure element it belongs to allows one to discriminate nucleotides described with the same symbol (a dot or a bracket) but belonging to a different secondary structure element. A BEAR representation of RNA secondary structure is a string, having the same length of the associated RNA sequence but, at the same time, encodes for more structural information than the standard dot-bracket notation. For instance, the BEAR encoding discerns unpaired nucleotides belonging to a loop and to a bulge. Every RNA secondary structure encoded with dot-bracket notation can be easily translated into BEAR encoding in linear time. Exploiting this informative notation, we computed a substitution matrix of secondary structure elements using the Dayhoff approach[9]. In particular, we selected a subset of Rfam[2] multiple sequence alignments of homologous RNA, we folded it and then converted into BEAR encoding. As a result we obtained multiple sequence alignments of BEAR encoded RNAs. Finally, we used the sequence alignments of BEAR characters to compute transition frequencies among different secondary structure elements and their lengths obtaining the MBR substitution matrix. The substitution frequencies in MBR capture the type and amount of structural variation that structurally similar, homologous, and/or functionally related RNAs can tolerate. Therefore, among other applications, MBR rates can be used to align the SSEs of two related RNAs encoded using the BEAR representation, in a similar way in which amino acid or nucleotide substitution matrices are used to align two protein or RNA sequences. For this purpose, we implemented a modified version of the Needleman-Wunsch

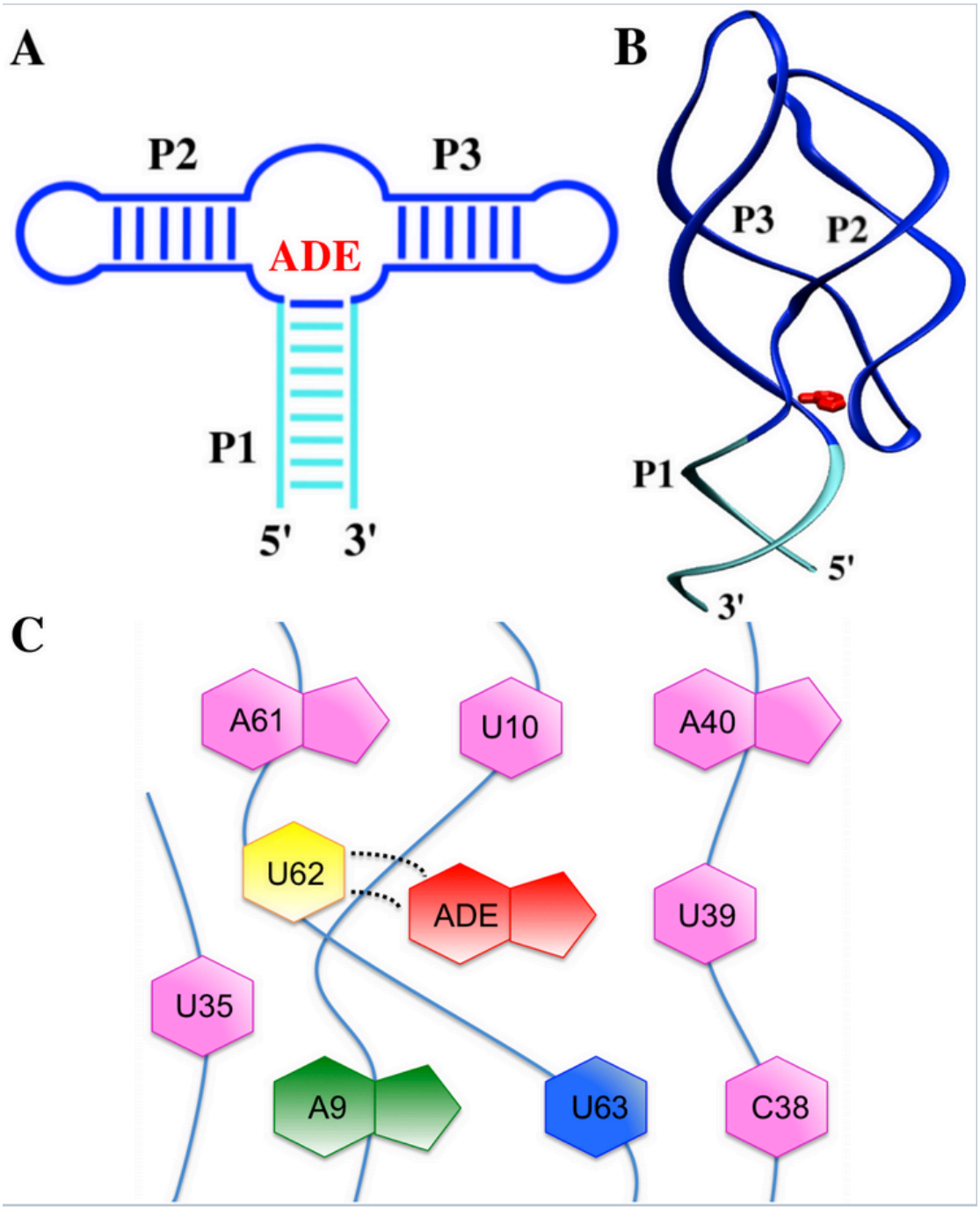
[10] algorithm for the global alignment of RNA BEAR strings guided by the substitution matrix log-odds ratios.
**Results:** The presented method represents a novel approach in describing and comparing RNAs. The structure is still represented as a string but the new alphabet allows storing information describing the structural context of the nucleotide. We used available curated sequence alignments as benchmark and we compared the reconstructed alignment accuracy with that obtained using other popular aligning tools, such as LocARNA[3], Gardenia[11], RNAdistance[12], RNAforester[12] and RNAStrAT[13]. We succeeded in aligning RNA structures in $O(n^2)$, which is lower than all the other tested methods, with comparable alignment accuracy. Additionally, our approach is more accurate when sequence divergence is higher. Hence, the BEAR encoding and the associated MBR matrix can be powerful tools for a better understanding of RNA sequence/structure relationships, evolution and function.
**Contact email:** citterich@uniroma2.it
**Supplementary information:**1. Gardner,P.P., Wilm,A. and Washietl,S. (2005) A benchmark of multiple sequence alignment programs upon structural RNAs. Nucleic Acids Res., 33, 2433–9. 2. Burge,S.W., Daub,J., Eberhardt,R., Tate,J., Barquist,L., Nawrocki,E.P., Eddy,S.R., Gardner,P.P. and Bateman,A. (2013) Rfam 11.0: 10 years of RNA families. Nucleic Acids Res., 41, D226–32. 3. Will,S., Reiche,K., Hofacker,I.L., Stadler,P.F. and Backofen,R. (2007) Inferring noncoding RNA families and classes by means of genome-scale structure-based clustering. PLoS Comput. Biol., 3, e65. 4. Havgaard,J.H., Torarinsson,E. and Gorodkin,J. (2007) Fast pairwise structural RNA alignments by pruning of the dynamical programming matrix. PLoS Comput. Biol., 3, 1896–908. 5. Harmanci,A.O., Sharma,G. and Mathews,D.H. (2007) Efficient pairwise RNA structure prediction using probabilistic alignment constraints in Dynalign. BMC Bioinformatics, 8, 130. 6. Taneda,A. (2010) Multi-objective pairwise RNA sequence alignment. Bioinforma. Oxf. Engl., 26, 2383–90. 7. Notredame,C. and Higgins,D.G. (1996) SAGA: sequence alignment by genetic algorithm. Nucleic Acids Res., 24, 1515–24. 8. Macke,T.J., Ecker,D.J., Gutell,R.R., Gautheret,D., Case,D. a and Sampath,R. (2001) RNAMotif, an RNA secondary structure definition and search algorithm. Nucleic Acids Res., 29, 4724–35. 9. Dayhoff,M., Schwartz,R. and Orcutt,B. (1978) A Model of Evolutionary Change in Proteins. Atlas Protein Seq. Struct., 5, 345–352. 10. Needleman,S.B. and Wunsch,C.D. (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. J. Mol. Biol., 48, 443–53. 11. Blin,G., Denise,A., Dulucq,S., Herrbach,C. and Touzet,H. (2007) Alignments of RNA structures. IEEEACM Trans. Comput. Biol. Bioinforma. IEEE ACM, 7, 309–22. 12. Lorenz,R., Bernhart,S.H., Höner Zu Siederdissen,C., Tafer,H., Flamm,C., Stadler,P.F. and Hofacker,I.L. (2011) ViennaRNA Package 2.0. Algorithms Mol. Biol. AMB, 6, 26. 13. Guignon,V., Chauve,C. and Hamel,S. (2008) RNA StrAT: RNA Structure Analysis Toolkit. 16th Annu. Int. Conf. Ldots.

**Structural and thermodynamics characterization of the ligand-aptamer interactions in the adenine riboswitch**

Di Palma F(1), Colizzi F(1), Bussi G(1)

*(1) Physics Area, Scuola Internazionale Superiore di Studi Avanzati, Trieste*

**Motivation:** Riboswitches are composed of an aptamer domain that binds the effector ligand and an expression platform that transduces the ligand-induced conformational "switch" into a regulation of gene expression. One among the most studied riboswitches is the adenine sensing one (A-riboswitch) cis-regulating the translation initiation of add gene in Vibrio vulnificus. The ligand-bound structure of its aptamer (PDB ID 1Y26) is a three-way junction composed of three stems (P1, P2, P3) with the ligand completely encapsulated into the structure (see figure). The structural mechanism regulating the conformational changes between the ON- and the OFF-state in the A-riboswitch, upon ligand binding, mostly remains to be elucidated. It has been proposed that the P1 stem, the terminal helix of the aptamer, is stabilized by the ligand, and this could be a common feature in many riboswitch classes. The role of the ligand in the structural organization of the aptamer has been investigated with in vitro and in silico techniques. A quantitative estimation of the energetic contributions associated to ligand binding, in particular regarding the role of direct P1-ligand interactions, is however still lacking. In this context, state-of-the-art free-energy methods combined with atomistic simulations can bridge the gap, providing an unparalleled perspective on the mechanism and dynamics of the biomolecular process of interest.
**Methods:** In this work, we used steered molecular dynamics (SMD) simulations to study the thermodynamics of the P1 stem formation in the presence (Holo) and in the absence (Apo) of the cognate ligand. We enforced the breaking of the P1 stem base pairs (bp) by steering its root-mean-square deviation (RMSD) from the native structure. The A9-U63 bp, which directly stacks with adenine, was used as a proxy for the P1 stability. The unbinding event of the A9-U63 bp was described as a function of the steered RMSD and the free-energy profiles were computed using the Jarzynski equality (Apo and Holo forms). Then, we employed a recent in-house-developed reweighting scheme to estimate free-energy profiles as function of other, non-steered observables. In this particular case we analyzed the number of hydrogen bonds between the two bases, thus distinguishing the states in which the bp is formed from states where they are not paired anymore.
**Results:** Our non-equilibrium simulations provide measurements of the stability of the A9-U63 bp and quantify the direct ligand-dependent stabilization of the pairing. The results suggest that the bp could spontaneously break in the Apo form, whereas the presence of the ligand, and its pairing with U63, stabilizes the stacked bp. The $\Delta\Delta G$ between the two forms is $-4.4 \pm 2$ kJ/mol (~2 kcal/mol). This value quantifies the thermodynamic stabilization to the formation of the base pair which directly interacts with adenine in the P1 stem. Our results are in nice agreement with thermodynamic data based on dsRNA melting experiments. From a structural point of view, one can consider the P1 stem as a helix, and the ligand, paired with U62, provides an additional bp to the stem. Differently, in the Apo form, U62 acts as a dangling end at the 5' of the helix. Our work provides the quantitative energetic estimates of the P1 stabilization upon ligand binding introducing also the atomistic details to structurally characterize the folding process of the add riboswitch. Moreover, the results we obtained can complement existing experimental data and shade a new light on the currently accepted model for the aptamer folding.
**Contact email:** fdipalma@sissa.it


A
P2
P3
ADE
P1
5'
3'
B
P3
P2
P1
5'
3'
C
A61
U10
A40
U62
ADE
U39
U35
A9
U63
C38

## Special session: Stochastic modeling

**Spatial and Temporal Order Beyond the Deterministic Limit: The Role of Stochastic Fluctuations in Population**

Fanelli D

*University of Florence, Italy*

Investigating the dynamical evolution of an ensemble made of microscopic entities in mutual interaction constitutes a rich and fascinating problem, of paramount importance and cross-disciplinary interest. Molecules, with their chemical properties and distinct diffusive abilities, can be ideally grouped into homogeneous families, whose concentrations vary continuously with position and time, as follows the governing dynamics. From biology to biomedicine, passing through physics and chemistry, the study of population dynamics is often tackled via a simplistic approach: the scrutinized families are assumed to be composed of an infinite collection of constitutive elements. However, single individual effects, stemming from the intimate discreteness of the analyzed medium, can prove crucial by modifying significantly the mean-field predictions, and so opening up the perspective for alternative explanations of a wide gallery of experimental observations. The stochastic component of the microscopic dynamics can in particular induce the emergence of regular macroscopic patterns, both in time and space. These aspects will be discussed with reference to specific case studies and shown to possibly translate in an alternative paradigm for patterns formation in biology. The case of the intracellular calcium oscillations will be in particular considered.

Oral presentation: Stochastic modelling 1

**A combination of transcriptional and microRNA regulation improves the stability of the relative concentrations of target genes**

Riba A (1), Bosia C (2), El Baroudi M (3), Ollino L (1), Caselle M (1)

*(1) Department of Physics and INFN, University of Torino, via Pietro Giuria 1, I-10125, Torino, Italy (2) Human Genetics Foundation (HuGeF), via Nizza 52, I-10126, Torino, Italy (3) National Research Council (CNR), Institute of Informatics and Telematics (IIT) and Institute of Clinical Physiology (IFC), Laboratory for Integrative System Medicine (LISM), Via Giuseppe Moruzzi 1, Pisa, I-56124, Italy*

**Motivation:** The interplay between transcriptional and post-transcriptional regulation attracted much interest in the past few years. As in the purely transcriptional regulatory network, motifs belonging to such mixed layer of interaction have been identified and mathematically characterized. MicroRNAs (miRNAs), small non-coding RNAs which post-transcriptionally regulate gene expression, play a pivotal role in these circuitries. So far the attention was mainly devoted to circuits in which miRNAs have only an auxiliary role. However, several important biological processes are actually controlled by miRNAs which play themselves the role of master regulators. The corresponding network motifs show a remarkable degree of topological enrichment in the mixed regulatory network. A major reason of interest in this type of circuits is the so called “sponge effect”, i.e. the appearance of indirect interactions among targets due to competition for miRNA binding. In data analysis from the Encyclopedia of DNA Elements (ENCODE) project revealed that two distinct classes of miRNA-controlled circuits were particularly enriched in the network. In the first class miRNAs target two interacting genes (which for example can dimerize). MiRNAs belonging to the second class target two transcription factors (TFs) which both regulate the same gene, one as proximal and one as distal regulator. This same topology was found to be over-represented in human glioblastoma combining bioinformatic analysis and expression data. Both these examples suggest a role of miRNAs in ensuring the stability and fine-tuning of the relative concentration of their targets. The topological enrichment is further magnified if one selects those motifs in which the two targets are linked by a transcriptional regulation. The resulting network motif is a FFL in which a miRNA regulates a TF and together with it one or more target (T) genes. In the following we shall denote these circuitries as “miRNA-controlled FeedForward loops” (micFFL).

**Methods:** For the stochastic analysis we introduce the reactions involved in the analized circuits and use them to simulate the models through the Gillespie’s direct algorithm. miRNA interaction with its target is assumed to be a titrative one as suggested by miRISC action while the interaction between transcription factor (TF) and target (T) is modelled as a Hill function. We aren't able to solve exactly the relative Chemical Master Equations and so, to work out the first two moments of the distributions we approximated it through the system size expansion (or linear noise approximation). All this machinery allows us to evaluate the correlations and variances of molecular species and to compare circuits with different topologies. We would want to construct a list of putative micFFLs miRNA-T and TF-T regulatory interactions and two approaches are used. For the miRNA-T side we integrated information obtained from four freely available databases of miRNA-T interactions, chosen so as to have the widest possible spectrum of different prediction strategies: doRiNA, microRNA.org, TargetScan and PITA. For the TF-T side we used two different strategies. In the first one we selected the TFs contained in the JASPAR database and used the corresponding Position Frequency Matrix to construct a search algorithm for transcription

factor binding sites within the target promoter regions. In the second approach we simply used as signatures of TF-T interactions the ChIP-seq results of the ENCODE project.
**Results:** We show that microRNAs are able to fine tune the relative concentration of their targets and this function is particularly effective for micFFL circuits. Using both deterministic and stochastic equations we show that these FFLs are indeed able not only to fine-tune the TF/target ratio to any desired value as a function of the miRNA concentration but also, thanks to the peculiar topology of the circuit, to ensure the stability of this ratio against stochastic fluctuations. These two effects are due to the interplay between the direct transcriptional regulation and the indirect TF/Target interaction due to competition of TF and target for miRNA binding. We then perform a genome wide search of these FFLs in the human regulatory network and show that they are characterized by a very peculiar enrichment pattern. In particular they are strongly enriched in all the situations in which the TF and its target have to be precisely kept at the same concentration notwithstanding the environmental noise. As an example we discuss the FFL involving E2F1 as Transcription Factor, RB1 as target and miR-17 family as master regulator. These FFLs ensure a tight control of the E2F/RB ratio which in turns ensures the stability of the transition from the G0/G1 to the S phase in quiescent cells.
**Contact email:** riba@to.infn.it

Oral presentation: Stochastic modelling 2

**Single-molecule modeling of mRNA decay**

Valleriani A(1), Sin C(1)

*(1) Department of Theory and Bio-Systems, Max Planck Institute of Colloids and Interfaces, Potsdam, Germany*

**Motivation:** The degradation of macromolecules like mRNAs and proteins is an orchestrated process with many participating complexes. Biochemical experiments can reveal the effects of single components of the degradation machinery, thereby providing also useful modeling backbones from a single-molecule perspective. As a prominent example one can take the relatively detailed scheme of mRNA degradation mediated by miRNA, with its sequence of events from the intact to the degraded mRNA. Decay experiments such as those obtained by interrupting transcription or by a pulse-chasing process, instead, can provide important time scales of the degradation process even if they do not address the single-molecule perspective. Recently, we have pointed out that the shape of the experimental decay curve can still unveil some aspects of the complex single-molecule aging and degradation process. In particular, the molecular aging process associated to the decay can be related to the biochemical steps of degradation. This important observation allows us to move beyond the assumption of simple exponential decay and provides tools to test the validity of the qualitative backbone models coming from biochemistry studies.
**Methods:** We model mRNA decay from a single-molecule perspective by assuming that each mRNA molecule jumps stochastically from one biochemical state to another and by mapping this process into a continuous time Markov chain. The advantage of this approach is that it can be easily scaled to include several complex phenomena and that it can be solved analytically. After formulating a few alternative parsimonious models, data from mRNA decay experiments are fitted in order to determine the rates of the transitions between the states of the Markov chain. We apply this procedure to published data about the decay pattern of a known target of miRNA decay.
**Results:** The first striking feature is that the data do not support the standard degradation model proposed in the literature. Nevertheless, based on our analysis we are able to propose an extension of the standard model that is able to faithfully fit the data. From the analysis of the result we are then able to derive a time scale of mRNA maturation. Finally, we believe that our study provides enough material to formulate further experiments to find the missing link in the standard model on miRNA mediated degradation.
**Contact email:** angelo.valleriani@mpikg.mpg.de

Oral presentation: Stochastic modelling 3

**High-Throughput Quantification and Stochastic Modeling of NF- κB Dynamics.**

Zambrano S(1,3) Molina N(2) Bianchi M(1) Agresti A(3)

*(1) Università Vita-Salute San Raffaele, Milan, Italy (2) Synthetic and Systems Biology Center. University of Edinburgh (UK). (3) Ospedale San Raffaele, Genetics and Cell Biology Division, Milan, Italy*

**Motivation:** The members of the NF-κB family of transcription factors (homo/hetero dimers, being p65 the must abundant monomer) control hundreds of genes that play a pivotal role in a variety of cell processes that are strongly involved in different steps of inflammation. Upon inflammatory stimuli (e.g. with TNF-α or LPS), IκB inhibitor proteins that constrain NF-κB in the cytoplasm of resting cells are degraded, and NF-κB relocates into the nucleus where it drives the expression of many genes, including those encoding the IκB inhibitors. This negative feedback loop, as predicted by mathematical models, causes oscillatory dynamics of the nuclear amount of NF-κB. Our key aim is to understand the role of these oscillations. In doing so, we can also gain important clues on the role played by this common kind of regulatory feedbacks in cell regulation. All this requires large amounts of data and the use of mathematical models.
**Methods:** We have measured the time evolution of the nuclear to total ratio GFP-p65 in knock-in mouse embryonic fibroblasts using time-lapse imaging. In order to have an automated an unbiased measure of NF-kB dynamics at single-cell level for hundreds of cells, we have developed a software that is able to track hundreds of cells and measure the evolution of the nuclear to total ratio of NF-kB for each of them for up to 15 hours. In order to have a deeper insight on the system behaviour, we have developed a simple mathematical model of NF-kB dynamics. Our 6-species model adds to the well-known 3-species telegraph model the minimum number of species needed in order to encapsulate the basic regulations and feedbacks present in they system. To explore the system, we use deterministic simulations, stochastic simulations and what we call hybrid simulations, in which only the gene state is modeled stochastically.
**Results:** Using our software, we analyzed NF-κB dynamics in over 2000 cells exposed to different concentrations of TNF-α. We reproduced known features of the NF-κB system, such as the heterogeneity of the response in the cell population upon stimulations and we confirmed that a fraction of the responding cells does not oscillate. Our results then confirm that NF-κB dynamics of is strongly heterogeneous. Our method has allowed us to unveil intriguing new features: a basal amount of nuclear NF-κB is present in unstimulated cells, and at any time a fraction of unstimulated cells shows spontaneous random activation of the NF-κB system. We have found that our simple model of NF-kB dynamics is already able to reproduce many of these phenomena. In particular we find that in the model the stochastic gene activation is responsible for most of the dynamic heterogeneity, something which can be checked by comparing stochastic and hybrid simulations. Furthermore, in our model stochastic gene activation by itself gives rise to spontaneous activation of NF-κB. Our results also suggest that feedbacks can play a key role in order to produce patterns of gene activation that are more complex than those predicted by the telegraph model.
**Contact email:** samuel.zambrano@gmail.com

Oral presentation: Stochastic modelling 4

**Enhancing simulation of chemical reactions at mesoscales**

Caravagna G (1), Cazzaniga P(2), Nobile MS(1) Pescini D(3), Re A(1)

*(1) Department of Informatics, Systems and Communication, University of Milano Bicocca, Milano (2) Department of human and social sciences, University of Bergamo, Bergamo (3) Department of statistics and quantitative methods, University of Milano Bicocca, Milano*

**Motivation:** In the field of simulation of biochemical systems, it is fundamental to provide an accurate description of the temporal evolution of a model of chemical reactions, under some specific hypothesis on the system components. Thus, multiple approaches have been developed and are now widely used. Among these, recent approaches tried to enhance our ability of simulating multiscale systems, that is the class of chemical reaction systems characterized by the compresence of many scales within the involved molecular counts, or within the kinetics of the involved reactions. In these conditions, the occurrence of both fast and slow reactions, together with molecular species in high and low amounts, make simulation particularly costly from a computational perspectives and, unless some non-trivial mathematics is involved, accurate simulation might be prohibitive. Thus, very often, the usual Gillespie-like approaches fail to be used in this context, and specific techniques have to be developed. On the one hand, however, to preserve the accuracy of the simulated dynamics it is necessary to account for the stochasticity within the system. On the other hand, it is possible to speed up the simulations exploiting deterministic approaches, but these fail in describing the stochastic fluctuations related to the molecular species occurring in small quantities within the system. To overcome these limitations of the classic simulation methods, hybrid techniques have been introduced. These algorithms exploit a partitioning of the chemical system into fast and slow reactions (or according to the molecular concentration) thus simulating the different partitions with deterministic and stochastic algorithms, respectively (see, for instance, [Salis, JCP 2005]). In order to provide a more accurate simulation of the system dynamics, and to reduce the computational costs of the existing simulation algorithms, we present here a hybrid algorithm that takes into account an intermediate scale.

**Methods:** In this work we present a novel hybrid algorithm that accounts for the mesoscale; indeed, the biochemical system under investigation is here partitioned in three different sets: slow reactions that are simulated by means of the stochastic simulation algorithm [Gillespie, JCP 1977], intermediate reactions with the Chemical Langevin Equation, and fast reactions with Ordinary Differential Equations. The rationale behind this partition strategy aims to: (1) provide a correct description of the dynamics of molecular species occurring in very low amounts, taking into account the noise and the stochastic fluctuations; (2) give an accurate description of the noisy dynamics of the intermediate partition of the system, by applying a more efficient simulation method; (3) achieve a fast simulation of the dynamics when molecular amounts are high and noise is not fundamental in the system dynamics. Moreover, we provide the proper mathematical formalization of the mesoscale system enriching the Chemical Master Equation given in [Salis, JCP 2005] correctly taking into account the corresponding Langevin process.

**Results:** The validity of our hybrid simulation algorithm has been assessed by comparing the molecular distributions obtained by using the hybrid algorithm presented in [Salis, JCP 2005] with those produced by using our method. Moreover, the performances of our simulation strategy have been analyzed by comparing its running time with the other simulation

algorithms. All these tests have been realized by simulating ad-hoc biochemical systems in which there is a clear separation among the reactions involved.
**Contact email:** dario.pescini@unimib.it

## Posters

**Even numbered** posters are displayed from the beginning of the meeting until after the morning coffee break of the **27th**.

**Odd numbered** posters are displayed from after lunch on the **27th** until the end of the meeting.

Poster: Algorithms for Bioinformatics – AB1

**ASSIST: a local structural comparison tool for protein function recognition**

Caprari S(1), Toti D(2), Viet Hung L(1), Di Stefano M(3), Polticelli F(1,4)

*(1)Department of Sciences, University Roma Tre, Roma (2)Department of Computer and Electric Engineering and Applied Mathematics, University of Salerno, Fisciano (3)Artificial Solutions, Barcelona (4)National Institute of Nuclear Physics, Roma Tre Section, Roma*

**Motivation:** In recent years, structural genomic initiatives were focussed on the determination of the three-dimensional structure of as many "hypothetical" proteins as possible in order to gain insight into their biological function and/or catalytic activity. At the same time, the availability of a significant number of enzymes' three-dimensional structures has allowed to code structure-function relationships in databases such as the Catalytic Site Atlas. In this context, the purpose of the present work was the development of a novel and versatile tool for identifying structural similarity between an input protein and known proteins/enzymes functional sites. The program, named ASSIST (which stands for Active Site Similarity Search Tool), was designed to predict the catalytic activity, if any, of proteins whose function is unknown. ASSIST uses a geometric hashing technique to find the largest subset of similar residues between an input protein and known active sites/structural motifs.
**Methods:** ASSIST is based on a local geometric/chemical comparison algorithm whose core functionality resides in trying to identify, within a input protein whose function is unknown, substructures potentially similar to known catalytic sites. The program searches for sets of similar substructures, or "alignments", between an input protein and a list of known catalytic residues, which can be either functional sites or structural motifs. The execution flow of the system is composed of four main phases: a preliminary phase, where known catalytic sites are retrieved and stored; a preprocessing phase, where the information coming from residues pairs in the input structure is stored in a hash table; an alignment recognition phase, where the data structure created during the preprocessing phase is compared with the known catalytic sites previously acquired; a result display phase, where the alignments found are presented both in tabular form and in graphics form using an embedded JMol window. ASSIST is coded in Java and thus can be executed on any any operating system equipped with a suitable Java Virtual Machine (JVM).
**Results:** Tests on 54 randomly chosen enzymes belonging to the six classes of the Enzyme Commission classification were performed to assess the effectiveness of ASSIST in recognizing the catalytic activity of an enzyme. This analysis revealed that the program is able to correctly predict the catalytic activity of the enzymes analyzed in the vast majority of the cases. Indeed, in 60% of the cases, ASSIST is able to predict the exact catalytic activity of the input protein. Furthermore in an additional 20% of the cases the program identifies a close functional homolog of the input protein. Particularly interesting, is the case of the protein BfR192 which does not share any significant sequence or structural homology with proteins whose function is known. In this case, the exclusively structure-based approach of ASSIST is able to recognize the local structural similarity between BfR192 and the proline-specific aminopeptidase active site even in a different sequence/structure context. ASSIST is freely available for download at the URL http://www.computationalbiology.it/software/ASSISTv1.zip.
**Contact email:** fabio.polticelli@uniroma3.it
**URL:** http://www.computational biology.it/ASSISTv1.zip

Poster: Algorithms for Bioinformatics - AB2

**A family of new algorithms for species delimitation in the incoming "meta –omics"/ "Bioinformatics" Era.**

Cardinali G(1),Corte L(1), Roscini L(1),Duong Vu (2), Robert V(2)

*(1)Department of Pharmaceutical Sciences - University of Perugia, Perugia Italy (2) Centraalbureau voor Schimmelcultures - Utrecht - NL.*

**Motivation:** Biodiversity is regarded as the most precious and probably the most fragile resource of our planet. Its preservation and exploitation call for joint efforts in all scientific and technological areas involved. Microbial Biodiversity , although representing the major reservoir of biological diversity, is even more complex and less understood that that of plants and animals, due to long lasting technical and logical problems. The introduction of Next Generation Sequencing opens an era in which the complexity of the microbial world can be disclosed. However, for the first time ever, data analysis (dry) requires much longer times than those necessary for the microbial and molecular (wet) operations. Moreover, the logical and biological problems related to the lack of a consensus in microbial species definition, creates serious problems in data comparison, sharing and re-use. In this framework, the dispute between phenetics and cladism is a further problem which complicates the scenario, especially because in microbes the extent of cladogenesis and anagenesis is not that postulated and often found in higher eukaryotes, as witnessed by relatively low values of Consistency Index in phylogenetic reconstructions. This stalling situation calls for comparative analyses to establish the effective limits for the use of the two approaches and – possibly- some hybrid situation taking advantages of the best aspects of their most prominent aspects. It is particularly relevant to highlight that cladistic algorithms normally require much more computation intensity than those proposed in phonetics. Aim of the work presented here is to present a set of novel algorithms intended to allow: 1. determination of discontinuities (gaps) between taxa 2. definition of the best representative strain within the species 3. Assessment of the identification quality within the species 4. rapid case-by-case comparison of the cladistic vs. phenetic approach.
**Methods:** Data refers to the D1/D2 domain of the gene encoding for ribosomal 26S rDNA (shortly LSU) and to the ITS complex (ITS1, 5.8S, ITS2) in fungi and particularly in yeast. The choice was motivated by the fact that these two sequences have been largely used in species identification and classification over the last two decades. Data were retrieved from the CBS (Centraalbureau voor Schimmelcultures – Utrecht) database through the BioloMICS software (Bioaware - http://www.bio-aware.com -), or directly from the CBS site (http://www.cbs.knaw.nl/) . Pairwise distance matrixes were calculated with the "optimistic reverse algorithm", using BioloMICS. This algorithms allows to obtain distances without previous global alignment and with automatic correction for the query sequence direction and length. All other calculations were carried out in the free environment R (www.cran.org), producing scripts that will be presented separately.
**Results:** 1. Distribution discontinuities were analyses with the "ContinuityTest" script, showing the cumulative distances of up to 1000 strains from a seed strain. The strains employed in the test were chosen as the most similar to the seed strain after a BLAST search. Results showed that discontinuity, if any, are present at distances of around 4% which is a much larger range than that normally covered by a yeast species. 2. The definition of the most representative strain within a species is crucial for several analyses such as the quality of the strain identification. Three algorithms were tested for this purpose: a. MDA (Minimal Distance Analysis) b. PDA (PCoA based Distance Analysis) c. MeDA (Mean Distance

Analysis) Results showed the outcome of the three approaches in species with different strain distributions and densities. 3. The quality of strain identification at the species level was carried out with the “SpeciesTest” algorithm implementing two approaches: one on the distance from the most representative strain and one on the distances within a network represented by the distances among all strains of the species. 4. Finally, CLAPHE, is the algorithm used to assess the quality of the outcomes generated by the cladistic approach with the increase of the taxonomic range taken into consideration. This algorithm is based on the reiterative calculation of the Consistency Index in strain groups re-sampled according to the distance from a reference strain

**Contact email:** gianluigi.cardinali@unipg.it

Poster: Algorithms for Bioinformatics – AB3

**MONSTER v1.0: a novel procedure to extract and search for RNA non-branching structures**

Fiscon G(1,2), Paci P(1), Colombo T(3), Iannello G(4)

*(1) Institute for System Analysis and Computer Science "Antonio Ruberti" (IASI), CNR, Viale Manzoni 30, 00185 Rome, Italy. (2) Department of Computer, Control and Management Engineering (DIAG), "Sapienza" University of Rome, Viale Ariosto 25 00185, Rome, Italy. (3) Institute for Computing Applications "Mauro Picone" (IAC), CNR, Via dei Taurini 19, 00185 Rome, Italy. (4) Centro integrato di ricerca, Università Campus Bio-medico di Roma, Via Alvaro del Portillo 21 00128, Rome, Italy.*

**Motivation:** RNA structure prediction and structural motifs analysis are challenging tasks in the investigation of RNA functions. A case in point is the study of the recently appreciated long non-coding RNAs (lncRNAs), that are generally defined as non-coding RNAs longer than 200 nucleotides. These novel genes appear often deregulated in cancer and are emerging as new players of transcriptional and post-transcriptional regulation. However, the vast majority of them remain functionally uncharacterized yet. Among others mechanisms of action, specific substructures are likely to be instrumental for functioning for at least that subset of this variegated RNA class acting as scaffolds or guide for proteins. This observation may not be restricted only to lncRNA but be extended to any type of RNAs. Thus, a functionally characterized RNA (reference RNA) can be used to infer the function of others that are functionally unknown (target RNAs), based on shared structural motifs. However, current tools predicting RNA secondary structure and identifying structural similarities are not immediately suitable to deal with lncRNAs because of their long sequence and the lack of multiple alignments.
**Methods:** We propose a novel pipeline (MONSTER v1.0) to detect structural motifs shared between two RNAs, choosing to characterize the folding of an RNA sequence by means of a sequence-structure descriptor (i.e., an array of non-overlapping non-branching structures positioned on the RNA sequence). Specifically, we developed two core modules: (i) nbRSSP_extractor, to assign a unique structure to the reference RNA encoded by a set of non-branching structures; and (ii) SSD_finder, to detect structural motifs that the given target RNA has in common with the reference. Then, we integrated these novel algorithms with already existing software to reach a coherent pipeline able to perform the following two main tasks: the prediction of RNA secondary structures (integration of nbRSSP_extractor and RNALfold) and a search engine for chains of matches (integration of SSD_finder and Structator).
**Results:** To test our procedure, we evaluated the performances of the two core modules (nbRSSP_extractor and SSD_finder), using a dataset of RNAs with known structures and a class of ncRNA families obtained from online freely available database (i.e., RNAstrand v2.0 and Rfam 11.0, respectively). In particular, we tested the performances of our method both in predicting secondary structures of well-characterized RNAs with known structures (rRNAs) and into enabling the identification of the members of the Rfam families. Our structure prediction step (nbRSSP_extractor together with RNALfold) has lower computational costs in comparison with other state-of-the-arts prediction tools (i.e., Rfold and RNAfold). Moreover, our chaining algorithm (SSD finder) shows several strengths, including high specificity, high computational efficiency, and high sensitivity to identify chains of matches. This result has been achieved thanks to the definition of a new score function that takes into account the relative distances as well as the number of the unbranched secondary structures that have been identified as common between a given pair of reference and target RNA.

**Contact email:** fiscon@dis.uniroma1.it

**The cleverSuite Approach for Protein Characterization**

Petr Klus (1), Gian Gaetano Tartaglia (1)

*(1) Centre for Genome Regulation, Barcelona, Catalunia*

**Motivation:** The recent shift towards high-throughput screening is posing new challenges for the interpretation of experimental results. Here we propose the cleverSuite approach for large-scale characterization of protein groups.
**Methods:** The central part of the cleverSuite is the cleverMachine, an algorithm that performs statistics on protein sequences by comparing their physico-chemical propensities. The second element is called cleverClassifier and builds on top of the models generated by the cleverMachine to allow classification of new datasets.
**Results:** We applied the cleverSuite to predict secondary structure properties, solubility, chaperone requirements and RNA-binding abilities. Using cross-validation and independent datasets, the cleverSuite reproduces experimental findings with great accuracy and provides models that can be used for future investigations.
**Contact email:** petr@tartaglialab.com

Poster: Algorithms for Bioinformatics – AB5

**Characterizing evolving protein communities**

Mahmoud H(1,3), Masulli F(1,2,3), Rovetta S(1,3), G. Russo(2)

*(1) DIBRIS, Dipartimento di Informatica, Bioingegneria, Robotica e Ingegneria dei Sistemi, University of Genova (Italy) (2) Sbarro Institute for Cancer Research and Molecular Medicine College of Science and Technology, Temple University, Philadelphia, PA (USA) (3) GNCS - INDAM (Italy)*

**Motivation:** Proteins bind together forming aggregations reflecting biological processes. This process is represented by protein-protein interaction networks. Protein ties may evolve producing protein complexes or mutate due to possible virus infections or during disease progress. Moreover, proteomic data are very complex and live in high dimensional manifolds. Drug discovery targets identification and control of such changes and of their regulation factors with the least possible side effects.
**Methods:** We characterize the protein network as a set of dense linked sub graphs and employ a fuzzy spectral community detection approach in order to infer the core network and identify the evolved interactions, considering possible overlapping structure.
**Results:** We did our experiments on protein-protein interaction networks considering possible nodes removal due to network evolution. The main contribution of this work is employing fuzzy and spectral based approaches for identifying overlapping evolutionary protein communities with an application in sparse interacting protein networks.
**Contact email:** hassan.mahmoud@unige.it
**Supplementary information:**Work partially funded by a grant of the University of Genova. Hassan Mahmoud is a PhD student in Computer Science at DIBRIS, University of Genova.

Poster: Algorithms for Bioinformatics – AB6

**Multi-objective evolutionary gene clustering for GO-based semantic similarity maximization**

Norberto Diaz-Diaz(1), Federico Divina(1), Marco Masseroli(2)

*(1) School of Engineering, Pablo de Olavide University, Carretera Utrera Km 1, 41013 Sevilla, Spain (2) Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano, Piazza Leonardo da Vinci 32, 20133 Milano, Italy*

**Motivation:** Many biomolecular experiments and bioinformatics techniques select sets of genes or gene products according to different biological conditions or discriminative parameter values. Very often it is important to assess the functional coherence and meaning of such sets. Towards this aim, gene and gene product annotations to a variety of terminologies and ontologies are generally used. Several quantitative techniques exist to leverage these annotations in order to quantify the semantic similarity of genes; however very few ones exist to highlight the most representative biological features of a group of genes. GFD (Gene Ontology (GO) based Functional Dissimilarity) was proposed as a novel method to evaluate groups of genes by measuring gene set dissimilarity based on the most relevant (common and specific) functional annotation of the whole set. The GFD demonstrated to perform well when all genes in the considered set have one or more functions in common. Yet, no evaluation of the GFD is available for sets of genes where no function is shared among all genes of the set (e.g., when the gene set is composed of sub-groups of multiple function genes that share one or more functions with the other genes of the same subgroup, but none with the genes of other subgroups). Here we tested the performance of the GFD in this general scenario, which is frequent with sets of many genes. We demonstrated that the GFD can be leveraged to cluster the initial gene set in sub-groups of similar genes by using a multi-objective evolutionary method.
**Methods:** In order to detect clusters of GO-based similar genes, we defined a multi-objective evolutionary algorithm (MOEA), in which the fitness of an individual of the population considered by the algorithm is based on the Pareto dominance concept. To this aim, three objectives have been used: 1) average volume of the clusters, 2) average GFD of the clusters, and 3) number of clusters. In particular, the algorithm aims at maximizing the first objective, while minimizing the other two. Since such objectives are in contrast with each other, the use of a MOEA is justified. An individual of the population considered by the MOEA is represented as a matrix, where columns correspond to genes and rows to clusters. While the number of columns is fixed, the number of rows can vary, depending on the number of clusters that the MOEA solution defines. An element eij of the matrix is set to 1 if gene i belongs to cluster j, otherwise it is set to 0. Thus, the algorithm can also determine the appropriate number of clusters.
**Results:** We applied the developed method to yeast (Saccharomyces cerevisiae S288c) genes annotated to KEGG metabolic pathways, where pathway annotations define the group of genes that take part in each pathway. We focused the GFD on the GO Molecular Function (MF) sub-ontology, which describes the several molecular activities that occur in metabolic pathways. Thus, in the same metabolic pathway there are genes that are involved in the same molecular function and others that take part in other functions, with some genes involved in multiple functions. By evaluating the GFD for the considered genes annotated to GO MFs and leveraging it in the developed MOEA, we identified several homogeneous clusters of genes annotated to the same metabolic pathway. As expected, in some of these clusters/pathways (e.g., ABC transporter) multiple function genes share one molecular

function and one or more subgroups of these genes share another more specific annotation. Additionally, the MOEA also properly identified clusters/pathways (e.g., Fatty acid biosynthesis) that contain one gene with GO MF annotation(s) shared with the other genes of the cluster, which in turn do not share any GO MF annotation among them. Most of all, the MOEA could also correctly find clusters that include functionally similar genes which do not share any function. For instance, the found cluster of genes involved in the "Arachidonic acid metabolism" pathway consists of six genes annotated to different GO MFs, whose similarity according to the GFD is 0.3667 (the lowest the GFD is in the [0-1] range, the most similar the genes in the set are). Furthermore, the MOEA could obtain two subgroups of this cluster, each of them containing three genes sharing the same annotation(s). The first subgroup contains the TGL4, ECM38 and LAP2 Saccharomyces cerevisiae S288c genes, all annotated to the GO MF "hydrolase activity" (GO:0016787). The second subgroup is formed by the GPX1, GPX2 and HYR1 Saccharomyces cerevisiae S288c genes sharing the GO MF "glutathione peroxidase activity" (GO:0004602) annotation. In both cases the shared GO MF was identified by the GFD as the most relevant (common and specific) GO MF annotation of the subgroup Thus, our method demonstrated to be capable of detecting subgroups of genes that share one or more functions, independently from the fact that some function common to all the considered genes exist.

**Contact email:** marco.masseroli@polimi.it

**A full fine-grained parallelization of the multi-objective evolutionary approach NSGA-II for the CUDA architecture**

Pasquale G(1), Mosca E(2), Merelli I(2), Milanesi L(2), Clematis A(1), D'Agostino D(1)

*(1) Institute of Applied Mathematics and Information Technologies, Italian National Council of Research, Section of Genoa, Italy (2) Institute of Biomedical Technologies, Italian National Council of Research, Section of Milan, Italy*

**Motivation:** Many bioinformatics and systems biology challenges, in particular those related to big data analysis, can be formulated as optimization problems and therefore are addressed using heuristics. Beside the typical optimization problem, formulated with respect to a single target, the possibility of optimizing multiple objectives is rapidly becoming more appealing. Recently, several tasks, such as feature classification, gene clustering, SNP selection, pathways analysis, sequence alignment, structure prediction, subnetwork identification, flux balance analysis, are being faced also with multi-objective (MO) optimization approaches. In this context, MO Evolutionary Algorithms (MOEAs) are one of the most widely used classes of methods to solve MO optimization problems. However, these methods can be particularly demanding from the computational point of view and, therefore, effective parallel implementations are needed. This fact, together with the recent wide spreading of powerful and programmable low-cost Graphics Processing Units (GPUs), promoted the production of works in the literature that focus on the parallelization of one or more computational phases among the steps characterizing MOEAs (typically, objective evaluation, fitness assignment, selection, crossover and mutation). One of the first proposed and most famous MOEAs is the Non-dominated Sorting Genetic Algorithm (NSGA), which has been recently improved and renamed as Fast Non-dominated Sorting Genetic Algorithm or NSGA-II. Among the parallel implementations of NSGA-II existing in the literature, none attempts to parallelize the fitness assignment step, i.e., the ranking of non-dominated solutions through the 'fast non-dominated sorting' procedure. This is the core of NSGA-II and, although it is the hardest to parallelize because of its iterative nature, it occupies alone more than 90% of the total computational time needed by the algorithm.
**Methods:** In this work, we present a full fine-grained parallelization of NSGA-II for the CUDA platform. We not only describe in detail new parallel implementations of the most investigated steps, namely, objective evaluation, binary tournament selection, Simulated Binary Crossover (SBX crossover) and polynomial mutation, but also propose different methods to parallelize the fast non-dominated sorting algorithm of NSGA-II. A first contribution indeed is to show useful techniques to map on a Single-Instruction Multiple-Data (SIMD) architecture the typical operators of GAs, through the development of custom kernels and the exploitation of general-purpose parallel libraries. A second contribution is more related to dominance-based MOEAs and NSGA-II in particular, and consists in showing how to accelerate their heaviest computational part.
**Results:** Simulations on benchmark multi-objective optimization problems are executed for increasing population sizes and problem complexity, to show both the convergence properties and the performance of the proposed implementation. Preliminary results on a population of 4096 individuals show that, using an Nvidia GeForce GTX 580 device, our parallelization of NSGA-II's fast non-dominated sorting algorithm provides speedups of about 7-10x against the implementation available on the Kanpur Genetic Algorithms Laboratory website run on an Intel Xeon E5645 CPU, which is based on a highly optimized algorithm for a sequential execution. Moreover, if considering a direct implementation of

the fast non-dominated sorting procedure, the speedups of our parallelization are of the order of 130x.

**Contact email:** daniele.dagostino@ge.imati.cnr.it

Poster: Algorithms for Bioinformatics – AB8

**A new method for RNA secondary structure motifs discovery**

Pietrosanto M(1), Mattei E(1), Helmer-Citterich M(1), Ferrè F(1)

*(1)Centre for Molecular Bioinformatics, Department of Biology, University of Rome "Tor Vergata", Via della Ricerca Scientifica, 00133, Rome, ITALY*

**Motivation:** Functional regions of RNAs are often related to recurrent patterns in sequence and/or secondary structure. These recurrences are generally called motifs, and they have been found to play an important role in RNA folding and interaction with other molecules. Of particular importance is the interaction with RNA-binding proteins (RBP), which is involved in the regulation of a large number of cellular processes. The development of new high throughput techniques to identify RBP targets (such as CLIP-seq [1]) increases the demand for motif finding tools able to reduce the high number of false positives produced by such techniques and to find the correct motifs. Among the available motif-finding tools, the majority focuses on sequence patterns[2, 3], sometimes including secondary structure as additional constraints to improve their performance[4]. Nonetheless, secondary structures may have their motifs too, and these motifs may be concurrent to their sequence counterparts or even be independent from them. During the last years some effort was put into research of structural motifs using advanced methods which require long pipelines and/or high computational efforts [5, 6]. Here we present a novel method for structural motif discovery taking advantage of a new encoding for RNA secondary structure named BEAR (Brand nEw Alphabet for RNAs)[Mattei et al. submitted]. In BEAR, RNA secondary structures are described as a string encoded with a structural alphabet describing each RNA sub-structure (i.e.: loop, interior loop, bulge, stem) and its size. This representation allows us to adapt methods developed for sequence analysis on BEAR-encoded secondary structures.
**Methods:** Secondary structure representation is a bottleneck in RNA structural motif discovery tools. The way in which the RNA is described affects both the algorithmic complexity and its accuracy. The classic dot-bracket notation is generally not suitable, and more complex descriptions, such as tree-based representation are preferred[7]. Nevertheless, working with tree-based representations increases the complexity of the algorithm and thus the computational time. On the other hand, sequence motif discovery tools can exploit the powerful resources offered by string theory. We present an approach for discovering structural motifs that combines the knowledge of sequence motif-finding algorithms to a new string encoding of secondary structure. Thanks to the BEAR encoding, that uses a different set of characters for each RNA secondary structure element and size, the classical dot-bracket notation can be converted into a string of characters storing secondary structure information. This new context-aware representation allows us to use approaches that are more similar to known methods of motif discovery (such as MEME[3]) or to novel methods that exploit suffix trees (such as DRIMust[8]). In particular, the method we present uses a simulated annealing approach (similar to that of GLAM2[9]) that explores the space of candidate motifs in order to find the one with the minimum score. This approach prevents the risk of finding local minima by sometimes admitting unfavourable moves in the scoring space of candidate motifs. This procedure is borrowed from physics, as it is analogous to the crystallisation in a cooling material. In the definition of the scoring function we used MBR (Matrix of BEAR-encoded RNA) to weight the substitution score between different structural elements (e.g. loops, stems, internal loops) or between two elements of the same type but having different size. The substitution scores in MBR capture the type and amount of structural variation that structurally similar, homologous, and/or

functionally related RNAs can tolerate. The scores were computed on alignments of BEAR-encoded RNAs obtained by encoding RNA alignments stored in Rfam database[10] using BEAR.

**Results:** We used Rfam database to test the ability of our method to recall known motifs. In particular, we extracted three datasets, a dataset of sequence motifs with little or no structural conservation, a dataset of structural motifs having little or no sequence conservation, and a dataset of sequence and structural motifs. Preliminary results show a promising ability of our method to identify motifs in each of the three datasets. In other words, our method is able to discriminate between structural or sequence motifs and reduce the number of false positives if compared to other known methods. Moreover, this is done with far less computational complexity with respect to other structural motif-finding algorithms[5, 6].

**Contact email:** fabrizio.ferre@uniroma2.it

**Supplementary information:**1. Uren,P.J., Bahrami-Samani,E., Burns,S.C., Qiao,M., Karginov,F.V., Hodges,E., Hannon,G.J., Sanford,J.R., Penalva,L.O.F. and Smith,A.D. (2012) Site identification in high-throughput RNA-protein interaction data. Bioinforma. Oxf. Engl., 28, 3013–20. 2. Marsan,L. and Sagot,M.-F. (2000) Algorithms for extracting structured motifs using a suffix tree with an application to promoter and regulatory site consensus identification. J. Comput. Biol., 7, 345–362. 3. Bailey,T.L. and Elkan,C. (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. Proc. Int. Conf. Intell. Syst. Mol. Biol. ISMB Int. Conf. Intell. Syst. Mol. Biol., 2, 28–36. 4. Hiller,M., Pudimat,R., Busch,A. and Backofen,R. (2006) Using RNA secondary structures to guide sequence motif finding towards single-stranded regions. Nucleic Acids Res., 34, e117. 5. Heyne,S., Costa,F., Rose,D. and Backofen,R. (2012) GraphClust: alignment-free structural clustering of local RNA secondary structures. Bioinforma. Oxf. Engl., 28, i224–232. 6. Kazan,H., Ray,D., Chan,E.T., Hughes,T.R. and Morris,Q. (2010) RNAcontext: a new method for learning the sequence and structure binding preferences of RNA-binding proteins. PLoS Comput. Biol., 6, e1000832. 7. Le,S.Y., Nussinov,R. and Maizel,J.V. (1989) Tree graphs of RNA secondary structures and their comparisons. Comput. Biomed. Res. Int. J., 22, 461–473. 8. Leibovich,L., Paz,I., Yakhini,Z. and Mandel-Gutfreund,Y. (2013) DRIMust: a web server for discovering rank imbalanced motifs using suffix trees. Nucleic Acids Res., 41, W174–179. 9. Frith,M.C., Saunders,N.F.W., Kobe,B. and Bailey,T.L. (2008) Discovering sequence motifs with arbitrary insertions and deletions. PLoS Comput. Biol., 4, e1000071. 10. Burge,S.W., Daub,J., Eberhardt,R., Tate,J., Barquist,L., Nawrocki,E.P., Eddy,S.R., Gardner,P.P. and Bateman,A. (2013) Rfam 11.0: 10 years of RNA families. Nucleic Acids Res., 41, D226–32.

Poster: Algorithms for Bioinformatics – AB9

**Memory Efficient String Graph Construction**

Previtali M(1), Della Vedova G(1), Pirola Y(1), Rizzi R(1), Bonizzoni P(1)

*(1) Dipartimento di Informatica Sistemistica e Comunicazione, University of Milan-Bicocca, Milan*

**Motivation:** In the last few years the cost of sequencing DNA and RNA has plummeted and, at the same time, the quantity of available data has exploded. Anyway, the price to pay for those advances is that data is available in the form of short (mostly 35-100bp long) sequences, called reads, that are much harder to assemble into a genome or a set of transcripts than that obtained with previous technologies. The strategy to cope with the hurdles posed by NGS data has been twofold: (1) to focus on resequencing projects, where a reference genome is exploited to avoid a de novo assembly step, (2) to design a different class of programs (such as Velvet and Trinity) capable of handling NGS reads but requiring huge amount of data. Two related computational tools, that have been successfully used to align NGS reads against a reference genome, are the Burrows-Wheeler Transform (BWT) and the FM-Index, that couple small RAM consumption with good running time (see for example the read aligners FM, BWA, and Bowtie). Those tools have found a first important application in an assembly method called SGA, that can construct the String Graph (Myers 2005) where each vertex is a read and two reads are connected if they are maximally overlapping (therefore the main difference between string graph and de Bruijn graphs is that in the latter graph vertices are kmers instead of reads). Although SGA copes with the problems that arise from the large number of reads we have to manage, it forces the user to store the whole FM-Index in main memory, therefore it can be run only on a high-end machine – notice that BWT-based aligners have to keep the FM-index of the reference genome in main memory, while SGA keeps the much larger FM-index of all reads in memory. Moreover datasets are forecast to grow at least as quickly as hardware capabilites.
**Methods:** In this abstract we propose a method to lessen the memory usage of SGA, by exploiting an external memory algorithm to compute the overlaps between reads without requiring the BWT to be held in big chunks in main memory, and accessing to it sequentially a number of times that is linear with the maximum length of the reads in our set. Our method follows the approach of (Bauer, Cox and Rosone 2011) to compute in external memory the BWT of a set of reads and the approach of (Lam et al. 2009) to compute a bidirectional BWT that is necessary to compute the overlap between two reads. The main notion of our method is that of Q-interval, that corresponds to the set of (sorted) suffixes of the input reads that begins with the substring Q – this idea has been introduced in (Cox, Jakobi, Rosone, and Schulz-Trieglaff 2012). The main contributions of our approach are (1) an incremental (i.e. Q is extended by a character at a time) and disk-based algorithm to compute all Q-intervals shared by at least two reads, and (2) and efficient test to determine if Q is an overlap between two reads that has to be represented by an edge in the string graph (i.e. if it is a maximal overlap between those reads). Moreover, we extend multiple Q-intervals in parallel and our disk access strategy does not require to compute the FM-index.
**Results:** The memory requirements of our method are really small. In particular it needs to store in main memory only the string graph and an arbitrarily small (and user-defined) portion of the BWT. On the other hand, relying on external memory (although efficiently) requires an analysis of the I/O time. Given that we store both BWT and Q-intervals on external memory the I/O time will be related to them. We only need to read from disk l

times the BWT, where l is the maximum length of a read, and we need to store on disk every Q-interval with at least two suffixes.
**Contact email:** m.previtali3@campus.unimib.it

Poster: Algorithms for Bioinformatics – AB10

**Reconstructing tree-like cancer progression models with a probabilistic causation-based shrinkage estimator**

Loes Olde Loohuis (1), Giulio Caravagna (2), Alex Graudenzi (2), Daniele Ramazzotti (2), Giancarlo Mauri (2), Marco Antoniotti (2), Bud Mishra (3)

*(1) Department of Computer Science City University New York, The Graduate Center New York, USA (2) Dipartimento di Informatica Sistemistica e Comunicazione Università degli Studi Milano-Bicocca Milano, Italy (3) Courant Institute of Mathematical Sciences New York University New York, USA*

**Motivation:** Cancer progression can be modelled in terms of a sequence of discrete steps, where the tumor acquires certain distinct properties at each state. Different progression sequences are possible, but some are more common than others, and not every order is viable. In the last two decades, many specific genes and genetic mechanisms that are involved in different types of cancer have been identified, but, unfortunately, the causal and temporal relations among the genetic events driving cancer progression remain largely elusive. The main reason for this state of affairs is that information revealed in the data is usually obtained only at one (or a few) points in time, rather than over the course of the disease. Extracting this dynamic information from the available cross-sectional data is challenging, and the combination of mathematical, statistical and computational techniques is needed. The state-of-the-art techniques to reconstruct tree models of progression for accumulative processes such as cancer, seek to estimate causation by combining correlation and a frequentist notion of temporal priority. In this work we define a novel theoretical framework to reconstruct such models based on the probabilistic notion of causation defined by Suppes, which differ fundamentally from that based on correlation.
**Methods:** We consider a general reconstruction setting complicated by the presence of noise in the data, owing to the intrinsic variability of biological processes as well as experimental or measurement errors. To gain immunity to noise in the reconstruction performance we use a shrinkage estimator. The set-up of the reconstruction problem is as follows. Assuming that we have a set G of n mutations (events, in probabilistic terminology) and m samples, we can represent a cross-sectional dataset as an m x n binary matrix. In this matrix, an entry (k,l) = 1 if the mutation l was observed in sample k, and 0 otherwise. The problem we solve is to extract a set of directed edges E yielding a progression tree T from this matrix. More precisely, we seek to reconstruct a rooted tree that satisfies: (i) each node has at most one incoming edge, (ii) the root has no incoming edges (iii) there are no cycles. In particular, we state that a causes b if it occurs more frequently and if its presence raises the probability of observing b the most considering all the events in the model.
**Results:** We made substantial use of synthetic data to evaluate the performance of our method. We included a form of noise in generating the datasets, in order to account for (i) the realistic presence of biological noise (such as the one provided by bystander mutations, genetic heterogeneity, etc.) and (ii) experimental errors. The noise denotes the probability that any event assumes a random value (with uniform probability), after sampling from the generating topology. On synthetic data, we show that our approach outperforms the state-of-the-art techniques and, we show that our method is efficient even with a relatively low number of samples and its performance quickly converges to its asymptote as the number of samples increases. Our analysis suggests the applicability of the method even to small datasets of real patients.
**Contact email:** daniele.ramazzotti@disco.unimib.it

Poster: Algorithms for Bioinformatics – AB11

**A graphical platform for the identification of differentially expressed genes**

Russo F(1), Angelini C(1)

*(1) Istituto per le Applicazioni del Calcolo "Mauro Picone", Consiglio Nazionale delle Ricerche (CNR), Napoli*

**Motivation:** We present an open source graphical platform to analyse RNA-Seq data in order to identify differentially expressed genes across multiple biological conditions in RNA-Seq experiments. This graphical platform is useful for those that are not expert shell-users and for those that are expert as well, since it reduces the analysis time drastically.
**Methods:** The platform is written in R and it is open source. It includes several methods, such as: DESeq, DESeq2, NOISeq, edgeR, baySeq. Moreover, there multiple normalization procedures available to the user and a large number of functions to explore the data before and after the application of statistical hypothesis tests.
**Results:** We show the functionalities of this platform by means of several results coming from real data case studies.
**Contact email:** f.russo@na.iac.cnr.it

**A group lasso-based Knowledge Driven Variable Selection for the identification of small relevant functional groups**

Barbieri M(1), Squillario M(1), Zycinski G(1) and Barla A(1)

*(1) Department of Informatics, Bioengineering, Robotics and Systems Engineering, University of Genova, Genova*

**Motivation:** Interpretability of high-throughput data analysis results is often a difficult task. Usually, methods for the identification of relevant molecular variables provide just a ranked list of discriminant features. Therefore, to better capture which gene modules, groups or pathways are involved in the condition of interest (e.g., diseased vs control, treated vs not-treated), a subsequent enrichment procedure is required. Several approaches and tools have been proposed to tackle this issue, all using sources of biological knowledge stored in public repositories such as the Gene Ontology, KEGG or Reactome. Independently of the implementation, enrichment tools use the biological knowledge a posteriori to functionally characterize the discriminant list of molecular variables selected in the preliminary phase of feature selection. In this context, the KDVS method was recently proposed to overcome the drawback of using the biological knowledge a posteriori, when the feature selection has already been performed. The main idea is to apply the feature selection separately on each functional group of variables defined by the chosen source of prior knowledge. The current implementation of KDVS uses Gene Ontology as source of biological knowledge and a sparse regularization method, namely l1l2, to identify the relevant variables. The Gene Ontology is composed of three domains: Biological Process (BP), Molecular Function (MF) and Cellular Component (CC). Each domain is a Directed Acyclic Graph and can be viewed as a tree-like hierarchical structure, with the root node representing the most general function or process in the domain and with leaves nodes corresponding to the most specific functions or processes. KDVS selects as relevant (or enriched) terms, those terms that are associated with a high prediction accuracy.
**Methods:** KDVS applies l1l2 to larger terms, associating them with both a list of relevant molecular variables and with a estimated prediction accuracy. For smaller terms, the number of variables belonging to that term is usually similar to the sample size and KDVS skips the variable selection step and just evaluates the prediction of a linear regression model (Ordinary Least Squares). The main disadvantage, in this case, is that each small and very specific term is considered independently to the other terms. Therefore, KDVS usually does not select many terms of this type, because such specific terms, considered independently, are not able to discriminate between cases and control in multifactorial and complex diseases such as tumors or neurodegenerative diseases. In this work, we propose a variation to the original KDVS implementation that considers small GO terms altogether using group lasso. In this way, small but relevant terms that would have been discarded if considered individually, may be selected by the algorithm while preserving their independency.
**Results:** We apply the modified version of KDVS to a gene expression dataset of controls and patients affected by Parkinson's Disease (PD). The data are publicly available on the Gene Expression Omnibus (GEO, GSE6613) and consist of 33 cases and 22 controls, each associated to 22283 probesets. We focus only on smaller terms and compare the results of the original and modified KDVS. Moreover, we also validate the results considering, as benchmark, a group of BP terms already known to be associated with PD.
**Contact email:** margherita.squillario@unige.it

Poster: Algorithms for Bioinformatics – AB13

**JAVA-GOEA: GO enrichment analysis pipeline for RNA-seq data**

Tripathi KP (1), Cassandra R (1), Ambrosone A (2), Tortiglione C (2), Tino A (2), Guarracino MR (1)

*(1) Laboratory for Genomics, Transcriptomics and Proteomics, High Performance Computing and Networking Institute, National Research Council, Via P. Castellino, 111, Naples, Italy; (2) NanoBioMolecular Group, Istituto di Cibernetica "Eduardo Caianiello" CNR via Campi Flegrei, 34 80014 Pozzuoli Napoli*

**Motivation:** RNA-Seq is a new tool which utilizes the maximum power of high-throughput sequencing to measure RNA transcript counts at an extraordinary accuracy. It provides quantitative means of exploring the transcriptome of an organism of interest. Even if an organism lacks the referenced genome, de novo transcriptome assembly and differential expression can be performed. There is numerous software relevant to computational reconstruction of transcript structures and quantification [1]. However, interpreting this extremely large data into biological knowledge is a big problem, and biologist-friendly tools to analyze them are poorly lacking. There is a need for automated pipeline, which effectively translates RNA-seq results into biological interpretations such as gene or protein annotation for assembled reads and GO enrichment analysis. Previously researchers published methodologies [5,6], which performs the GO term, enzyme and domain annotations on transcriptome, but these analyses are not comprehensive as they do not include annotations for pathways such as KEGG, BioCarta as well as they also do not provide any information on protein-protein interactions.
**Methods:** Presently, we are working on the development of Java pipeline to annotate an individual differentially expressed transcript, and carry out GO enrichment analysis of expression profiles, under the different treatment condition for organisms, which lacks the referenced genome. In this pipeline, we are embedding Blast, QuickGo and DAVID tools (Database for Annotation, Visualization and Integrated Discovery) for gene ontology annotation and enrichment analysis. Our pipeline maps the assembled transcripts to referenced genome as well as Swiss-Prot/Uni-Prot protein id. Gene ontology annotations are carried out by QuickGo [4] and DAVID tools [2,3].
**Results:** The pipeline offers a report on statistical analysis of GO enrichment. It enables a biologist to identify enriched biological themes, particularly GO terms related to biological process, molecular functions and cellular locations. Utilization of DAVID tools in our pipeline allow to map annotated transcript on KEGG, Panther pathways. It also provide information about the protein-protein interactions for the annotated transcripts. Through our pipeline, we are providing an automated protocol to cluster differentially expressed transcripts based on functional annotation. All these utilities in our pipeline, deliver a platform for the biologist, to understand the homologous RNA-Seq data in a biological sense and in a straight forward way.
**Contact email:** kumpar@na.icar.cnr.it
**Supplementary information:**1) Nature Methods 2013; 10, 1177–1184 2) Nature Protoc. 2009;4(1):44-57 3) Nucleic Acids Res. 2009;37(1):1-13 4) Bioinformatics. 2009; 25(22):3045-6 5) BMC Genomics 2012, 13(Suppl 7):S9

Poster: Algorithms for Bioinformatics – AB14

**CONSRANK: A WEB TOOL FOR THE ANALYSIS AND RANKING OF DOCKING MODELS**

Chermak E(1), Petta A(2), Serra L(2), Vangone A(3), Scarano V(2), Cavallo L (1,3), Oliva R (4)

*(1) KAUST Catalysis Center, King Abdullah University of Science and Technology, Saudi Arabia. (2) Dipartimento di Informatica, Università di Salerno, Italy (3) Dept. of Chemistry and Biology, University of Salerno, Italy (4) Dept. of Applied Sciences, University "Parthenope" of Naples, Italy.*

**Motivation:** Correctly scoring docking models to rank native-like conformations before the incorrect ones is still an open problem, which is also object of assessment in the Critical Assessment of PRedicted Interactions (CAPRI). We recently proposed a novel approach to the scoring and ranking of docking models, CONSRANK, which is based on the conservation (or frequency) of the inter-residue contacts in the docking decoys ensemble. CONSRANK is a consensus-based algorithm, which ranks models based on their ability to match the most conserved contacts in the ensemble they belong to. Application of CONSRANK to over 100 targets from the three docking decoys benchmarks (DOCKGROUND, RosettaDock and CAPRI) showed a very good performance, both in terms of native-like solutions ranked in the top positions, and the Area Under the ROC Curve. Here we introduce a web tool for CONSRANK, to make it easily available to the scientific community.

**Methods:** 0

**Results:** The CONSRANK server output includes: i) a list of observed inter-residue contacts in the models ensemble with relative conservation rate; ii) values of parameters reflecting overall conservation of inter-residue contacts in the models ensemble; iii) the ranking of the models based on the CONSRANK normalized score. Furthermore, CONSRANK gives as output a "consensus map", i.e. an intermolecular contact map where the conservation of inter-residue contacts is reported on a gray scale. This is an interactive map that can be zoomed and navigated to visualize the identity of the residue pairs corresponding to a given contact and its conservation rate. An interactive 3D representation of the consensus map, where the third dimension is given by the conservation rate of each inter-residue contact is also provided. By clicking on a specific model, further analyses will be performed on it and 2D and 3D inter-molecular contact maps will be generated and visualized superimposed to the consensus ones.

**Contact email:** romina.oliva@uniparthenope.it

**hLGDB: a database of human lysosomal genes and their regulation**

Brozzi A (1), Urbanelli L (1), Germain PL (2), Magini A (1), Emiliani C (1)

*(1) Department of Experimental Medicine and Biochemical Sciences, University of Perugia, Via del Giochetto, 06123 Perugia, Italy. (2) Scuola Europea di Medicina Molecolare (SEMM), SEMM at Istituto Europeo di Oncologia Via Adamello, 16 - 20139 Milano*

**Motivation:** Lysosomes are cytoplasmic organelles present in almost all eukaryotic cells, which play a fundamental role in key aspects of cellular homeostasis such as membrane repair, autophagy, endocitosis and protein metabolism. The characterization of the genes and enzymes constituting the lysosome represents a central issue to be addressed toward a better understanding of the biology of this organelle. In humans, mutations that cause lysosomal enzyme deficiencies result in >50 different disorders and severe pathologies. So far, many experimental efforts using different methodologies have been carried out to identity lysosomal genes.
**Methods:** The Human Lysosome Gene Database (hLGDB) is the first resource that provides a comprehensive and accessible census of the human genes belonging to the lysosomal system. This database was developed by collecting and annotating gene lists from many different sources. References to the studies that have identified each gene are provided together with cross databases gene related information. Special attention has been given to the regulation of the genes through microRNAs and the transcription factor EB. The hLGDB can be easily queried to retrieve, combine and analyze information on different lists of lysosomal genes and their regulation by microRNA (binding sites predicted by five different algorithms). The hLGDB is an open access dynamic project that will permit in the future to collapse in a unique publicly accessible resource all the available biological information about lysosome genes and their regulation. hLGDB is a MySQL 5.0.95 database (constructed in the fourth normal form, some redundancy being kept to increase retrieval performance), and the interface is built in PHP.
**Results:** hLGDB currently contains 435 genes. There are 16 sources of information divided in four main categories: Proteomic Studies, Databases, Reviews and System Biology Approaches. Each gene has been associated to its Official HGNC Gene Symbol and to its Entrez Gene ID (mappings were based on data provided by Entrez with a date stamp from the source of 7 March 2012). The gene transcripts associated to each gene are annotated accordingly to NCBI RefSeq or GenBank (release 57). miRNA target predictions were extracted from the tables downloaded from the websites of the different algorithms used to predict the binding between miRNA and gene transcripts. We paid special attention on balancing predictions, which were as follows: (i) more suitable to look for confirmatory evidence (TargetScanS); (ii) more suitable to identify any possible target for a particular miRNA, to form the basis for in vitro or in vivo experiments (picTar four-way and five-way); (iii) more suitable to find in silico evidence for the interaction between a miRNA and a gene of a certain family or function (PITA, miRanda). To increase miRNA-target mRNA information, experimentally verified miRNA targets from miRTarBase were also reported. Coordinated Lysosomal Expression and Regulation (CLEAR) is a nucleotide motif (GTCACGTGAC) found to be highly enriched in the promoter set of lysosomal genes. We mapped this motif on both strands on the human genome (hg19) by means of fuzznuc utility of the EMBOSS package allowing one single mismatch. The binding sites of the TFEB come from a Chip-seq experiment carried out on HeLa cell lines. Researchers may benefit from

hLGDB because they have in a single reference to the broadest compendium of lysosomal gene lists. They can search for miRNA targets combining up to six different methods. Results of miRNA targets may be directly compared with other transcriptional regulation elements such as the distance from the TSS of TFEB binding site or the distance to a CLEAR sequence to identify common features of regulation.

**Contact email:** alessandro.brozzi@gmail.com

**URL:** http://lysosome.unipg.it

Poster: Biological Databases – BD2

**Integration of available multiple annotation data and detection of new annotations**

Canakoglu A, Masseroli M

*Dipartimento di Elettronica Informazione e Bioingegneria, Politecnico di Milano, Milano, Piazza Leonardo da Vinci 32, 20133 Milano, Italy*

**Motivation:** With the improvement in the biomolecular engineering and in the sequencing techniques, the number available biomolecular interaction data and the annotated data between the different biomolecular entities, such as DNA sequence, gene, protein and transcript, and their biomedical features have been increasing. Such large datasets need to be curated and analyzed either by domain experts or automated curation algorithms. Also to improve the efficacy of the information, data colltected from different sources should be integrated together. Although some of the annotation, curated by domain experts, are available in different sources sparsely, therefore the integration is a paramount tool. With data available nowadays in the different heterogeneous databases, the annotations can be identified, by using the overlapping relationships between the available data. This can be realized by combining the data across different data sources and also by using database cross reference information.
**Methods:** In Politecnico di Milano University, we have integrated data from several heterogeneous biomolecular annotations data sources with data warehousing approach, and we created the Genomic and Proteomic Data Warehouse (GPDW). After the data integration from various data sources with quality controls, we execute our inference method which is a transitive relationship between various biomolecular entities and biomedical features. As a base point, we are using the gene and their encoded proteins with the other features. Although this method may seem very basic, it is effective and used for the other domains but yet it has not been for the integrated annotation data in biomolecular field yet. By combining the data from multiple data sources and their cross database identification information, we increase our method's expressive power and the quality of our inferred annotations. By using this method, we detect new annotation and fill the missing annotations in the databases.
**Results:** Currently, the GPDW integrates data from several well-known databases carefully selected, including Entrez Gene, UniProt, IntAct, MINT, Expasy Enzyme, BioCyc, KEGG, Reactome, GO, GOA and OMIM. At the time of writing in the Genomic and Proteomic Data Warehouse (GPDW), there are 9,537,645 genes of 9,631 organisms, 38,960,202 proteins of 338,004 species and 30,784,393 associations between gene and their encoded proteins. Our method is based on mainly genes and their encoded proteins, and we inferred annotations between gene or protein and the other features defined in the data warehouse, such as Gene Ontology. For the validation of our method, we found a new annotation between the human gene IGF2 (Insulin-like growth factor 2 - somatomedin A) and the Gene Ontology "Insulin-like growth factor binding molecular function". It is inferred from the IGF2 gene encodes IG2R_HUMAN (the EMBL/GenBank protein IDs AAL55889, AAY40360) proteins which is in the Entrez Gene data bank. EMBL/GenBank protein IDs are defined as database cross reference to the protein P09565 (Putative insulin-like growth factor 2-associated protein) defined in the UniProt database. In major databases the protein P09565 (Putative insulin-like growth factor 2-associated protein) is not defined as encoded by the IGF2 human gene. By using the currently available ones, the new annotation Insulin-like growth factor binding to the IGF2 human gene is inferred. We added 120,669 new annotations to between gene and Gene Ontology which is currently available 1,272,168 annotations by using

63,080,137 protein to Gene Ontology annotations. Our transitive relationship based inference approach will be used as in-silico identification method to reveal the new annotations. The newly detected annotations needs to have biological validation. However, the detected annotations enrich the integrated ones already available and help comprehensive searches to answer complex biomedical questions. The integrated and new annotated data is available by the web based interface Genomic and Proteomic Knowledge Base (GPKB) which is using the GPDW as backend. It is publicly available at http://www.bioinformatics.dei.polimi.it/GPKB/ with an easy-to-use Web interface.

**Contact email:** canakoglu@elet.polimi.it

**URL**: http://www.bioinformatics.dei.polimi.it/GPKB/

Poster: Biological Databases – BD3

**A Cloud-based Approach to a Semantic Network in Cell Biology**

Dessì N(1), Diaz G(2), Milia G(1), Pascariello E(1)

*(1)Department of Mathematics and Computer Science, University of Cagliari, Cagliari (2) Department of Biomedical Sciences, University of Cagliari, Cagliari*

**Motivation:** Semantic networks in Biology started out as a natural way to provide intuitive and useful representations for modelling semantic knowledge and inference in this domain. In these networks links between any two nodes are directly associated in some way in order to express not only taxonomical but also functional and causal relationships. Equipped with some kind of procedural component which is expected to supports both network management operations and knowledge inference tasks, a semantic network offers little or no general insights without an efficient data structure which supports its structural principles. There is a fair consensus about using database ideas to store the network and a relatively less agreement about the data models for representing its structure, specially for large-scale semantic network. One avenue of particular interest is the possibility of implementing the network on a cloud platform using a NoSQL database. Cloud paradigm is especially attractive for large scale, elastic novel applications. NoSQL databases are a new generation of databases featured by a schema-free data model and driven by the idea that not every data management problem is best solved using a relational databases. To evaluate the potential for cross fertilization among these two emerging technologies within a semantic network, we investigate about the following aspects: -To what extent the structure of a semantic network can be supported by the schema-free data model that characterizes NoSQL databases? -To what extent can cloud computing improve processing tasks in terms of elastic and scalable applications?
**Methods:** With the aim to give a practical answer to the above questions, we present a cloud based web application for designing and the exploring a semantic network in Cell Biology. Our application exploits Google App Engine, a cloud platform which includes a NoSQL database supporting the management of data objects as entities. Entities are clustered in collections, namely kinds, which categorise the entities they group. Each entity has a key that uniquely identifies it within its kind and one or more properties. A property can be a string, an integer or a reference to another entity. Entities grouped by the same kind can have different properties. The web application consists of a knowledge base and a set of procedural modules. Specifically, the knowledge base is modelled by a semantic network whose nodes are semantic types representing named entities of the knowledge domain. Each named entity belongs to a single high-level category. Labelled links represent relationships among semantic types (i.e. causal, functional, inverse functional etc.). The semantic network consists of 531 semantic types (nodes) belonging 7 different high-level categories and 12 distinct relationships. The knowledge base is stored by means of two kinds. The first kind is a collection of entities which consists of couples (semantic type, category) and asserts knowledge about the categorization of the semantic types. The second kind maps the network structure into a collection of data objects where each object contains: a couple of identifiers which reference two semantic types (namely A and B) belonging to the first kind, the label which specifies the type of relationship among these entities (A vs B), an integer denoting the strength of this relationship. Python modules support the procedural component allowing the network management and interactive visualization of the network structure. When the user queries about a semantic type, the

application returns a labelled graph which visualizes the entities the semantic type is linked to.

**Results:** We provided the scientific community with a resource that can be easily interrogated to obtain structural and functional information about Cell Biology domain. Within a single application, the semantic network administrator may easily perform updates and users retain the advantage of browsing the network from the web. Our approach demonstrates that NoSQL databases and Cloud Computing may have potentially significant implication for developing semantic networks in Cell Biology in terms of structure and knowledge representation and performance of the procedural component. In particularly, the unique characteristics of NoSQL databases imply a different way of designing semantic networks while the cloud environment allows to take advantage of the ability to scale automatically physical resources as the number of nodes or relationships increases. This scalability is related both to the load (i.e. it is easy to expand the network to accommodate new nodes) and the geography (i.e. the web application maintains usefulness and usability regardless of how far its users are).

**Contact email:** dessi@unica.it

Poster: Biological Databases – BD4

**A Food Additive 3D Repository for in silico screening in Food Chemistry: The case of Androgen Receptor**

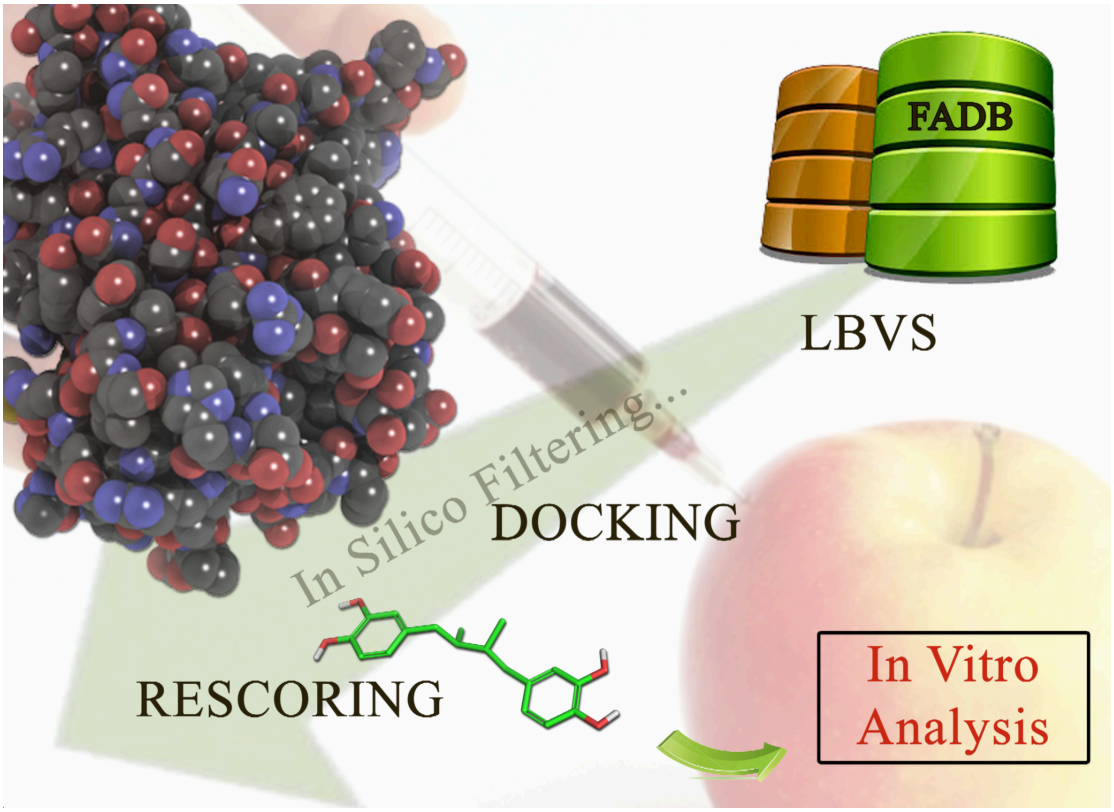
Ginex T(1) and Cozzini P(1)

*(1) Department of Food Science, University of Parma, Parma*

**Motivation:** As defined by the Codex Alimentarius Commission, food additives category includes a wide variety of substances that are added to food for technological and organoleptic purpose in the manufacture, processing, preparation, treatment, packing, packaging, transport or holding of such food results. From a legislative point of view, allowed additives defined as GRAS ("generally recognized as safe") are considered safe under the conditions of its intended use while additives not included in GRAS list must undergo toxicity testing to prove their safety. Even if FDA refers to food additives as GRAS, how can we be sure about their safety? In this regard, EFSA is re-evaluating all food additives by 2020, especially some food colours, to update risk assessment for food additives consumption (http://www.efsa.europa.eu/en/press/news/120130b.htm). Previous safety assessments will be reviewed and updated to take into account new scientific information pointing at a possible concern for health. Unfortunately, in vitro safety assessment is very expensive and challenging due to plenty of data to analyse. In this contest, in silico techniques as high-throughput virtual screening and molecular docking simulations could be a valid and innovative support for more efficient preliminary data evaluations. Despite the wide potentialities of this new modus operandi, the lack of specific, efficient and available free dataset of 3D small molecules (focused on alimentary field) is the rate-limiting step to the development of computational toxicology in food safety area.
**Methods:** Here, we present the first draft of FADB (Food Additives Data Base), the 3D version of EAFUS list of 3969 food additives dataset. Dataset population was reached through automatic retrieval of SD files from PubChem by using Microsoft Visual Studio 2010 C# Tool. In particular, CAS RN was used for recursive inquiry on PubChem home page. In this way, a total of 2649 molecules were collected whereas the remaining CAS RNs (25%) for which there were no "molecular feedback" in PubChem database has been manually edited with the support of TorchV10Lite software (www.cresset-group.com). All molecules were subjected to structural check for correct assignment of atom and bond types whereas inorganic ions in salts and other unnecessary components were deleted. Moreover, also molecular redundancies were excluded. Given a biological system, FADB can be used for in silico preliminary evaluations of toxicity in order to select chemicals that have necessary and sufficient conditions to perturb specific physiological equilibria. In our case, the dataset was used for preliminary toxicological studies on AR-LBP. The procedure can be summarized as follows: LBVS with FLAP (http://www.moldiscovery.com/) for a preliminary filtering, Docking of selected candidates with GOLD (http://www.ccdc.cam.ac.uk/pages/Home.aspx) and Rescoring of all poses generated during docking simulations with HINT Scoring Function (http://www.edusoft-lc.com/hint/).
**Results:** By applying the filtering protocol described above on Androgen Receptor, three potential binders for AR have been identified: they are the antioxidant NorDihydroGuaiaretic Acid (NDGA), the antioxidant Phloretin and the flavouring agent 1-(2-Hydroxyphenyl)-3-(pyridine-4-yl)-propan-1-one. In particular, the latter belongs to the class of flavourings for which The Codex Alimetarius Commission has requested additional data to complete the evaluation (ftp://ftp.fao.org/codex/meetings/ccfa/ccfa45/fa45_15e.pdf). Here are reported binding modes and affinity values of these three chemicals for AR: the selection was based

on the assumption that compounds with binding affinity higher than the reference compound testosterone and a suitable coverage of the active site could be able to influence AR activity in a ligand-dependent manner. In light of these results, further in vitro studies needs to be performed for a proper risk assessment.

**Contact email:** tiziana.ginex@studenti.unipr.it



Poster: Biological Databases – BD5

**A knowledge-base for the vitis-vinifera functional analysis**

Pulvirenti A (1)∗,+, Giugno R (1)*,+, Distefano R (1), Pigola G (3), Mongiovì (2), Giudice G (4), Vendramin V (3), Lombardo A (5), Cattonaro F (3), Ferro A (1)

*(1) Department of Clinical and Molecular Biomedicine, University of Catania. (2) Department of Mathematics and Computer Science, University of Catania. (3) IGA Technology Services. (4) Cardiovascular Development and Repair Department, Centro Nacional de Investigaciones Cardiovasculares (CNIC), Melchor Fernández Almagro 3, 28029 Madrid, Spain (5) Parco Scientifico e Tecnologico, Catania. ∗ Corresponding author + Equal Contributors*

**Motivation:** Vitis vinifera L. (Grapevine) is the most important fruit species in the modern world. Wine and table grapes sales contribute significantly to the economy of major wine producing countries. The two main goals in wine production concern quality and safety [1]. To properly deal with these aspects and to gain biological knowledge about cultivars, a genomic approach is the most reliable strategy. The recently published grapevine genome sequence offers the opportunity to study the potential roles of genes and microRNAs in fruit maturation and other physiological and pathological processes [2]. Although several systems allowing the analysis of plant genomes have been reported [3], none of them has been designed specifically for the functional analysis of grapevine genomes of cultivars under environmental stress in connection to microRNAs data.
**Methods:** Here we introduce a novel knowledge base, called BIOWINE, designed for the functional analysis of Vitis vinifera genomes of cultivars present in Sicily. The system allows the analysis of RNA-seq experiments (i.e. exome and short-noncoding RNAs) of two different grapevine cultivars present in Sicily, namely Nero d'Avola and Nerello Mascalese. Each cultivar was taken under different climatic conditions of phenological phases, diseases, and geographic location. The BIOWINE web interface is equipped with data analysis modules for grapevine genomes. In particular users may analyze the current genome assembly (Genoscope 12X) together with the RNA-seq data through a customized version of GBrowse. The web interface allows users to perform gene set enrichment by exploiting third-party databases such as: KEGG Pathway, Gene Ontology, mirBase, UNIPROT, and PlantGDB.
**Results:** BIOWINE is a knowledge base implementing a set of bioinformatics tools for the analysis of grapewine genomes. The system aims to increase our understanding of the grapevine varieties and species, with particular attention to Sicilian products, from the point of view of adaptability to different climatic conditions, phenological phases, diseases, and geographic location.
**Contact email:** apulvirenti@dmi.unict.it, giugno@dmi.unict.it
**Supplementary information:**[1] Vivier MA, Pretorius IS (2002) Genetically tailored grapevines for the wine industry. Trend in Biotechnology 20: 472–478. [2] Mica E, Piccolo V, Delledonne D, Ferrarini A, Pezzotti M, Casati C, Del Fabbro C, Valle G, Policriti A, Morgante M, Pesole G, Enrico Pè ME and Horner DS. High throughput approaches reveal splicing of primary microRNA transcripts and tissue specific expression of mature microRNAs in Vitis vinifera, BMC Genomics 2010, 11:109 [3] PlantGDB, http://www.plantgdb.org/VvGDB/

Poster: Biological Databases – BD6

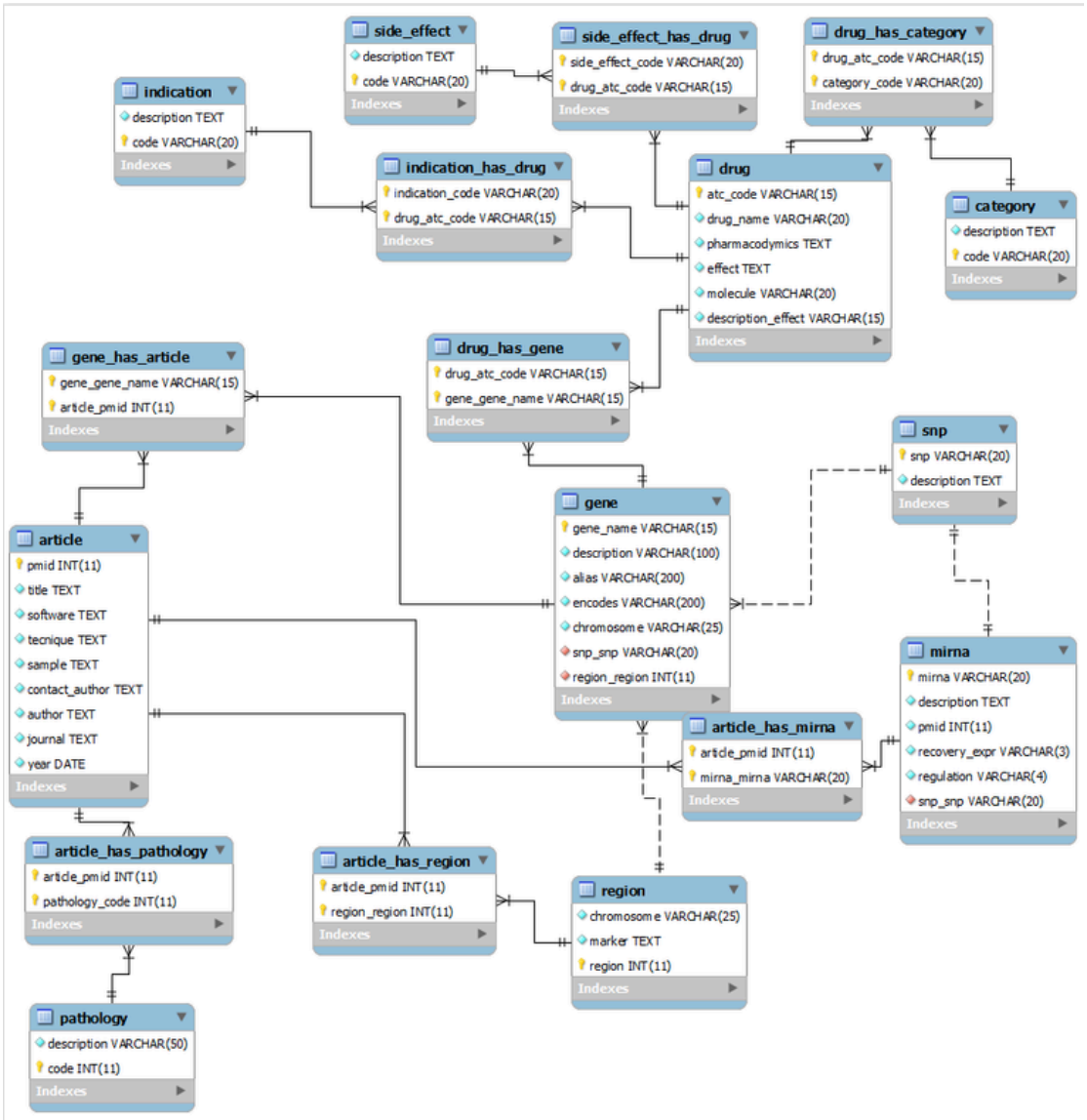
**OCDB: the first overall database collecting genes, miRNAs and drugs for Obsessive-Compulsive Disorder**

Privitera AP (1), Russo F (2), Distefano R (1), Ferro A (1), Pulvirenti A (1), Giugno R (1)

*(1) Department of Clinical and Molecular Biomedicine, University of Catania. (2) Laboratory of Integrative System Medicine (LISM), Institute of Informatics and Telematics (IIT) and Institute of Clinical Physiology (IFC), National Research Council (CNR), Pisa*

**Motivation:** Obsessive-compulsive disorder (OCD) is a psychiatric condition characterized by intrusive and unwilling thoughts (obsessions) which give rise to anxiety. The patients feel forced to perform a behavior (compulsions) related to the obsession. They suffer very much since they struggle to resist to their compulsions while feel to be required to perform them. The World Health Organization ranks OCD as 1 of 10 most disabling medical conditions worldwide(1). The OCD is a pathology, among all anxiety disorders, for which hereditary component has been demonstrated. Consequently online resources collecting and integrating scientific discoveries and genetic evidences about OCD would be very helpful.
**Methods:** Several studies supporting the existence of genetic susceptibility factors have been reported (2). The most studied genes are SLC1A1, SLC6A4, SLC6A2, HTR2A, HTR2C and,more recently, NTRK3 and SLITRK1. Very few studies reported the role of miRNAs down-regulating disease candidate genes and drug targets in anxiety disorders and OCD(3)(4) . Here we propose a manually curated database, OCDB, which will be soon available on line showing relationship between OCD–associated genes, OCD–miRNAs and OCD–drugs with particular emphasis on drug targets and their polymorphisms. We maintain the therapeutical and side effects of the drugs, miRNA and drug targets polymorphisms, miRNA binding and information about the candidate disease genes (location, alias, their use as putative biomarkers, and so on). Relevant pathologies (OCD or associated disease such as social phobia and schizophrenia) and experimental conditions (i.e., number of samples, used softwares, and so on) will be also reported. The database has been created using MySQL, Apache and PHP. It contains relations among 45 genes, 24 miRNAs and 7 drugs. Several tools for advanced search and browsing have been integrated. An export function to download search results in various formats (CSV, XLS, TXT) will be available.
**Results:** Our database is an essential resource to improve knowledge on this pathology and to support genome-wide analysis studies. For instance using our database the gene HTR2C results to be simultaneously a target of the Clomipramina drug and of miR-22(4), and contain polymorphisms. Such information was separately reported in 5 publications. Finally, genetic scientific results concerning OCD related to autism and other anxiety disorder are also reported. For all these reasons OCDB may represent an important knowledge base for researchers working on anxiety disorders.
**Contact email:** giugno@dmi.unict.it
**Supplementary information:**1. Murray C, Lopez A. Global Burden of Disease: a Comprehensive Assessment of Mortality and Morbidity From Diseases, Injuries and Risk Factors in 1990 and Projected to 2020. Vol. 1. Cambridge, Mass: Harvard University Press; 1996. 2. Bloch MH, Pittenger C. The Genetics of Obsessive-Compulsive Disorder. Curr Psychiatry Rev. 2010 May 1;6(2):91-103. 3. Muiños-Gimeno M, Guidi M, Kagerbauer B, Martín-Santos R, Navinés R, Alonso P, Menchón JM, Gratacòs M, Estivill X, Espinosa-Parrilla Y. “Allele variants in functional MicroRNA target sites of the neurotrophin-3 receptor gene (NTRK3) as susceptibility factors for anxiety disorders.” Hum Mutat. 2009 Jul;30(7):1062-71. doi: 10.1002/humu.21005. 4. Muiños-Gimeno M, Espinosa-Parrilla Y, Guidi M, Kagerbauer B,

Sipilä T, Maron E, Pettai K, Kananen L, Navinés R, Martín-Santos R, Gratacòs M, Metspalu A, Hovatta I, Estivill X. Human microRNAs miR-22, miR-138-2, miR-148a, and miR-488 are associated with panic disorder and regulate several anxiety candidate genes and related pathways. Biol Psychiatry. 2011 Mar 15;69(6):526-33. doi: 10.1016/j.biopsych.2010.10.010. Epub 2010 Dec 17. Note that OCDB will be soon on line. Authors will communicate the web address as soon as possible.



Poster: Biological Databases – BD7

**A new version of GALT protein database**

d'Acierno A(1), Facchiano A(1), Marabotti A(2)

*(1) Institute of Food Science, CNR, Avellino (2) Department of Chemistry and Biology, University of Salerno, Fisciano (SA)*

**Motivation:** Galactose-1-phosphate uridylyltransferase (GALT) is the second enzyme in the metabolic pathway of galactose catabolism, and its impairment is related to the rare genetic disease classic galactosemia. More than 260 variations are known, of which about 60% are missense variations having a direct impact on protein structure and function. In addition, this disease is characterized by compound heterozygosis i.e. galactosemic people often show a combination of mutations. A clear relationship between genotype and phenotype is still not clear, also because the full characterization of the effect of each variation at protein level is hard to achieve by experimental methods. In order to assist in the interpretation of the effect of each mutation at molecular level, in the past we developed a model of the 3D structure of the human enzyme, which has not yet been obtained by experimental approaches [1], and we also simulated the effect of many known missense mutations, in terms of their effect on enzyme activity, stability, intersubunit association and other structural features [2]. This information has been collected in a Web-accessible database that allowed the automatic interrogation of data available on the wild type protein, and the creation of a static HTML page with a summary of the effects of each variation analyzed [3]. During the following years, the growth of available information on new GALT variations [4-7] and especially the need to perform more accurate and effective interrogation pushed us to reorganize our database. Therefore, here we present GALT Protein Database 2.0 and all its new features.
**Methods:** The domain under study has been first modeled using an entity-relationship (ER) model, having in mind of realizing a database capable of storing data not just for the GALT protein. Next, the ER diagram has been translated into a relational logical model, and several indexes and materialized views have been created to make queries simpler and faster. The web application has been realized using a Model-View-Controller design paradigm, and Struts 2 has been adopted as developing framework. Thus, JAVA has been adopted as coding language for the model layer and JSP has been used for the presentation layer. The available running version uses PostgreSQL as DBMS and Tomcat as application server. The creation of the models of the missense variants and the analysis of their structural properties have been made as described previously [2].
**Results:** The new Web application is still composed by two main sections (the first one for wild type protein, the second one for variants). Data available for wild type protein have been increased with respect to previous database version, and the interface with users has been changed. In fact, in order to improve readability by non-specialists, data for each residue are first presented synthetically with a simplified level of interpretation, and then, integral information is shown only on request, by clicking on the residue number. The section devoted to GALT variations has been deeply changed. All structural and functional information on the variants are now stored into the database, and can be extracted according to queries based on structural properties. In addition, a new organization of data has allowed to manage separately variations on the two chains of this dimeric enzyme: in this way, also data about compound heterodimeric proteins can be managed. The DBMS is now also in charge of handling PDB models of the variant protein and, when the variation involves the active site, of the complex with the substrate. These PDB files can be

downloaded by users, or visualized interactively using Jmol (http://www.jmol.org). The Web page created for each variant is no longer static but dynamically created, including not only information related to the variation, but also the comparison with the corresponding results obtained for wild-type protein. Additional sections of the Web applications are still present and have been enriched. The present application is only the first step towards the development of a more general platform that could collect structural data for virtually any mutant form of any protein. Our future plans include also the interactive creation of variants, and the creation and analysis not only of missense variations, but also of more complicated indels or nonsense variations.

**Contact email:** amarabotti@unisa.it

**Supplementary information:**Acknowledgements This work is supported by a grant from 'Fondi di Ateneo per la Ricerca di Base (FARB)'2013, ORSA130225, and has been partially supported by the Flagship InterOmics Project (PB.P05, funded and supported by the Italian Ministry of Education, University and Research and Italian National Research Council organizations). References: [1] Marabotti A, Facchiano A. (2005) Homology modeling studies on human galactose-1-phosphate uridylyltransferase and on its galactosemia-related mutant Q188R provide an explanation of molecular effects of the mutation on homo- and heterodimers. J Med Chem, 48, 773-779. [2] Facchiano,A. Marabotti,A. (2010) Analysis of galactosemia-linked mutations of GALT enzyme using a computational biology approach. Protein Eng Des Sel, 23, 103-113. [3] d'Acierno,A. et al. (2009) GALT protein database, a bioinformatics resource for the management and analysis of structural features of a galactosemia-related protein and its mutants. Genomics Proteomics Bioinformatics, 7, 71-76. [4] Boutron,A. et al. (2012) Mutation spectrum in the French cohort of galactosemic patients and structural simulation of 27 novel missense variations. Mol Genet Metab, 107, 438-447. [5] Singh,R. et al. (2012) Frequency distribution of Q188R, N314D, Duarte 1, and Duarte 2 GALT variant alleles in an Indian galactosemia population. Biochem Genet, 50, 871-880. [6] Tang,M. et al. (2012) Correlation assessment among clinical phenotypes, expression analysis and molecular modeling of 14 novel variations in the human galactose-1-phosphate uridylyltransferase gene. Hum Mutat, 33, 1107-1115. [7] Coelho,A.I. et al. (2013) A frequent splicing mutation and novel missense mutations color the updated mutational spectrum of classic galactosemia in Portugal. J Inherit Metab Dis, June 8, 2013. DOI: 10.1007/s10545-013-9623-1.

Poster: Biological Databases – BD8

**FastSemSim: fast and easy evaluation of semantic similarity measures on biomedical ontologies**

Mina M(1), Sanavia T(2)

*(1) MPBA, Fondazione Bruno Kessler, Trento (2) Department of Information Engineering, University of Padova, Padova*

**Motivation:** Although biological databases addressed the problem of providing consistent and formal descriptions of genes and proteins using different data structures, the quantification of the functional similarity between genes or proteins exploiting these data is still a challenge. Several semantic similarity measures (almost 50) have been proposed to compare either the terms from biomedical ontologies or the genes/proteins annotated with them, using different approaches to consider the relationships among the annotations (Guzzi et al., 2012). Their implementations can be distinguished in two categories: on-line and off-line. For on-line analyses, G-SESAME (Zhidian et al., 2009), FunSimMat (Schlicker and Albrecht, 2010 and ProteInOn (Pesquita et al., 2009) are the most recently proposed and provide, beyond the classical information-theoretic based metrics, new approaches based on term specificity. On the other hand, almost all off-line tools are developed in R language: GOSemSim (Yu et al., 2010) and csbl.go (Ovaska et al., 2008), are among the most used. However, current applications are not specifically designed to manage huge amounts of data in order to support genome/proteome-wide analyses. Recently, new integrative approaches in the analysis of high-throughput data has proven that the integration of prior knowledge from biomedical ontologies is a useful resource to improve the identification of expression patterns (Di Camillo et al., 2012) and to provide more stable biomarker lists in classification problems (Sanavia et al., 2012). Therefore, more efforts are required to develop applications which are able to be both scalable across genome/proteome-wide data and enough flexible to provide a user-friendly platform to calculate these similarities and to integrate them within new computational pipelines.

**Methods:** FastSemSim is both a Python library and an end-user application, featuring an intuitive graphical user interface (GUI). As input data, the library requires the ontology graph and the gene/protein annotations. The current version of FastSemSim handles both obo and obo-xml daily updated files and supports any multi-rooted ontology (as long as it is acyclic). The library was implemented with a modular architecture: a core component includes all the routines for parsing the ontology and the annotations to extract common features for the measures (e.g. Information Content), whereas all the semantic similarity measures were developed on top of the core library as independent modules. The library currently supports 16 different measures, both pairwise and groupwise. While groupwise measures can directly evaluate gene/protein similarities considering the corresponding sets of terms, pairwise measures need a "mixing strategy" (Guzzi et al., 2012). The three most used strategies were implemented: the average (avg) and the maximum (max) of all term pairwise similarities, and the average of similarities between best matching terms (Best Match Average). FastSemSim supports inter-ontology and inter-category (e.g. Molecular Functions and Biological Process in Gene Ontology) relationships, and provides several filtering functions which allow the user to perform organism-specific analysis or to work only with specific evidence codes. In addition, a GUI is provided to easily calculate the similarity measures, characterized by a user-friendly front-end to load the ontology and the annotation files, to input the query and to select the output parameters. Both the library and the interface are

compatible with Python 2.x and were tested on Microsoft Windows, OS X and different Linux distributions.

**Results:** Compared with the most used off-line and on-line available tools for semantic similarities, FastSemSim shows the highest coverage of implemented semantic similarity measures, enabling the systematic evaluation of different measures (Cho et al., 2013). Scalability was tested on Gene Ontology (GO) annotations for the categories Biological Process and Molecular Function across the proteomes of several organisms (Human, Mouse, Fly and Yeast). Resnik measure (Resnik, 1999) combined with the max mixing strategy, conventionally used for protein-based studies, was applied. FastSemSim was able to accomplish the analysis for all the proteomes, ranging between 2 minutes for the Yeast proteome (6380 GO annotated proteins) and 5 hours and 18 minutes for Human proteome (45576 GO annotated proteins). Available R applications do not provide an efficient implementation able to deal with more than 1000 genes/proteins efficiently. FastSemSim proved to meet the requirements of handling huge amounts of data. Moreover, the Python implementation and the modular architecture of the library can be easily exploited to both integrate semantic similarity within computational pipelines and extend the library with new measures.

**Contact email:** mamina@fbk.eu

Poster: Biological Databases – BD9

**PPStruct: a database of plant protein stuctures and annotations**

Potenza E(1), Coller E(2), Hirsh L(1), Di Domenico T(1), Cestaro A(2), Tosatto SCE(1)

*(1) Department of Biomedical Sciences, Universita' di Padova, Padova (2) Computational Biology Department, Fondazione Edmund Mach, Trento*

**Motivation:** During the last ten years, the development of high-throughput sequencing, has generated a huge amount of genome sequences. Giving biological meaning to this data depends entirely on the capacity to develop instruments for its interpretation and organization. Moreover, once the protein sequences have been identified, functional annotation requires dedicated usage of an enormous amount of bioinformatics resources and specialized databases. Sequence annotation is often inaccurate and reliable predictions can only be obtained by using structure based functional annotation methods. These methods require the three-dimensional structure of the identified proteins. The experimental solution of protein structures is very time consuming and cannot be applied to all proteins in a genome, but has to be replaced with computational homology models. It is currently estimated that well over half of the known protein sequences can be predicted in this way. Plant genomics, despite its importance, started later than animal genomics. Currently the are less than ten plant genomes available in a genome browsers and few more at the "draft genome" level. In light of this limited amount of available data, any consideration regarding peculiar plant characteristics has to be considered temporary and seen with caution. Plant genomes were so far mostly annotated by hand, with an enormous expenditure of financial and human resources. Genome annotation for plants has to transit from prevalently manual towards fully automated annotation, with possible manual supervision, and is in serious need for the creation of new tools to permit this transition.
**Methods:** PPStruct database and website was designed with a multi-tier architecture, using separate modules for data management, data processing and presentation functions. To simplify development and maintenance, all tiers handle the common JSON (JavaScript Object Notation) format, thereby eliminating the need for data conversion. The MongoDB database engine is used for data storage and Node.js as middleware between data and presentation. PPStruct exposes its resources through RESTful web services, by using the Restify library for Node.js. The Angular.js framework and Bootstrap library were selected to provide the overall look-and-feel. Additional information is added to entries by querying the PDB and UNIPROT web services. Currently the genomes available at the database were annotated for the following features: - Domain assignment: InterPro tools set (Hunter et al., 2009) - Secondary structure: fastSS (Walsh et al., unpublished) - Disordered regions: MobiDB (Di Domenico et al., Bioinformatics 2012) - Homology modelling: HOMER (URL: http://protein.bio.unipd.it/homer/)
**Results:** Here we present PPStruct, a pipeline and a database dedicated to plant functional annotation. Our effort takes into account several specific aspects exploiting plant differences. The protein structure level is brought into play with the aim to better explain the effects of phenotypic differences at the molecular level. Reliable models are built for each gene transcript identified and the models will be used to better define the function of each protein. PPStruct website is currently under development but will be available soon from URL: http://ppstruct.bio.unipd.it/
**Contact email:** emilio.potenza@bio.unipd.it

Poster: Biological Databases – BD10

**Pfam: the protein families database.**

Finn RD(1), Bateman A(2), Clements J(1), Coggill P(2,3), Eberhardt RY(2,3), Eddy SR(1), Heger A(4), Hetherington K(3), Holm L(5), Mistry J(2), Sonnhammer EL(6), Tate J(2,3), Punta M.(2,3)

*(1) HHMI Janelia Farm Research Campus, Ashburn, VA USA (2) European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Hinxton, Cambridge UK (3) Wellcome Trust Sanger Institute, Hinxton, Cambridge UK (4) MRC Functional Genomics Unit, Department of Physiology, Anatomy and Genetics, University of Oxford UK (5) Institute of Biotechnology and Department of Biological and Environmental Sciences, University of Helsinki Finland (6) Stockholm Bioinformatics Center, Swedish eScience Research Center, Department of Biochemistry and Biophysics, Stockholm University Sweden*

**Motivation:** Pfam [1], available via servers in the UK (http://pfam.sanger.ac.uk/) and the USA (http://pfam.janelia.org/), is a database of protein families that is widely used for homology-based functional annotation. We describe our latest release (27.0) and discuss present challenges.
**Methods:** Pfam families are defined by two alignments ('seed' and 'full') and a profile hidden Markov model (HMM). The seed alignment from which the profile HMM is built needs to be of high quality, as it provides the basis for the position-specific amino-acid frequencies, gap and length parameters in the model [2]. The seed alignment for each Pfam family is a set of curator-defined sequences that are representative for that family. Pfam profile HMMs are searched against a large sequence collection, based on UniProtKB [3], to find all instances of the family. Sequence regions that score above the curated threshold that is set for each family to eliminate false positives (the so-called gathering threshold) are aligned to the profile HMM to produce the full alignment. Curated entries are referred to as Pfam-A entries. The profile HMMs are built and searched using the HMMER software suite (http://hmmer.janelia.org). Pfam data are available in a variety of formats, which include flatfiles (derived from the MySQL database) and relational table dumps, both of which can be downloaded from the FTP site (ftp://ftp.sanger.ac.uk/pub/databases/Pfam). The Pfam website provides different ways to access the database content, providing both graphical representations of and interactive access to the data.
**Results:** The current release of Pfam, version 27.0, contains 14,831 Pfam-A families. Compared with Pfam 26.0, there has been an increase of 1,159 families. The Pfam-A families in release 27.0 match 79.9% of the 23.2 million sequences and 58% of the 7.6 billion residues in the underlying sequence database. This corresponds to a negligible percentage increase in sequence and residue coverage with resepct to release 26.0 (<0.5%), but reflects a significant amount of curation effort. Indeed, in the 20% of sequences and 42% of residues that remain un-annotated, Pfam is facing what appears to be a highly diverse set of protein regions, where large families are rare and clade-specific homologous groups abound. This new scenario has influenced Pfam strategies for building new families over the last couple of years. During this period, curation has focused mainly on human and PDB protein sequences, that is, sequences that are potentially of higher significance to the community of biologists. Our selection and targeting of human protein regions, has been described extensively in [4] and on our blog http://xfam.wordpress.com/2013/05/07/pfam-targets-conserved-human-regions/ and http://xfam.wordpress.com/2013/05/14/case-studies-from-the-list-of-human-regions-not-in-pfam-27-0/. The explosion in the number of protein sequences being released in public repositories such as UniProt has brought about new important challenges for Pfam. Among these, is the increasing difficulty in aiding human

interpretation of multiple sequence alignments through visualization. For example, the largest Pfam-A family (version 27.0) with >363,000 matches to the profile HMM is the ABC transporters family (ABC_tran, accession PF00005)—its full alignment is thus too large to be usefully visualised. The seed alignment, by contrast, contains just 55 representative sequences, which may be an insufficient number to represent the sequence diversity within the family. To provide more useable samples of the sequence diversity within a family, we now additionally calculate model-matches for four additional sequence sets, based on Representative Proteomes (RPs) [5]. We use the RP sequence sets constructed using co-membership thresholds of 75, 55, 35 and 15%, giving a range of sequence redundancy for each family. Using RPs has the advantage that it still allows for organism-specific copy numbers to be assessed, a feature that can be lost when using global non-redundancy thresholds on an entire sequence database. However, the major advantage for Pfam is the dramatic reduction in the size of the family full alignments. For the ABC_tran family, the RP alignments range in size from approximately a quarter of the size of the full alignment to less than one tenth. Other changes introduced in the latest release include faster interactive DNA searches, mark up of protein disordered regions (using the prediction program IUPred[6]) and removal of sequences potentially derived from spurious open reading frames according to the AntiFam database [7]. [1] Finn et al. Nucleic Acids Res. 2013 (Epub ahead of print) [2] Krogh et al. J. Mol. Biol. 1994;235:1501-1531. [3] UniProt Consortium. Nucleic Acids Res. 2012;40:D71-D75. [4] Mistry et al. Database 2013;2013:bat023. [5] Chen et al. PLoS One 2011;6:e18910 [6] Dosztányi et al. Bioinformatics 2005;21:3433-3434. [7] Eberhardt et al. Database 2012;2012:bas003.

Poster: Biological Databases – BD11

**miMETA: an online meta-analysis tool for the miRandola database**

Di Bella S(1), Russo F(2,3), Nigita G(4), Pulvirenti A(1), Giugno R(1), Ferro A(1)

*(1) Department of of Clinical and Molecular Biomedicine, University of Catania, Catania, Italy (2)Laboratory of Integrative System Medicine (LISM),Institute of Informatics and Telematics (IIT) and Institute of Clinical Physiology (IFC), National Research Council (CNR), Pisa, Italy (3) Department of Informatics, University of Pisa, Pisa, Italy (4) Department of Mathematics and Computer Science, University of Catania, Catania, Italy*

**Motivation:** MicroRNAs (miRNAs) are small non-coding RNAs that act as post-trancriptional regulators of protein coding genes. A lot of miRNAs have been found to correlate well with many human cancers and other diseases. Recently, miRNAs have been found in human body fluids such as plasma, serum and urine and they are considered potential non-invasive biomarkers. For this reason we recently developed miRandola, the first extracellular circulating miRNAs database. To improve the usability of the database as a key source to inform clinical practice, in the present study we introduce miMETA, a meta-analysis tool for miRandola.
**Methods:** 0
**Results:** miMETA incorporates two R packages, Mada ('Meta-Analysis of Diagnostic Accuracy') and Metafor ('Meta-Analysis Package for R'). Mada provides functions for diagnostic meta-analysis, Metafor provides a comprehensive collection of functions for conducting meta-analysis in R. Standard methods for diagnostic accuracy meta-analysis have been applied: sensitivity, specificity, diagnostic odds ratio (DOR), log odds ratio, and the area under the curve (AUC). The AUC represents an analytical performance summary displaying sensitivity-specificity trade off. The methodological quality of each study can be assessed by QUADAS (QUality Assessment for studies of Diagnostic Accuracy), a 14-questions evidence-based quality assessment tool used for diagnostic accuracy studies evaluation. When a criterion is fulfilled, unclear or not achieved a score of 1, 0 or –1 is assigned respectively. Only studies with higher QUADAS score (≥ 10) are accepted. Users may select a disease related to extracellular miRNAs annotated in the miRandola database. Next they may decide to apply the QUADAS test. Finally information on the number of patients with True Positive (TP), False Positive (FP), True Negative (TN), False Negative (FN) tests and related miRNAs, for each study may be uploaded. The results page reports the meta-analysis graphs and extracellular miRNAs annotations for the specific desease. Links to miRandola and PubMed together with a table summarizing the main statistics for each study are provided. A case study for the circulating miRNAs in Acute Myocardial Infaction (AMI) patients from Lippi et al. is reported. Nine studies, totaling 1295 subjects (418 controls; 32%), were included in their meta-analysis showing the importance of this kind of studies in the diagnostic performance of these novel molecular biomarkers. By using miMETA users can obtain information about the reported potential biomarkers in the miRandola database and compare the results of their meta-analysis with available data of the database (information about extracellular miRNAs, diseases, cellular and extracellular miRNA expression, the potential biomarker role etc). For example, in the above meta-analysis miR-499 is a sensitive biomarker for AMI. The corresponding miMETA miRNA output-table reports links to miRandola with all the entries for miR-499. This miRNA is reported as a candidate biomarker for AMI and other cardiovascular diseases such as troponin-positive acute coronary syndrome and diastolic dysfunction. miRandola will be updated with information of case-control studies in order to perform meta-analysis for each disease as soon as they are

published. Finally, availability of miMETA, the first online tool allowing extracellular miRNA data meta-analysis, should encourage researchers to upload their own data into miRandola.
**Contact email:** francesco.russo@iit.cnr.it
**Supplementary information:**Russo F et al. (2012) miRandola: Extracellular Circulating MicroRNAs Database. PLoS ONE 7(10): e47786. Lippi G et al. (2013) Circulating microRNAs (miRs) for diagnosing acute myocardial infarction: meta-analysis of available studies. Int J Cardiol. 167(1):277-8. Whiting P et al. (2003) The development of QUADAS: a tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews. BMC Med Res Methodol. 3: 25.

**Global characterization of alternative splicing events in long non coding RNAs.**

Colantoni A(1), Helmer-Citterich M(1), Ferrè F(1)

*(1) Centre for Molecular Bioinformatics, Department of Biology, University of Rome Tor Vergata, Via della Ricerca Scientifica snc, 00133 Rome, Italy*

**Motivation:** Long non coding RNAs (lncRNAs) are long transcripts that share many characteristics with mRNAs (most of them are transcribed by RNA Polimerase II, are polyadenylated, have a multi-exonic structure and have standard canonical splicing signals) but lack protein coding potential. They are generally expressed at lower levels than protein coding genes. LncRNA genes show a modest conservation when compared with protein coding genes; however their promoters and exons evolve less quickly than neutrally evolving regions. Interestingly, more than 25% of lncRNA genes are associated to two or more transcripts [1]: this means that alternative splicing (AS) plays an important role in the definition of the expression profiles of these genes, and possibly in their functional modulation. Notwithstanding the enormous interest recently raised in understanding lncRNAs functions and biology, still the current knowledge of these important regulators of many cellular processes is remarkably incomplete. The aim of this work is to verify whether the general conclusions that have been drawn about the conservation and the expression profiles of AS events in protein coding genes are valid also for lncRNAs, or if new paradigms are needed.
**Methods:** A set of 3502 human lncRNA loci with two or more associated transcripts was obtained from the Gencode lncRNA annotation v15 [2]. Due to the low expression levels of lncRNAs, the annotation of their boundaries could be incomplete. For this reason we decided to analyze only alternative splicing events that affect internal exons. Among these events, we identified ~ 800 >40 nt long conditional exons-i.e. exons that are either included or skipped, and ~ 680 constitutive exons. We measured the conservation of conditional and constitutive exons using phastCons scores [3] calculated on a set of 33 placental mammals. We also measured the average splice site strength of 3' and 5' splice sites of conditional and constitutive exons of lncRNAs using the MaxEntScan software [4]. A similar analysis was also conducted on ~ 260 conditional 3' splice sites and ~ 220 conditional 5' splice sites. To study the differential usage of conditional exons across tissues, we used the publicly available Body Map 2.0 RNA sequencing data generated for a set of 16 human tissues and mapped the reads to the hg19 genome using Bowtie2 [5] and Tophat2 [6]. Reads mapped to exon-exon junctions that support the inclusion or the exclusion of the lncRNA conditonal exons were used to compute the level of inclusion of each exon across all the examined tissues.
**Results:** We observed that, for lncRNAs, the average strength of constitutive splice sites is higher than that of conditional splice sites, and this holds true also for protein coding genes [7][8]. With regard to protein coding genes, conditional exons have been shown to be less conserved than constitutive exons. Comparing the average conservation of lncRNA conditional exons vs. constitutive exons, we found that conditional exons are less conserved than constitutive exons, similarly to what has been observed for mRNA [9](p-value < 0.05). In proteins, constitutive splice sites are more conserved than conditional splice sites, but the intronic regions flanking the former are usually less conserved than those flanking the latter [10]. We found that this conservation pattern is maintained also in lncRNAs. In general, we found that the evolutionary features that distinguish alternative from constitutive splicing events in protein coding genes are found also in lncRNA genes: this suggests that the mechanisms and machinery that have been associated to AS of pre-mRNA could be involved

in the AS of lncRNAs as well. To draw ulterior parallels between these two classes of genes we also determined the differential expression patterns of lncRNA conditional exons. Splicing variant-specific expression patterns, as detected by next generation transcriptome sequencing, offered a detailed picture describing the cellular usage of different lncRNA variants carrying different sets of features, providing a comprehensive overview on how and why these molecules are shaped by splicing and offering clues about their evolution.

**Contact email:** fabrizio.ferre@uniroma2.it

**Supplementary information:**References 1. Derrien T, Johnson R, Bussotti G, Tanzer A, Djebali S, Tilgner H, Guernec G, Martin D, Merkel A, Knowles DG, Lagarde J, Veeravalli L, Ruan X, Ruan Y, Lassmann T, Carninci P, Brown JB, Lipovich L, Gonzalez JM, Thomas M, Davis CA, Shiekhattar R, Gingeras TR, Hubbard TJ, Notredame C, Harrow J, Guigo R: The GENCODE v7 catalog of human long noncoding RNAs: Analysis of their gene structure, evolution, and expression. 2012:1775–1789. 2. Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, Aken BL, Barrell D, Zadissa A, Searle S, Barnes I, Bignell A, Boychenko V, Hunt T, Kay M, Mukherjee G, Rajan J, Despacio-reyes G, Saunders G, Tanzer A, Steward C, Harte R, Lin M, Derrien T, Chrast J, Walters N, Balasubramanian S, Pei B, Tress M, Rodriguez JM, et al.: GENCODE: The reference human genome annotation for The ENCODE Project. 2012:1760–1774. 3. Felsenstein J, Churchill GA: A Hidden Markov Model approach to variation among sites in rate of evolution. Molecular Biology and Evolution 1996, 13:93–104. 4. Yeo G, Burge CB: Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. Journal of computational biology : a journal of computational molecular cell biology 2004, 11:377–94. 5. Langmead B, Salzberg SL: Fast gapped-read alignment with Bowtie 2. Nature methods 2012, 9:357–9. 6. Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL: TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome biology 2013, 14:R36. 7. Clark F, Thanaraj TA: Categorization and characterization of transcript-confirmed constitutively and alternatively spliced introns and exons from human. Human molecular genetics 2002, 11:451–64. 8. Koren E, Lev-Maor G, Ast G: The emergence of alternative 3’ and 5' splice site exons from constitutive exons. PLoS computational biology 2007, 3:e95. 9. Nurtdinov RN, Neverov AD, Favorov A V, Mironov AA, Gelfand MS: Conserved and species-specific alternative splicing in mammalian genomes. BMC evolutionary biology 2007, 7:249. 10. Chen L, Zheng S: Identify alternative splicing events based on position-specific evolutionary conservation. PloS one 2008, 3:e2806.

**Evaluating effects of RNA editing on protein-coding genes**

Grassi L(1), Leoni G(1), Tramontano A(1,2)

*(1) Physics Department, Sapienza University of Rome, Piazzale Aldo Moro, 5 I-00185 Roma, Italy. (2) Istituto Pasteur - FondazioneCenci Bolognetti, Sapienza University of Rome, Piazzale Aldo Moro, 5 I-00185 Roma, Italy.*

**Motivation:** RNA editing is a post-transcriptional process that introduces changes in RNA transcripts. It was first observed more than 25 years ago in kinetoplastid protozoa. The insertion or the deletion of many uridine nucleotides in the mitochondrial mRNA of the trypanosomes has functional effects at level of the edited proteins. The most prevalent type of RNA editing in the animal kingdom is mediated by the adenosine deaminase acting on RNA (ADAR) enzyme, it converts adenosines to inosines (A→I editing) in double-stranded RNA (dsRNA) substrates. Inosine is equivalent to guanosine; for this reason when the editing occurs at level of coding sequences it can induce a non-synonymous substitution. Several RNA-Seq studies identified many previously unreported A→I transitions at level of protein-coding genes in H. sapiens, C.elegans and D. melanogaster. Interestingly this scenario indicated new fascinating possibilities to grow the complexity of eukaryotic genomes. We aim to explore the effects of the amino acid substitutions caused by RNA editing in order to understand whether they can modulate protein functionality.
**Methods:** Two datasets (one relative to D. melanogaster and one relative to H. sapiens) are analysed. In order to exclude biases due to multiple mapping reads we restrict the analysis to the events overlapping single transcripts or multiple transcripts with unique coding sequence. Firstly we ask whether the editing acts preferentially on particular amino acid residues, preferring some non synonymous substitutions to other. In order to answer this question we compare the empirical substitutions with the ones obtained through a random model. This is realized by randomly mutating adenines to guanosines in the transcripts under examination. Cases of editing with synonymous amino acid substitutions are reported in literature at level of highly conserved residues. This would imply an important modulation of protein functionality caused by RNA editing. For this reason we want to test whether conservation is a general property shared also by the residues of the two datasets considered in our study. We calculate the Shannon entropy of edited residues by using multi-alignments of two groups of orthologoue proteins: one made by the 1to1 orthologs for related species and one less stringent including the orthologs defined by Homologene. Furthermore we investigate other specific properties of the edited residues as their solvent accessibility; their conformation in secondary structure and their tendency to be affected by post-transcriptional modifications.
**Results:** In this study we find interesting properties of edited residues. There are non synonymous substitutions occurring more than expected by chance both in H. sapiens and in D. melanogaster. Even though non synonymous mutations have different behaviours in human and in drosophila, the relative z-scores, obtained by comparing real and random frequencies, are positively correlated. Interestingly in both species, the most favoured amino acid substitutions are the ones not drastically altering their physical-chemical properties. Furthermore in both datasets, with both orthology groups, Shannon entropy distributions of edited and non-edited residues are very similar, indicating a non peculiar conservation of edited residues. All these evidences suggest that amino acid substitutions caused by RNA editing can only have a limited effect on protein functionality.
**Contact email:** luigra@gmail.com

Poster: Genomics, Transcriptomics and NGS – GTN3

**Building Custom frameworks for the ANalysis of VAriantS of many samples: the CANVAS tool.**

Benelli M(1), Contini E(1), Paganini I (2), Pippucci T (3), Papi L(2), Torricelli F(1)

*(1) Diagnostic Genetic Unit, Careggi University Hospital, Florence, Italy (2) Medical Genetics Unit, Department of Biomedical, Experimental and Clinical Sciences, University of Florence, Florence, Italy (3) Medical Genetics Unit, Department of Medical and Surgical Sciences, University of Bologna, Bologna, Italy*

**Motivation:** The comprehension of the molecular and cellular basis of diseases is vital for understanding the mechanisms of disease pathogenesis, improving diagnosis and decide on treatments. The high throughput sequencing (HTS) of genomes and exomes is an increasingly used tool that aids researchers in rapidly identifying genetic variation among many sequenced samples. However, selecting the subgroup of variants underlying disease remains challenging. Here we present a new software package that allows researchers to design prioritization schemes based on custom research criteria. Our variant analysis framework can be used to search for candidate alterations at gene and/or variant level. Applications of our tool include the identification of de novo mutations, family-based studies, detection of somatic variations as well as population studies.
**Methods:** The CANVAS package is developed in perl and R scripts. The required inputs of CANVAS are the variants of the samples as VCF >= 4.1 format and a file that users have to create to describe the relationships among input samples. An optional file can be used to specify a set of filters that can be applied to the variants of each sample (e.g, genotype = heterozygous, minimum depth = 10, 1000g frequency < 0.01, etc). The file containing the relationships among samples is aimed at the design of custom prioritization schemes. After VCF files are imported and metadata are collected, variants are annotated by ANNOVAR (Wang et al., 2010). The analysis can be performed at gene or variants level and follows the rules (prioritization schemes) reported in the relationship file. For istance, in family-based studies, one or more variants shared among family members may be sought as candidates (variant analysis). On the other hand, in cancer genomics studies one may be interested in looking at the most frequently altered genes to search for cadidate driver mutations (gene analysis). The output of CANVAS is a tab-delimited file containing all the relevant information about the prioritized variants/genes. These include gene names, variants information (genomic coordinates, Amino Acid changes) and information about samples that share variants/genes.
**Results:** To test the suitability of our tools, we applied CANVAS in a project comprising the whole exome sequencing of 35 familial cases of schwannomatosis. CANVAS generated a list of prioritized variants that helped clinician identify the genes involved in the diseases (the paper is in preparation). In conclusion, we have developed CANVAS, a flexible tool to perform custom proritization analysis of the variants emerging from the high throughput sequencing of many samples. Our tool is aimed at exploring and interpreting the hundreads of thousands of variants that are usually found in genome or exome sequencing projects. CANVAS will facilitate discovery of the genetic basis of human diseases including unsolved Mendelian disorders and cancer genomics studies. CANVAS will be soon available under GPL3 license.
**Contact email:** matteo.benelli@gmail.com

**Tomato genome annotation: genome peculiarity or miss-annotation**

Bostan H(1),Colantuono C(1),Chiusano ML(1)

*(1) Deptartment of Agricultural Sciences, University of Naples Federico II, Via Università 100, 80055 Portici, Italy*

**Motivation:** The release of the Tomato genome sequence in 2010 by the "International Tomato Genome Sequencing Consortium" and of its preliminary gene annotation performed by the International Tomato Annotation Group (iTAG)[Tomato Consortium, 2012] was a precious moment for those involved in tomato research, but also for all plant scientist, since a new reference plant genome, highly relevant in food and agriculture sciences, was made available. However, the completion of a genome sequencing effort is never at an end, and the need of a reliable annotation is fundamental to fully exploit the acquired knowledge. Therefore in the presented work, we focused on a supplementary effort to screen the current available annotation and fix its possible limits in support of all users.
**Methods:** In this paper, the remapping of the iTAG predicted genes along the tomato genome was made independently using transcript to genome alignment software [Gremme, G., et al, 2005;Slater, G.and Birney, E., 2005]. Each mapped cDNA sequence was then compared and labelled according to the reference gene annotation. Further analysis on the resulting data revealed ambiguous or miss-located loci.
**Results:** Our analysis, based on the remapping of the iTAG predicted transcripts onto the tomato genome, revealed ambiguous, repeated and miss-predicted loci which have been traced to support the flourishing of investigation on Tomato now that the genome sequence is available.
**Contact email:** bostanict.net@gmail.com

**Integrating Genetic Variants within the i2b2 Framework: the NoSQL way.**

Gabetta M(1), Limongelli I(1), Rizzo E(1), Segagni D(2), Bellazzi R(1)

*(1) 1Biomedical Informatics Labs "Mario Stefanelli", Department of Electrical, Computer and Biomedical Engineering, University of Pavia, Italy (2) IRCCS Fondazione Salvatore Maugeri Hospital, Lab of Computer and Systems engineering for Clinical Research, Pavia, Italy*

**Motivation:** Next generation sequencing data generates large collection of variants that need to be properly organized and queried in databases and datawarehouses. This abstract describes the efforts made in order to integrate, leveraging NoSQL technologies, this information with the phenotypical data managed by the i2b2 framework. Sample variants are frequently collected in the Variant Calling Format (VCF) as the results of variant callers. Often these data pass through annotation step with tools like ANNOVAR. Therefore variants are integrated with information like: gene names, effects on primary protein structure, prediction scores and many others. Once genetic annotations are available, it is worth to integrate them inside a data warehouse optimized for querying phenotypical data, like i2b2, in order to link these two interconnected aspects. The reasons for choosing NoSQL technologies, and in particular Apache CouchDB, for managing the genetic data are many, among them we can mention: • They allow the data model to stay flexible (schemaless) so that applications are affected by possible changes in this model considerably less than their SQL-based counterparts. • The data are stored in a format that is very near to their original one. CouchDB, in particular, stores data into JSON documents (one document per variant in our case); this makes the information stored inside the database easily readable by humans (no need to combine data from different tables). Documents are also easily writable because the only entity to be modified is the document itself. • Many experiences conducted so far have proven that NoSQL technology is more suitable to scale when the data volume increases. However, NoSQL, and CouchDB in particular, have some limitations compared to traditional SQL databases, the most important being the fact that they do not provide a standard query language. Queries have to be pre-designed and are kept up to date by the database engine; this obviously makes the query process extremely fast but forces the application designer to foresee the axes on which the stored dataset will be queried.
**Methods:** The i2b2 extension we have developed so far is composed by two main parts: the first is the pipeline to populate the DB, which starting from VCF files, runs ANNOVAR, creates the JSONs with ANNOVAR's output and finally stores these files in CouchDB. The JSON structure is based on an object model specifically designed for this task. The second part is an i2b2 Cell, called NoSQL-NGS Cell, which, along with its plugin for the i2b2 Webclient, allows exploring the variants associated with a Patient Set previously achieved with the i2b2 query process. Despite CouchDB can be queried with REST methods, so that the Webclient could also communicate directly with the database, the presence of a mediator (i.e. the NoSQL-NGS Cell) has proven to be essential because not all the possible client-side queries can be managed with a single query on CouchDB and the assignment to a browser-hosted client of the (potentially) demanding task of aggregating partial results is, in general, a bad practice.
**Results:** To date the system has been preliminarily tested on standard desktop machine: ranging from a small set of mutations up to the equivalent of 25 exomes (about 550.000 variants), while the set up of the environment scales linearly with the volume of data managed, the query time remains almost instantaneous (<1sec). The whole system has been also deployed on the Amazon Web Services (AWS) environment. Preliminary tests confirm

our expectations: once the querying directions have been set, the querying performance and, accordingly, the user experience are very promising. The system was tested on 55 exome variant sets (about 22.000 variants each) from 1000 Genome Project that were uploaded and annotated in less than 1 hour and half, within 3 pre-designed query data generation time. This allowed query variants for genomic interval, exonic function (missense, nonsense etc.) and Polyphen2 score in an average of 2-3 seconds.
**Contact email:** matteo.gabetta@unipv.it

**Annocript: a flexible pipeline for transcriptome annotation also capable to identify noncoding transcripts**

Musacchia F(1), Basu S(1), Salvemini M(2), Sanges R(1)

*(1) Laboratory of Animal Physiology and Evolution, Stazione Zoologica "Anton Dohrn", Napoli*
*(2) Department of Biology, University of Naples "Federico II", Napoli*

**Motivation:** Transcriptome sequencing (RNA-seq) promises to give a comprehensive view of whole transcriptomes, allowing the collection and the study of all the RNA isoforms produced in a given organism, tissue or cell even in the absence of a reference genome
**Methods:** The pipeline was developed using Perl, MySQL and R and allows the creation of a comprehensive user-friendly table containing all the annotations produced for each transcript. The proteins most similar to the transcripts are given by the BLASTX
**Results:** Currently, Annocript is a flexible stand-alone pipeline, highly configurable by the user, capable to create a database with state-of-the-art annotations. It generates an output containing all the collected results and contains plug-in procedures to manage and data-mine results such as: non-coding sequences, putative translations, abundance of GO classes, functional enrichment analyses, enzymes and domains composition. An HTML page can be created containing summary plots and descriptive statistics of a given transcriptome. GFF3 outputs allow the integration of the annotations in genome browsers and easy file exchange trough current standard formats and a GFF database permits to rapidly access results using bioinformatics API such as BIOPERL
**Contact email:** francesco.musacchia@szn.it
**Supplementary information:**References

**Assessing the quality of a reference genome model by paralog and single copy gene analyses: basic bioinformatics for intriguing results.**

Sangiovanni M(1), Vigilante A(2) and Chiusano ML(3)

*(1) Department of Electrical Engineering and Information Technology, University of Naples Federico II, Napoli, Italy (2) UCL Genetics Institute, Department of Genetics, Evolution and Environment, University College London, UK (3) Department of Agricultural Sciences, University of Naples Federico II, Napoli, Italy*

**Motivation:** With the sequencing technologies becoming much cheaper and powerful, the need for reliable annotation tools and references for comparative genomics is even more striking. Nevertheless, reference genomes show still not clarified complexity in terms of the events that shaped their organization. As an example, the Arabidopsis thaliana genome, published in 2000, is considered the reference model in Plant sciences. In principle, it should be exhaustively annotated and well understood in terms of its evolutionary history. However, this genome underwent ancient whole genome duplications, followed by gene reduction, diploidization events and extended rearrangements, which relocated and split up the retained portions. This intricacy makes the identification of paralogs and single copy genes a mandatory step to understand the organization and evolution of the A. thaliana genome, and to improve its exploitation as a reference for comparative genomics in Plants. The intrinsic genome complexity of A. thaliana as well as the lack of a reference bioinformatics pipeline make the exhaustive identification of paralogs and singleton genes still a daunting task in genomics.
**Methods:** 0
**Results:** We describe here a complete computational strategy to detect both duplicated and single copy genes in a genome, discussing all the methodological issues that may strongly affect the results, their quality and their reliability. This approach was used to analyze the organization of Arabidopsis nuclear protein coding genes and, besides classifying computationally defined paralogs into networks and single copy genes into different classes, it unraveled further intriguing aspects concerning the quality of the genome annotation and the gene relationships in this reference plant species. Since our results on Arabidopsis may be useful for Plant comparative genomics and genome functional analyses, we make them accessible to the scientific community through the web site pARsi (paralogs and singleton genes browser for Arabidopsis).
**Contact email:** chiusano@unina.it

Poster: Genomics, Transcriptomics and NGS – GTN8

**Bioinformatics analysis of NGS data for viral pathogenesis factors identification.**

Catalano D(1) Cillo F(2) Tecce T(3) Finetti-Sialer M.M(1)

*(1)CNR-Istituto di Bioscienze e Biorisorse Consiglio Nazionale delle Ricerche, Via Amendola 165/A 70126, Bari, Italy (3)Università degli studi di Bari Dipartimento di Bioscienze, Biotecnologie e Biofarmaceutica, Via Orabona 70126, Bari Italy (2)CNR-Istituto di Virologia Vegetale, Consiglio Nazionale delle Ricerche, Via Amendola 165/A 70126, Bari, Italy*

**Motivation:** Potato virus Y (PVY) is a severe plant pathogen that has a worldwide distribution. Its host range includes important solanaceous crops such as tobacco, potato, tomato and pepper. PVY presents three main strains (PVYO, PVYN and PVYC) and several recombinant forms among different strains have been described. PVY consists of a single-stranded positive sense RNA genome of about 9800 bases in length. In the genome the terminal regions are untranslated (UTR) and between them there is a single large open reading frame coding for an unique polyprotein. During the infection the polyprotein is cleaved into ten products by three virus-encoded proteases. Recent studies revealed a high level of protein and nucleotide sequence conservation among 77 PVY genomes. In particular, two strain, PVY-SON41 and PVYc-to, showed 91 % and 95 % of genomic and protein conservation, respectively. However, when tomato (Solanum lycopersicum cv. UC82) was infected by these two isolates the disease phenotype appeared extremely different. RNA silencing, or post-transcriptional gene silencing (PTGS) is a conserved mechanism present in a broad range of eukaryotes, acting in plants as an immune antiviral system. Successful viral infections is often associated to suppression or evasion of the host induced silencing response. Small interfering RNAs (siRNAs) accumulate in plants infected by viruses and provide specificity to this RNA-mediated immune system. Taking into account that RNA silencing occurs in plants after perfect or nearly perfect siRNA- mRNA matching, we hypothesized that different viral sRNA produced by the two different PVY strains could have different gene targets, and that this difference originated the different symptoms observed in tomato plants. For this purpose, aim of the present study was the bioinformatic analysis of the two PVY strains genomes and the small RNA generated after infection, in order to identify the different regions responsible for the disease phenotypes observed.
**Methods:** A Bio-informatic pipeline was used to extract the complete datasets of 21-mers generated from the full genomes of PVYc-to and PVY-SON41 strains. The procedure scans the entire genomic RNA sequence shifted by one base at the time and generates two complete 21-mer datasets, one for each viral isolate. The two datasets were loaded in the MySQL database and compared by a Smith–Waterman algorithm embedded in a Perl script procedure to obtain the perfect match and dissimilar 21-mers in the viral genomes isolates. Next generation sequencing (NGS) data of a small RNA library extracted from infected tomato plants challenged with both PVY strains were mapped on the viral genomes by Bowtie. The Z-score was employed to identify the occurrence of PVYc-to 21-mers statistically significant and used for target analysis on tomato transcriptome (Solgenomics ITAG2.3), by BLAST in a first step, followed with RNAhybrid analyses for the minimum free energy calculation.
**Results:** Genomic regions of PVY isolates were explored for their potential to express virus-derived small RNAs (vsiRNA) that could suppress accumulation of host mRNAs, on the base of perfect or quasi-perfect sequence complementarity, which in turn would lead theoretically to dysfunctional biological processes thus explaining different disease phenotypes. Despite the high nucleotide sequence identity shared by the two viral genomes

(> 91%), the 21-mer datasets produced a high degree of variation. The isolate-specific putative mRNA targets of vsiRNA in tomato were 379 and 444 for PVYc-to and SON41 respectively. The NGS data analysis showed than PVYc-to vsiRNA are most abundant than SON41, vsiRNA induced by virus infection was predominantly 21 bp length, while most abundant endogenous sRNA was 24bp length. Statistical analysis of NGS data suggested that several viral region of PVYc-to ("hot spots") are over-represented in the vsiRNA datasets. Putative vsiRNAs within these regions could be involved in RNA silencing of transcription factors genes with a known role concerning leaf development and symmetry. These genes, if differentially modulated during the PVY infections, may cause the leaf malformations observed as major differential symptoms between the two isolates. To validate the bioinformatic approach, target genes with highly significant complementarity were selected from the in silico data for experimental assays in tomato plants mechanically infected by the two PVY isolates. Some host transcription factors active in vegetative and flower development, and leaf morphogenesis were selected among the PVYc-to vsiRNAs possible targets. Assays performed on inoculated plants showed a down regulation of FBP2 and ANAC74 that are of particular interest because they fall into the “hot spots” identified by the in silico analyses. This is the first of a series of assays needed to screen possible siRNAs whose interaction with the host could give rise to the characteristic phenotypes of the PVY isolates under study.

**Contact email:** domenico.catalano@ibbr.cnr.it

**Whole genome sequencing of disease and carriage isolates of Non-Typeable Haemophilus influenzae identifies discrete population structure.**

De Chiara M(1), Hood D (2,3), Muzzi A (1), Pickard D (4), Pizza M (1), Dougan G (4), Rappuoli R (1), Moxon R (3), Soriani M (1), and Donati C (1,5)

*(1) Novartis Vaccines and Diagnostics, via Fiorentina 1, 53100 Siena Italy (2) Nuffield department of Clinical Medicine, University of Oxford, John Radcliffe Hospital, Headington, Oxford, OX3 9DU, UK (3) Molecular Genetics Unit, Medical Research Council Harwell, Oxfordshire, OX11 0RD, UK (4) Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK. (5) School of Pathology and Laboratory Medicine,The University of Western Australia (M504), 35 Stirling Highway, CRAWLEY WA 6009, Australia (6) University of Oxford, Department of Paediatrics, Medical Sciences Division, John Radcliffe Hospital, Headington, Oxford, OX3 9DU, UK (7) Department of Computational Biology, Research and Innovation Centre, Fondazione Edmund Mach, Via E. Mach 1, 38010 San Michele all'Adige, Trento, Italy*

**Motivation:** Despite numerous successful preclinical studies, the feasibility of developing an effective and broadly protective vaccine against those isolates of Haemophilus influenzae that do not express a polysaccharide capsule (Non-Typeable Haemophilus influenzae – NTHi) still remains a challenge. One of the main hurdles lies in the genetic diversity of the species, which renders extremely difficult the identification of candidate antigens able to protect against heterologous strains.
**Methods:** We sequenced the genomes of a collection of diverse isolates of NTHi, representative of both carriage and disease and of the diversity of the natural population.Using Discrete Analysis of Principal Component we analyzed of the distribution of polymorphic sites in the core genome and of the accessory genome to characterize the genetic variability and assess whether a population structure of NTHi could be defined.
**Results:** We found six defined distinct evolutionary clades which supported a predominantly clonal evolution of NTHi, with the majority of genetic information transmitted vertically within lineages. A correlation between the population structure and the presence of selected surface-associated proteins and lipooligosaccharide structure, known to contribute to virulence, was found. This high-resolution, genome-based population structure of NTHi provides the foundation to obtain a better understanding of NTHi adaptation to the host, as well as its commensal and virulence behavior, that could facilitate intervention strategies against disease caused by this important human pathogen.
**Contact email:** matteo.de_chiara@novartis.com

**Exploration of hormone-independent Estrogen Receptor alpha activity by integrative and comparative approaches**

Ferrero G (1,3), Caizzi L (1,2,5), Miano V (1,2), Balbo G (1,4), Cordero F (1,4), De Bortoli M (1,3)

*(1) Center for Molecular Systems Biology, University of Turin, Orbassano, Italy; (2) Bioindustry Park Silvano Fumero, Colleretto Giacosa, Turin, Italy; (3) Dept. of Clinical and Biological Sciences, University of Turin, Orbassano, Italy; (4) Dept. of Computer Science, University of Turin, Turin, Italy; (5) Dept. of Gene Regulation and Stem Cells, CRG Center for Genomic Regulation, Barcelona, Spain.*

**Motivation:** In last years, many high-throughput (HT) experiments have been conducted to understand gene expression. However, integrative computational strategies able to organize and join several sources of knowledge are still missing. These approaches can enhance radically the amount of information obtainable from a HT experiment interfacing it with heterogeneous biological contexts and the other layers of gene expression regulation. Although Estrogen Receptor α (ERα) is a hormone-activated transcription factor, evidence of hormone-independent activity of ERα (apoERα) is increasing. Constitutive functions of ERα that discern estrogenic signaling, poses new questions on roles of this factor in physiological and pathological contexts, including breast cancer. We performed an analysis of ChIP-Seq and RNA-Seq in hormone deprived MCF-7 cells, transfected with ERα siRNA (siERα) and control [1]. In relation with these data we developed an integrative analysis considering several data stored in ENCODE database, including chromatin accessibility and epigenetic marks. Moreover, we compare our results with respect to public ChIP-Seq and expression data of multiple grown and treatment conditions of breast cancer cell lines and primary samples

**Methods:** ChIP-Seq peak calling was performed on IgG samples. For significant peaks (p-value <1E-5), we computed a second p-value compared to siERα sample. RNA-Seq data was aligned with Tophat and differential test performed with RegionMiner, DESeq and EdgeR. Public ChIP-Seq datasets were downloaded from GEO, Cistrome or author materials and overlapped with Galaxy and GSC. Public expression datasets were analyzed with GEO2R while gene-set analysis performed with GSEA. GREAT was used to functionally annotated each dataset and GOSemSim to perform the semantic similarity analysis. Motif discovery and affinity analysis were performed with MEME and TRAP algorithms, respectively.

**Results:** ChIP-Seq and RNA-Seq experiments resulted in 4,232 apoERα peaks enriched on IgG (aERBS) and 912 apoERα regulated genes. We compared these results with those reported in studies of 17β-estradiol (E2)- and antiestrogen-treated or long term hormone deprived (LTED) models. In general, aERBS are occupied by ERα (> 50% of regions) in 75% of the conditions that we compared; however, only a marginal overlap was observed considering the gene expression data. For each condition, functional peculiarities emerged in the semantic similarity analysis of enriched Gene Ontology terms for each ChIP-Seq dataset. Notably, aERBS mapped closer to development related genes, while E2 sets to genes involved in the lipid metabolism. From these observations, we defined a model for genomic occupancy consensus of each biological condition considered. These models were compared each other taking into account the genomic distribution and the degree of overlap with binding of other transcription factors, co-regulators and histone modifications. FoxA1, Gata3, Ap2γ, p300 and CBP were predicted to co-occupy aERBS suggesting the architecture of the network of factors cooperating with apoERα on maintenance of basal transcription.

Furthermore, even in ethanol treated cells peaks of DNaseI-HS, H3K4me1 and H3K27ac overlapped with aERBS. Finally, when we sort peaks considering their p-values of aERBS over siERα sample, the most significant ones were show to be enriched in conserved full-Estrogen Receptor Element (fERE) motifs and higher binding affinity. Notably, these aERBS mapped closer to epithelial development genes, E-cadherin signaling related networks and genes down-regulated upon apoERα silencing. By integrating our ChIP-Seq and expression data with those already published , it is possible obtain a global view of the dynamics at the basis of regulation mechanisms. In that way, we were able to compare basal activity of ERα with multiple contexts of genomic occupancy and gene expression regulation. This strategy allowed us to isolate peculiarities in apoERα activity and we proposed it also in investigation of other complex biological questions.

**Contact email:** giulioferrero1@gmail.com

**A computational workflow to study lentiviral integrations in the human genome using a targeted genomic sequencing approach**

Gandolfi F(1), Moiani A(1),Vega-Czarny N(2), Mavilio F(1), Gyapay G(2).

*(1)Genethon, Evry, France (2)CEA-GENOSCOPE-Centre National de Séquençage, Evry 91057, France*

**Motivation:** Like other retroviruses, Lentiviruses can stably integrate in the host genome through reverse-transcription of genomic RNA and insertion of double-strand cDNA mediated by the viral integrase. Due to this peculiar characteristic, design of efficient lentivirus-based gene transfer vectors is to date one of the most active fields in gene therapy. Several studies demonstrated that Lentiviruses integrate non-randomly into the mammalian genome, with preference for transcription units. Hence, clinical application of these vectors poses risks due to their potential to interfere with normal gene expression at the post-transcriptional level. Despite these efforts, molecular mechanisms of lentiviral vector integration still remain poorly understood. Thus, analysis of its integration patterns in clinically relevant cells is important to better comprehend basic retro-virology and minimize the risk in the current vector-based gene-delivery systems. Next Generation Sequencing technologies have made it possible to deeply investigate the interactions between retroviruses and the host cell genome, allowing whole-genome analysis of the integration profile in a deeper manner. Therefore, the demand for an appropriate bioinformatic pipeline that can automatically predict and annotate vector integration sites is rising. However, bioinformatic analysis of NGS-data poses a number of technical issues (memory usage, quality of the alignment, multiple matches) that have to be taken into account. These factors make the definition of appropriate computational strategies not trivial and can affect the accuracy in the precise prediction of the insertion positions.
**Methods:** In our work, we analyzed integration site distribution on a pool of human hematopoietic progenitor cells (CD34+) infected with an HIV-derived lentiviral vector. To retrieve integration sites across the genome we performed a sequence double-capture technique using a probe designed on the LTR sequence of the virus. Randomly fragmented DNA were then captured, PCR-amplified and sequenced on an Illumina® MiSeq platform using paired-end sequencing technology. Sequencing data were processed by an automatized bioinformatic pipeline specifically developed to analyze and localize the viral integration pattern. Our workflow is based on a double-mapping procedure where paired-end reads are firstly screened against the vector to filter out all the sequence parts representing the virus. Next, filtered sequences are aligned to the human reference genome and mapping information from both analyses are finally combined together to exactly infer junction site positions between the human and the viral sequence. To test our strategy, results were compared with an integration pattern of about 31000 integration sites experimentally identified in the same CD34+ cells pool using a restriction enzyme-based technique for library preparation coupled with Roche454 single-end sequencing. The same pipeline was also used to annotate integration site distributions in order to better characterize genomic regions and features that can be distinctive of HIV insertional events. Finally, statistically rigorous definition of integration clusters was also adopted in order to identify and characterize high-density regions of insertional events.
**Results:** Sequencing data of two different paired-end libraries from the same cell population were processed through the pipeline in parallel manner. Through these two analyses, we were able to precisely map 87992 and 115952 junction-points between the virus and the

genome, corresponding respectively to 4894 and 4002 unique integration sites. The two lists defined together a final set of 7825 unique integration events. Genome-wide analysis of integration sites density distribution showed high concordance when compared to the corresponding profile obtained from the Roche454 single-end sequencing, thus supporting the feasibility and the efficiency of the approach. Annotation analysis of integration sites allowed us to identify a number of genes and coding regions that could represent preferential targets of the HIV-integration complex. Interestingly, genomic annotation of integration sites identified in the Roche454-sequencing study and in our analysis showed the same pattern of integration, confirming the tendency of HIV-derived vectors to integrate inside genes or in proximity of transcriptionally active regions. Furthermore, clustering analysis allowed us to identify 654 HIV-integration clusters. Among them, a significant fraction (63%) showed physical overlap with those mapped in the single-end sequencing study. Finally, new and previously identified clusters together with information on significant HIV-target genes and distinctive epigenetic signatures allowed us to deeply investigate genomic factors that define the mechanisms of lentiviral integration in human cells.
**Contact email:** fgandolfi@genethon.fr

**Defining nucleosome distribution and occupancy in mammalian cells with a reduced histone content**

Gatti Elena (1), Barozzi Iros (2), De Toma Ilario (1), Natoli Gioacchino (2), Bianchi Marco (1)(3), Agresti Alessandra (3)

*(1) Università Vita-Salute San Raffaele, Milan, Italy (2) Department of Experimental Oncology, European Institute of Oncology, Milan, Italy (3) Division of Genetics and Cell Biology, San Raffaele Research Institute, Milan, Italy*

**Motivation:** Mammalian cells lacking HMGB1, abundant non-histone chromatin protein, have 20-30% less histones and therefore nucleosomes (1). HMGB1 is normally found in the nucleus where it acts as a chromatin-binding factor and plays multiple roles in DNA transaction (2-3). HMGB1 also promotes nucleosome deposition, and thus can be thought as a "DNA chaperone" (2). Also yeast cells, lacking the functional homologous of Hmgb1, nhp6A/B, show a similar reduction. Moreover, Hmgb1-/- MEFs, HeLa cells in which HMGB1 has been knockdown by siRNA and the yeast nhp6a/b mutant display a very similar phenotype and are more susceptible to DNA damage. In yeast cells, the reduction in histone content results in a non-homogeneous change of nucleosome occupancy without gross modification of nucleosome position (1). The strict correlation between histone content and nucleosome number allows us to use histone content as a proxy for nucleosome number. We asked whether this phenotype is maintained also in complex genomes such as mammals. At the best of our knowledge, the genome-wide mapping of individual nucleosomes at single-base pair resolution (5-6) has been carried out in model organisms (7-8) and in higher organism9 with the aim of looking at particular features of nucleosome positions. We aim to study the distribution of nucleosomes and the relative occupancy both at genome-wide level and at particular regulatory regions (i.e. TSSs, enhancers…) in cell that have lost 20% of their nucleosomes. A high-resolution genome-wide analysis will help to determine nucleosome distribution and occupancy. With nucleosome distribution we intend where nucleosomes are located with respect to the genomic DNA sequence. The term nucleosome occupancy indicates the amount of time each specific DNA site spends wrapped in a nucleosome.

**Methods:** To this aim, we performed a deep sequencing analysis of micrococcal nuclease (MNase) protected nucleosomal DNA from G1-synchronized wild-type (wt) and Hmgb1-/- (ko) mouse embryonic fibroblasts (MEFs). Mono-nucleosomal DNA was isolated from agarose gel as a single 150bp band and used for Illumina library construction. Paired-end sequencing of 101bp reads on the HiSeq2000 platform has been performed at Institute of Applied Genomics (Udine). Each sample generated ca. 320M filtered, uniquely aligned and properly paired sequence reads. DNA reads were mapped to the reference mouse genome (mm10) with Bwa, reporting unique hits. Since available software are not always well suitable for the analysis of mammalian genomes, both public tools (8) and custom-made in house Perl and R scripts have been developed for nucleosome positioning and for measurement of occupancy.

**Results:** We run a lane of HiSeq2000 for each MNase digested sample; 80% and 87% of raw reads are mapped and properly paired for wt and ko respectively. In total we sequenced 2.0 and 2.4Gbp of mouse genome in wt and ko and we obtained a mean coverage of 17x and 20x respectively. We called the nucleosomes using DaNPOS8 and we produced heatmaps of nucleosome signals in 4kbp regions around TSSs of RefSeq genes. TSSs were also filtered on the basis of expression of genes detected through RNASeq experiment on MEFs. Heatmaps

ordered for nucleosome free region revealed phased +1, +2, +3 nucleosomes. Preliminary analysis on the differentially positioned nucleosomes between wt and ko and relative occupancy have been performed and we obtained indications of delocalization and fuzziness, two important properties of nucleosomes. In the next future we will verify the pattern of nucleosome positions in selected categories of sequences like promoters, splice junctions, repetitive sequences, telomeres, gene-rich and gene-poor regions. Challenging and appealing follow-up includes the correlation in wt and HMGB1 ko MEFs between transcriptional profiles determined through RNASeq analysis and the genome-wide nucleosome mapping in order to create an association between transcript abundance and alteration in chromatin structure. We thank IGA sequencing facility and Epigen funding. References 1.Celona B et al. PLoS Biol 9: e1001086 (2011). 2.Bianchi ME, Agresti A. Curr Opin Genet Dev 15: 496-506 (2005). 3.Stros M. Biochim Biophys Acta 1799 (2011). 4.Stillman DJ. Biochim Biophys Acta 1799: 175-180 (2011). 5.Zhang, Z. and Pugh, B.F. Cell, 144, 175-186 (2011). 6.Segal, E., et al. Nature, 442, 772-778 (2006). 7.Valouev, A., et al. Nature, 474, 516-520 (2011). 8.Chen, K et al. Genome research, 23(2), 341-351(2013).
**Contact email:** gatti.elena@hsr.it

**Improving miRNA-target prediction through tridimensional-structure modelling approach**

Leoni G. (1) Di Marino D. (1) Tramontano A. (1,2)

*(1) Department of Physics, Sapienza University, Rome, 00185, Italy (2) Istituto Pasteur—Fondazione Cenci Bolognetti, Sapienza University, Rome, 00185, Italy*

**Motivation:** miRNAs are short RNAs of 20-23 nucleotides able to regulate genome expression by binding specific target transcripts and promoting their degradation or their translational repression. The recognition of target by a miRNA occurs through a process that involves the interaction of miRNA and target with argonaute (AGO) proteins to form the RISC complex. Currently exist several computational methods for the prediction of miRNA interactions. Most of them analyse the sequence and the secondary structure of putative targets with the aim to recognize possible sites on transcripts that can be bound by miRNAs. The efficacy of currently used methods is limited by the still incomplete knowledge of rules governing the target recognition and by the enormous amount of nucleotidic sequences contained in transcriptome complementary to a specific miRNA sequence, resulting in high number of sites erroneously predicted as positive targets. Recently were characterized the crystallographic structures of AGO proteins providing new insights into the molecular mechanism of mRNA targeting by miRNA. Therefore, the development of a method able to describe and predict miRNA-targets including the information about the tertiary structure of miRNA-target complexes bound into AGO protein, could represent an additional tool that improve predictions made by existing algorithms.
**Methods:** Aim of this study is to analyse miRNA-targets binding interactions by modelling the tertiary structure of miRNA-target complexes fitted into AGO protein. An initial dataset of 137 positive and 137 negative experimentally validated, miRNA-target interactions were extracted from mirtarbase database. For each experimental evidence, the target 3'UTR was selected and the best putative target site was identified as the one with the best mirSVR score reported in microrna.org database. For each site, 1000 models of paired miRNA-target complexes were produced utilizing as template the partially solved duplex bound into the crystallographic structure of T. thermophilus AGO protein. The best model was defined by selecting the one with the minimal number of clashes with AGO protein and the maximum number of paired bases between miRNA and target. Finally the best complex was further minimized to estimate its capability to reach a local minimum of energy. All the final models were manually examined and from the total dataset were filtered 32 positive and 57 negative evidences with final best models detected as implausible because with a not reasonable conformation that severely overlaps to AGO backbone
**Results:** To better highlight the AGO contribution to duplex stabilization, we excluded from the analysis the models with a low number of interactions with AGO. Models of positive and negative dataset show different distributions of final energy (Wilcoxon test p-value =0.035) with positives showing energy lower than negatives. According to ROC analysis, the energy criterion is able to correctly identify the 63% of cases while other algorithms as PITA, Targetscan and mirSVR score are able to discriminate respectively only the 50%, the 40% and the 43% of cases. The obtained preliminary results raise the possibility to utilize energy estimates of the tertiary structure of miRNA-target complexes bound into AGO as a complementary tool to improve predictions made by current computational methods.
**Contact email:** anna.tramontano@uniroma1.it

**Investigating the potential of non-coding RNAs to regulate transcription through triple-helix formation**

Patavino C(1), Colantoni A(1), Pignotti D (1), Ferre'F(1), Helmer-Citterich M(1)

*(1) Centre for Molecular Bioinformatics, Department of Biology, University of Rome Tor Vergata, Via della Ricerca Scientifica snc, 00133 Rome, Italy*

**Motivation:** The ability of double-stranded DNA to form a triple-helical structure by hydrogen bonding with a third strand is well established. The triplex-formation involves a double-stranded nucleic acid such as duplex DNA and a single-stranded nucleic acid such as RNA. The 'triplex-forming oligonucleotide' (TFO) of the single-stranded nucleic acid binds to the triplex target sites (TTSs) according to specific rules (Hoogsteen hydrogen bonds). These constraints can be used to investigate the existence of putative TTS and TFO sequences pairs inside the genome. The potential participation of this complex in a variety of biological processes, like chromatin organization, DNA repair, transcriptional regulation and RNA processing, has been investigated in several studies to date [1]. As regards to transcriptional regulation, many research groups focused on the analysis of TTSs as components of putative regulatory sequences in cis and of TFOs localized on ncRNA molecules possibly acting in trans. Different computational studies analysed the abundance of putative TTS sequences in various genomes; these studies showed that TTSs are more abundant in promoter sequences than in transcribed or intergenic sequences in mouse and human genomes [2]. In the present study we have examined whether putative TTSs found in promoter regions are more likely to interact with TFO carried by ncRNA with respect to TTSs localized in intergenic regions. Furthermore, the propensity of some specific regions within the promoter to host putative TTS-TFO matches was investigated.
**Methods:** We obtained a set of promoters from UCSC [3], defined as regions of 2 kb upstream of the transcription start site of RefSeq [4] protein coding genes. The Ensembl [5] database was used to obtain a collection of ncRNA and intergenic regions sequences. The computational framework Triplexator [6] , which is able to predict putative TTSs in duplex nucleic acids, TFO sites in single-stranded sequences, and all the TTS-TFO pairs that satisfy the triplex-forming rules, was used in order to find putative unique TTSs inside the promoter and intergenic sequences, and to determine which of them was matching with at least one putative TFO carried by a ncRNA. A chi-squared test was performed to evaluate whether the difference between the percentage of matching TTSs in promoters vs intergenic regions was statistically significant and whether different regions of promoters show significant defference in the persentage of matching TTS-TFO pairs
**Results:** 37,06% of TTSs in promoters of human protein coding genes were found to have at least one match with a TFO carried by a ncRNA, yet a statistically not different percentage (36,04%) was observed for TTSs localized within intergenic regions. Hence, even if TTSs are more abundant in promoters, their localization within promoters is not associated with a greater probability to be bound by a ncRNA through the formation of a triple helix. However, after dividing promoters in four equally sized regions, we found that those TTSs that are closer to the transcription start sites match with putative TFOs with a higher frequency (40%) than other TTSs. Due to their proximity to the transcription start site, these TTSs are more likely to be involved in the regulation of gene expression. Additional controls are planned in order to validate these results, including a permutation of the ncRNA sequences. Furthermore, we are evaluating the evolutionary conservation of putative TTS/TFO pairs among different organisms.

**Contact email:** citterich@uniroma2.it
**Supplementary information:**References 1. Buske FA, Mattick JS, Bailey TL: Potential in vivo roles of nucleic acid triple-helices. RNA biology , 8:427–39. 2. Wu Q, Gaddis SS, MacLeod MC, Walborg EF, Thames HD, DiGiovanni J, Vasquez KM: High-affinity triplex-forming oligonucleotide target sequences in mammalian genomes. Molecular carcinogenesis 2007, 46:15–23. 3. Karolchik D, Barber GP, Casper J, Clawson H, Cline MS, Diekhans M, Dreszer TR, Fujita PA, Guruvadoo L, Haeussler M, Harte RA, Heitner S, Hinrichs AS, Learned K, Lee BT, Li CH, Raney BJ, Rhead B, Rosenbloom KR, Sloan CA, Speir ML, Zweig AS, Haussler D, Kuhn RM, Kent WJ: The UCSC Genome Browser database: 2014 update. Nucleic acids research 2013:gkt1168–. 4. Pruitt KD, Tatusova T, Maglott DR: NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic acids research 2005, 33:D501–4. 5. Hubbard T: The Ensembl genome database project. Nucleic Acids Research 2002, 30:38–41. 6. Buske FA, Bauer DC, Mattick JS, Bailey TL: Triplexator: detecting nucleic acid triple helices in genomic and transcriptomic data. Genome research 2012, 22:1372–81.

**INDIVIDUATION OF MODULES IN BIOLOGICAL NETWORKS BY MOST EFFICIENT SHORTEST PATHS AND COMMUNITY STRUCTURE**

Pepe D(1), Palluzzi F(2), Grassi M(1)

*(1) Department of Brain and Behavioral Sciences, University of Pavia, Pavia (2) Department of Electronics, Information and Bioengineering, Politecnico of Milano, Milano*

**Motivation:** Recently in biology and in medicine there was the passage from the reductionist approach, finalized to study of individual cellular components and their functions, to a network approach, based on the concept that it is possible to understand biological phenomena only by the characterization of interconnected patterns of relationships between the component involved. The philosophy of the last approach has ancient origins associated with one of the main Buddhist principles: all conditioned phenomena are radically interdependent and hence lack any kind of fixed or unchanging 'essence'. Biological networks are subjected to specific laws, as the small world phenomena, which affirms that there are relatively short path between any pair of nodes; the scale-free principle, with the consequence that there are few highly connected hubs; the local hypothesis i.e. the presence of modules, highly interlinked local regions in the network in which the components are involved in same biological processes. Considering these premises we propose an algorithm that starting from a list of genes and a network in which they are present, it is able to return a module through the most efficient shortest paths and the concept of community, subsets of vertices within which vertex–vertex connections are dense, but between which connections are less dense.
**Methods:** The algorithm can be conceptually divided in two parts: 1) computation of the most efficient shortest paths between every couple of nodes; 2) computation of the communities to which they belong. A measure of efficiency, proposed by Latora and Marchiori, is computed as the sum of the values of closeness centrality of a node, defined for the node n as the inverse of the average length of the shortest paths to/from all the other vertices in the graph. We used, differently, the measure of betweenness centrality, defined for a node n, as the ratio between the number of shortest paths between node i and node j in which the node n is involved and the total number of shortest paths between node i and node j. The criteria to choose the most efficient shortest paths consists in the maximize the sum of values of betweennes of the nodes involved in the shortest paths. For the individuation of the communities, we used the algorithm of Blondel et. al., based on the modularity measure and a hierarchial approach. In practice, initially it is created a community for each vertex. Then, for each subsequent step, vertices are re-assigned to communities in basis to the their contribution to modularity. The process ends when only a vertex is left or when it is not possible to improve further the modularity. To obtain the final module, we extract a subgraph composed by the nodes of interest resulting connected and by the nodes that are in the same communities. The analyses were performed with R.
**Results:** We applied the algorithm to find a disease module from the microarray experiment present in GEO with id GSE14580. The experiment was finalized to understand the mechanism of resistance to the anti-TNF-alpha drug, infliximab in patient affected by ulcerative colitis (UC). We compared two groups, patients affected treated with infliximab (8 samples) against control (6 samples). A differential analysis using a t-test with a Benjamini-Hochberg correction, allowed to individuate 1364 differentially expressed genes (DEGs). On these, a pathway analysis with the “Signaling Impact Pathway Analysis” (SPIA) revealed five perturbed KEGG pathways: cytokine-cytokine receptor interaction, cell adhesion molecules,

chemokine pathway, antigen processing and presentation, complement and coagulation cascades. On the DEGs in the perturbed pathways, we performed our analysis using as network the fusion of all KEGG pathways. We obtained a module of 402 nodes and 2961 edges. The goodness of the procedure was tested by a disease enrichment analysis based on terms in Disease Ontology (DO), that allowed to individuate which diseases are associated to the nodes/genes present in the module, and by a measure of semantic similarity of diseases, to verify how much the diseases enriched are similar to UC. The results were promising, considering that the enrichment and the measure of similarity revealed diseases associated with UC as autoimmune disease, peptic ulcer, esophagitis and side effects of infliximab as lupus erythematosus.

**Contact email:** danielepepe84@gmail.com

**Mitochondriome-Exome Wide Associations (MEWAs): a bioinformatics system for association studies between mitochondrial and nuclear variants**

Santorsola M(1,2*), Diroma MA(1*), Calabrese C(3), Clima R(3), Simone D(4,5), Guttà C(1), Gasparre G(3), Attimonelli M(1) *co-authors

*(1)Department of Biosciences, Biotechnologies and Biopharmaceutics, University of Bari, Bari 70126, Italy (2)Department of Science, University of Sannio, Benevento 82100, Italy (3)Department of Medical and Surgical Sciences, University of Bologna, Bologna 40138, Italy (4)Department of Biosciences, University of Milan, Milan 20133, Italy (5)Institute of Biomembranes and Bioenergetics (CNR) - Bari*

**Motivation:** The application of NGS sequencing techniques to molecular medicine has opened new scenarios of research providing the scientific community with large amounts of data. The largest fraction of high-throughput sequencing protocols is targeted to nuclear genome, totally neglecting the genetic material in mitochondria [1]. Nonetheless, DNA enrichment systems for exome sequencing (WES) allow a significant coverage of mitochondrial regions, even when designed for nuclear DNA exclusively [2,3]. Retrieving information from both nuclear and mitochondrial genomes offers the opportunity of a deep insight of their cross-talk [4] in healthy as in disease phenotypes (i.e. neurodegenerative pathologies, cancer etc.). To this aim here we present a protocol allowing the association of both nuclear and mitochondrial variants with a specific phenotype.
**Methods:** The protocol includes a first step of base calling, read alignment and variant calling on NGS and in particular WES data. The assembly of mitochondrial genomes is performed upon NumtS filtering and detection of heteroplasmic fractions, insertions and deletions (indels) [5]. Functional annotations are performed on both mitochondrial [5] and nuclear [6] variants identified thus allowing the prioritization of the most phenotipically relevant variants. The last step of the strategy is based on the application of statistical approaches finalized to the association process, which includes logistic regression [7,8,9].
**Results:** The designed protocol has been tested on WES data from the 1000 Genomes project, a population study whose samples were selected among healthy subjects. This preliminary analysis will allow the filtering out of associated polymorphisms. Further analyses are being applied on data available within the Sequence Read Archive (SRA), the database of Genotypes and Phenotypes (dbGaP) and The Cancer Genome Atlas (TCGA), with particular attention to colorectal cancer studies. Preliminary data regarding the association results will be presented. Associated variants statistically assessed will be annotated in an ad hoc designed catalogue.
**Contact email:** ma.santorsola@gmail.com; mariangeladiroma@gmail.com
**Supplementary information:**References [1] Pesole G, Allen JF, Lane N, Martin W, Rand DM, Schatz G, Saccone C. The neglected genome. EMBO Rep. 2012 Jun 1;13(6):473-4. doi: 10.1038/embor.2012.57. PubMed PMID: 22555611; PubMed Central PMCID: PMC3367242. [2] Picardi E, Pesole G: Mitochondrial genomes gleaned from human whole-exome sequencing. Nat Methods 2012, 9(6):523-524. [3] MA Diroma, C Calabrese, D Simone, M Santorsola, FM Calabrese, G Gasparre, M Attimonelli. Extraction and annotation of human mitochondrial genomes from 1000 Genomes Whole Exome Sequencing data. BMC Genomics, in press [4] Lane N. Evolution. The costs of breathing. Science. 2011 Oct 14;334(6053):184-5. doi: 10.1126/science.1214012. PubMed PMID: 21998376. [5] Calabrese C., Simone D., Diroma M.A., Santorsola M., Calabrese F.M., Gasparre G., Picardi E., Pesole G., Attimonelli M. MToolBox: a package for the analysis of human NGS mitochondrial data.

Abstract submitted to the Conference Proceedings of BITS 2013 [6] Wang K, Li M, Hakonarson H. ANNOVAR: Functional annotation of genetic variants from next-generation sequencing data Nucleic Acids Research, 38:e164, 2010 [7] Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ & Sham PC (2007) PLINK: a toolset for whole-genome association and population-based linkage analysis. American Journal of Human Genetics, 81. [8] B L Browning (2008) PRESTO: Rapid calculation of order statistic distributions and multiple-testing adjusted P-values via permutation for one and two-stage genetic association studies. BMC Bioinformatics 9:309. Available from http://www.biomedcentral.com/1471-2105/9/309. [9] Pahl R, Schäfer H. PERMORY: an LD-exploiting permutation test algorithm for powerful genome-wide association testing. Bioinformatics. 2010 Sep 1;26(17):2093-100. doi: 10.1093/bioinformatics/btq399. Epub 2010 Jul 6. PubMed PMID: 20605926

**Correction algorithm of pooled sequenced reads**

Cordero F(1), Beccuti M(2), Duma D(2), Ciardo G(2), Close TJ(3), Lonardi S(2)

*(1) Department of Computer Science, University di Torino, Torino, Italy (2) Department of Computer Science and Eng., University of California, Riverside, USA (3) Department of Botany & Plant Sciences, University of California, Riverside, USA*

**Motivation:** We recently showed how to take advantage of combinatorial pooling for clone-by-clone de-novo genome sequencing [1]. In our sequencing protocol, subsets of non-redundant genome-tiling BACs are chosen to form intersecting pools, then groups of pools are sequenced via standard multiplexing (barcoding). Sequenced reads can be assigned to specific BACs by relying on the structure of the pooling design: since the identity of each BAC is encoded within the pooling pattern, the identity of each read is similarly encoded within the pattern of pools in which it occurs. Finally, BACs are assembled individually, simplifying the problem of resolving genome-wide repetitive sequences. An advantage of our sequencing protocol is the potential to correct sequencing errors. This study investigates to what extent our protocol enables such error correction. Due to obvious needs in high-throughput sequencing technology applications, including de-novo assembly, the problem of correcting sequencing errors in short reads has been the object of intense research.
**Methods:** We deal with nucleotide reads. Given a read r, a k-mer alfa is any substring of r of length k. Also, a BAC (clone) is a 100–150kb fragment of the target genome replicated in E. coli. After identifying the BACs to be sequenced [1], we pool them according to a scheme that allows us to decode (assign) sequenced reads back to their corresponding BACs. The set of L pools, called signature, defines a unique pooling pattern for each variable, and can be used to retrieve its identity. In our algorithm, any k-mer occurring in a number of pools smaller than L is considered erroneous and we will attempt to correct it. This approach assumes that our DNA samples are covered by a sufficient number of reads, so that each genomic location is covered by several correct k-mers (mixed with a few corrupted k-mers). In practice, the depth of sequencing is not uniform and may vary significantly along the genome. When it is particularly low, it is possible for a correct k-mer to appear in fewer than L pools. For each pool p, we slide a window of size k on each read r of length at least k. We then build function poolcount, stored it in a hash table, where poolcount(alfa,p) is the number of times k-mer alfa (or its reverse complement, we choose the smallest of the two in lexicographic order as the representative) appears in pool p. The advantage of the pooling design in our sequencing protocol is that any k-mer alfa such that pools(alfa) = {p} (i.e., alfa occurs in just one pool), and even more so if poolcount(alfa, p) = 1, is likely to contain sequencing error(s). While we focus on these low-frequency k-mers to find errors, it turns out that they are responsible for the large majority of the entries in the hash table thus negatively affecting memory requirements. The solution is to not include them in the hash table when processing the reads to save space. Then, during the error correction step, any k-mer not found in the hash table is assumed to be incorrect. Specifically, the hash table does not contain k-mers appearing in fewer than l pools, where l is a user-defined parameter. At the other end, a k-mer alfa is deemed repetitive if |pools(alfa)| is major than h, where h is user-defined parameter. After building the hash table of k-mers appearing in at least l pools, we scan the reads again. If a k-mer alfa belonging to the current read r is not in the hash table, we attempt to correct it by changing the nucleotide at either its first or its last position into the other three possible nucleotides. The three variants are searched in the hash table:

if only one is present, then it is the correct version of alfa, assuming that alfa contains only one error. If multiple variants of alfa are instead found in the hash table, choosing among them becomes more difficult. Therefore, we need to analyze the entire read r to which alfa belongs. For any correct k-mer beta belonging to r, we expect pools(beta) to equal one BAC signature or the union of up to d BAC signatures.
**Results:** We tested our error correction method on short reads from the rice genome which is a fully sequenced 390Mb genome. We started from an MTP of 3,827 BACs selected from a real physical map library of 22,474 BACs. The average BAC length in the MTP was about 150kB. Overall, the BACs in the MTP spanned 91% of the rice genome. We pooled a subset of 2,197 of these BACs into 91 pools according to the parameters defined above. We compare the performance of our correction algorithm against the state-of-the-art error-correction method Racer. The high error correction accuracy achieved by our method illustrates another strong benefit of combinatorial-pooling-based sequencing, in addition to the cost savings due to simultaneously sequencing thousands of DNA samples.
**Contact email:** fcordero@di.unito.it

**Expression profile evaluation on mesenchymal stem cells by RNA-Seq and miRNA experiments**

Bergantino F(1), De Luca A(2), Roma C(1), Fenizia F(1), Gallo M(2), Frezzetti D(2), Costantini S(3), Normanno N(2)

*(1) Pharmacogenomic Laboratory, INT-Fondazione Pascale-Centro di Ricerche Oncologiche di Mercogliano (CROM), Mercogliano (AV), Italy (2) Cell Biology and Biotherapy Unit, INT-Fondazione Pascale, Naples, Italy (3) Drug Design and Systems Biology Laboratory, INT-Fondazione Pascale- Centro di Ricerche Oncologiche di Mercogliano (CROM), Mercogliano (AV), Italy*

**Motivation:** Next-Generation Sequencing (NGS) is a new sequencing technology that has improved the sequencing efficiency of biological molecules in comparison to traditional Sanger method, allowing to sequence up to a billion bases in a short time. NGS approaches can detect every possible type of genetic alteration such as gene polymorphisms, mutations, inversions, rearrangements and gene copy number alteration, and provide information on the cancer genome and transcriptome useful for understanding the pathogenesis of the diseases and improving the molecular diagnosis, the cancer classification, the identification of prognostic factors and the definition of therapeutic targets. In this context, the use of methods of NGS is favoring the development of personalized medicine for cancer treatment. In this work we have used bioinformatics approaches to analyze the data produced by NGS platform obtained on mesenchymal stem cells (MSC) treated with transforming growth factor-alpha (TGF-α) to modulate the expression of secreted factors capable to promote the proliferation and the migration of neoplastic cells within the tumor microenvironment.
**Methods:** RNA-Seq and miRNA experiments were performed on control and treated MSCs using SOLiD5500xl. Firstly, the good quality of the data was evaluated by Galaxy platform. Secondly, we aligned the reads to a reference genome (i.e. hg19) and we normalized the number of observed counts for the length of the element and the number of mapped reads, namely RPKM (Reads Per Kilobase per Million of mapped reads) with LifeScope™ Genomic Analysis Software, already implemented on SOLiD5500xl. Then, the differentially gene expression in untreated and treated samples was evaluated in terms of fold induction by Deseq tool in R package. Moreover, we performed on the differentially expressed genes a GO analysis and pathway evaluation by DAVID program to evaluate specific pathways and potential targets for therapeutic intervention. The same procedure was used to identify miRNAs that are differentially regulated after TGF-a treatment. Finally the presence of splicing isoforms was evaluated by LifeScope™.
**Results:** Experiments were repeated in quadruplicate and resulted very reproducible. Concerning RNA-seq experiment, the differential gene expression in untreated and treated samples showed that 10.068 genes were significantly differentially expressed. In details, 673 genes resulted to have a fold induction (FI)<0,5 and 967 FI>1,5. For these genes we found some significant GO terms for molecular function and two pathways (Cytokine-cytokine receptor interaction, Jak-STAT signaling pathway). Moreover, pathway analysis revealed that there were some genes of the EGF pathway suggest that EGFR stimulation in MSCs leads to activation of several signaling pathways that might be involved in tumor progression. In parallel, from the analysis of miRNA experiment, it is resulted that 799 miRNA were differentially expressed between untreated and treated cells. Focusing the attention on VEGFA, already object of study in our laboratory [1], we are evaluating the miRNA targets for VEGFA obtained using MirWalk algorithm. Finally, we evaluated the entire spliceosome

obtained by RNA-seq experiment and analysed in details the splicing isoforms for VEGFA evidencing that there were six isoforms, of which four present both in untreated and treated samples even if in different expression, and two present only in the treated samples. Therefore we can suppose that these isoforms could be used as new anticancer therapeutic targets.

**Contact email:** fbergantino@yahoo.it

**Supplementary information:**References [1] De Luca A, Lamura L, Gallo M, Maffia V, Normanno N. Mesenchymal stem cell-derived interleukin-6 and vascular endothelial growth factor promote breast cancer cell migration. J Cell Biochem. 2012 Nov;113(11):3363-70.

**RAP: RNA-Seq Analysis Pipeline, a web server for RNA-Seq analysis**

D'Antonio M. (1,2), D'Onorio De Meo P. (2), Picardi E. (1), Calogero R. (3), Castrignanò T. (2), Pesole G. (1,4,5)

*(1) Dipartimento di Bioscienze, Biotecnologie e Scienze Farmacologiche, University of Bari, Bari, Italy (2) CINECA - Consorzio interuniversitario per il calcolo automatico, Bologna, Italy (3) University of Turin, Turin, Italy (4) Istituto di Biomembrane e Bioenergetica, Consiglio Nazionale delle Ricerche, Bari, Italy (5) Center of Excellence in Genomics (CEGBA), Bari, Italy*

**Motivation:** In the last few years NGS (Next Generation Sequencing) platforms have greatly increased the opportunities of analysis with possibilities unimaginable a short time ago. NGS can be profitably used to investigate the gene expression process, estimating the nature and quantity of expressed mRNAs to determine which genes and isoforms are expressed and their levels [1]. The ability to identity and quantify expressed genes, under specific conditions, provides precious biological information even though doesn't explain the mechanism underling gene expression. Indeed, final products may be not easily predictable due to post-transcriptional phenomena as alternative splicing [2]. To better understand this process, a complete analysis pipeline is required, to exploit RNA-Seq data from many different points of view and adopting several computational strategies.
**Methods:** The analysis of RNA-Seq data can be challenging due to large computational and storage resources requirements. Moreover handling bioinformatics tools, data formats and command line applications can be tough for biologists without expertise in bioinformatics and IT. RAP (RNA-Seq Analysis Pipeline) is a web application implementing a fully automated analysis workflow, designed to integrate in-house developed scripts as well as open source analysis tools into one single pipeline. Using RAP the user can perform a complete NGS analysis without any specific technical competence nor directly using computational resources. Moreover RAP also offers an interface for results management, allowing the user to browser and filter the massive amount of data obtained from typical RNA-Seq experiments. RAP takes as input short-read datasets produced by Illumina sequencing platforms and several standard file formats (FASTQ, SRA, BAM and compressed formats). RAP pipeline is designed to analyze the data through a series of phases, each of them focused on a specific task. Following a phase of quality control and reads filtering, a main backbone is executed to align high quality reads to the reference genome and transcriptome (mapping both spliced and unspliced reads). Mapped reads are then assembled to reconstruct the expressed isoforms and to estimate the relative abundance at both gene and transcript level. Unmapped reads are aligned to a splice junctions library obtained from ASPicDB [3], a database of reliable annotations of alternative splicing patterns. This library contains both known and potential splice junctions, the latter obtained through a selective procedure of consecutive exon skipping. Cassette exons are also investigated adopting a specific analysis module to identify skipping events and calculate inclusion percentages. Reads still unmapped to the genome, transcriptome and junctions are analyzed to identify polyadenylation sites. PolyA tags are extracted, trimmed and aligned again to the genome. Furthermore data are analyzed in order to detect chimeric transcripts. Each phase of analysis can be customized configuring parameters or skipping branches not interesting to the user. While the pipeline is running the user can monitor the status of each step and access to intermediate files with preview and download facilities. The final results can be browsed directly through the web interface to query, filter and sort the results. Graphical results are also provided to help the user with a better interpretation of results. After the

completion of the main analysis, several operations can be executed. Differential expression can be calculated starting from an analysis result, selecting the lanes of interest. In case of biological replicates the user can assign a group to every set of replicates. Two different DE procedures are implemented. The user can also analyze differential usage of exon skipping events or polyadenylation sites.

**Results:** The main purpose of RAP is to provide to users a fully comprehensive RNA-Seq pipeline without any installation or IT requirement. The web interface provides an easy and intuitive access for data submission and results browsing. Users can access through RAP to several RNA-Seq algorithms, each integrated with other to maximize the overall quality and quantity of results. To better understand e manage the large amount of data and maximize the biological information extracted from them, the user can sort and restrict the number of final results filtering data by customizable thresholds, facilitating the identification of significant data.

**Contact email:** m.dantonio@cineca.it

**Supplementary information:**References 1. RNA-Seq: a revolutionary tool for transcriptomics. Wang Z, Gerstein M, Snyder M. January 2009, Nat Rev Genet. 2. Alternative Isoform Regulation in Human Tissue Transcriptomes. Wang E, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, Kingsmore SF, Schroth GP, Burge CB. November 2008, Nature. 3. ASPicDB: a database of annotated transcript and protein variants generated by alternative splicing. Martelli PL, D'Antonio M, Bonizzoni P, Castrignanò T, D'Erchia AM, D'Onorio De Meo P, Fariselli P, Finelli M, Licciulli F, Mangiulli M, Mignone F, Pavesi G, Picardi E, Rizzi R, Rossi I, Valletti A, Zauli A, Zambelli F, Casadio R, Pesole G. January 2011, Nucleic Acids Res (Database issue).

**miRNA profiling in Alzheimer's disease by next-generation sequencing**

Annese A(1), Manzari C(1), Chiara M(2), Zaffaroni G(2), Picardi E(1), D'Erchia AM(1), Horner D(2), Pesole G (1,3)

*1) Dipartimento di Bioscienze, Biotecnologie e Biofarmaceutica, Università degli Studi di Bari Aldo Moro, Via Orabona 4, 70126 Bari 2) Dipartimento di Bioscienze, Università degli Studi di Milano, Via Celoria 26, 20133 Milano, Italia 3) Istituto di Biomembrane e Bioenergetica, Consiglio Nazionale delle Ricerche, via Amendola 165/A, 70126 Bari*

**Motivation:** MicroRNAs (miRNAs) are 19-24 nucleotide-long non-coding RNAs that negatively regulate gene expression at the post-transcriptional level. Their profiles are significantly altered in neurodegenerative diseases such as Alzheimer's disease (AD),the most prevalent dementia worldwide. Compared to the major traditional methods used to measure the expression levels of miRNAs (i.e. Real-Time Reverse Transcription PCR (qRT-PCR) and Microarray), Next-Generation Sequencing technologies represent a more powerful tool for miRNA discovery and profiling. In this work, we used the miRNA-seq technology to investigate the pathological mechanisms of AD and to discover new diagnostics and prognostics markers of the disease and new drug targets.
**Methods:** Small RNA fractions were extracted using the mirVana™ miRNA Isolation Kit (Ambion, Life Technologies) from frozen samples of frontal and temporal gyrus of a cohort made of 8 AD cases and 5 controls matched for sex, age and post-mortem interval, obtained from the Netherlands Brain Bank, collected according to guidelines from legislative and ethical boards. 100 ng of small RNA was used as input for miRNA library construction following Illumina's TruSeq Small RNA Sample Preparation Protocol (Illumina). Barcoded miRNA libraries, checked for quality on a Bioanalyzer 2100 instrument and fluorimetrically quantified, were pooled and subjected to 1x50 nt sequencing on the Illumina MiSeq platform. Reads were mapped to known human miRNA precursors, allowing for known post-transcriptional modifications. Ambiguously mapping sequences were assigned to most probable genomic origins using an expectation maximization algorithm. Differentially expressed miRNAs were identified using both the DESeq and SAMSeq methodologies which generated substantially congruent conclusions. Probable functional implications of miRNA differential expression were evaluated though analysis of pathways (KEGG) enriched in predicted targets of differentially expressed miRNAs.
**Results:** Bioinformatics analyses of miRNA-seq data underline in both brain areas a strong deregulation of miR-132-3p and of three family-related miRNAs (miR-132-5p, miR-212-3p and miR-212-5p) encoded by the same miRNA cluster on chromosome 17. Deregulation of these miRNAs in AD brain may contribute to disease progression through aberrant regulation of different mRNA targets such as those involved in the Tau network, Amyloid Precursor Protein maturation and apoptosis.

**The Human Methylome Revealed by NGS Data Using ERNE on a Supercomputer**

Del Fabbro C(1), Prezza N(2), Tardivio F(2), De Paoli E(3), Policriti A(1,2)

*(1) Applied Genomic Institute, Udine (2) Department of Mathematics and Computer Sciences, University of Udine, Udine (3) Department of Agricultural and Environmental Sciences, University of Udine, Udine*

**Motivation:** The combination of next generation sequencing with bisulfite treatment (named BS-seq) has tremendously improved the ability to produce data for cytosine methylation analysis. Cytosine methylation is a DNA modification that has great impact on gene expression regulation and important implications for the biology and health of several living beings, including humans. Unmethylated cytosines are revealed by the BS treatment thanks to their conversion into uracils, which are then read as thymines after PCR and sequencing while methylated cytosines remain uncoverted and are read as regular cytosines by the sequencers. However, by converting unmethylated cytosines into thymines, BS-seq poses computational challenges to read alignment and aggravates the issue of multiple hits due to the ambiguity raised by the reduced sequence complexity. Aligning BS-treated reads is one or two orders of magnitude slower than standard DNA alignment, hence new software able to exploit the advantage of clusters of computers will be a standard in the next few years.
**Methods:** A parallel program called ERNE-PBS5 was designed to take advantage of the Fermi Italian supercomputer (which requires a minimum allocation of 1024 processors). ERNE-PBS5 is the evolution of the ERNE-BS5 alignment software and it is able to efficiently align BS-treated reads. Both the reads and reference sequence are partitioned among the cluster's nodes. Another ingredient is to use MPI (Message Passing Interface) to transmit input and output of the alignments performed. The new approach consists in adding communication among nodes during the alignment. When a node finds a better solution than the ones currently discovered, the possibility to communicate to other nodes reduces global search space and allows saving computation time. After the alignment, the methylome is reconstructed by the ERNE-METH software that calls the methylation level for each cytosine in the reference genome. ERNE is free software, distributed under the Open Source License (GPL V3) and can be downloaded at: http://erne.sourceforge.net.
**Results:** Here we present how it is possible to reveal the human methylome at single base resolution using data produced with the BS-seq protocol. Reads are aligned using ERNE-PBS5, an MPI-ready software able to take advantage of the communication capabilities within a cluster of computers. A classical approach with a single node would require from days to weeks to align the (huge) amount of data produced by next generation sequencers, while the use of the supercomputer reduces the waiting time to at most six hours. The alignments are also processed by ERNE-METH to call the methylation level for each cytosine in the genome.
**Contact email:** delfabbro@appliedgenomics.org

**Exploring isoMIRNAs in small RNA-seq analysis**

Grassi L.(1), D'Andrea D.(1), Tramontano A.(1,2)

*(1) Department of Physics, Sapienza University, Rome (2) Istituto Pasteur—Fondazione Cenci Bolognetti, Rome*

**Motivation:** Typically microRNAs (miRNAs) are annotated as single defined sequences (canonical). Despite this, many miRNAs display several length and/or sequence variants, named isomiRNAs. Once these variants were discovered, the scientific debate was about the extent of noise to their abundance estimation. They were indeed interpreted as consequences of sequencing or alignment artifacts, poor quality or degraded RNA or sloppy Drosha/Dicer excision. The results derived by high-throughput sequencing technologies confirmed that isomiRNAs are observable in many sample of different species. Despite the consistent presence in several studies, the biological relevance of isomiRNAs remains controversial. A small but growing number of studies suggest that, in certain cases, alternative isomeric forms may have different properties. At the same time the lack of tools able to test the specific effects of different variants leaves still open the debate about the isomiRNA functionality. Indeed, since the miRNA target recognition is circumscribed at the seed region placed at the five prime end, it is easy to imagine the effects that these variants can have on target selection. Otherwise it is more complex to understand the functional diversities among three prime variants. We developed a stand-alone software able to detect isomiRNAs in small RNA-Seq results and used it to explore the miRNA variants properties in order to better understand their biological relevance.
**Methods:** We analyzed the small RNA-Seq results derived from the study of He et al. (2012) relative to 5 cerebral cells/tissues. They consisted in cortical GABA+ neurons, cerebellar Purkinje cell, cortical excitatory neurons, whole neocortex and whole cerebellum, each one in three biological replicates. We developed a stand-alone software, named isomiRT, able to identify all possible isomiRNAs present in a given sample, starting from the results of small-RNA sequencing experiment. The tool is a stand-alone software that rapidly processes the reads (in fastq or fasta format) derived from sequencing results. In a first step the tool uses as reference the mature miRNAs annotated in miRbase, subsequently the non-mapped reads are compared with miRNA precursors in order to detect longer, shorter, overlapping or modified variants. The tool gives as result all the variants for each miRNA, reporting the occurrences and the sequences with the relative variations. Furthermore graphs relative to general statistics are produced, such as the distributions of variants, the mutations in seeds or the type of mutations.
**Results:** The first observable result is that canonical miRNAs not necessarily constitute the most expressed class of miRNA. Their occurrence vary among different cells/tissues and accounts at most for half of the expressed miRNAs in a given sample. Moreover the comparison of isomiR profiles reveals a great consistency among the biological replicates of a given sample and, at the same time, evident differences across the sample. This indicates that the phenomenon of isomiRNAs is cell/tissue specific and, at the same time, it exhibits strong repeatability. We calculate the Shannon Entropy and the tissue specificity of each miRNA variant in order to compare canonical miRNAs to other variants. From this comparison emerges that canonical miRNAs and other variants have similar tissue specificity. Finally, the majority of non-canonical miRNAs displays differences at the three prime ends of the miRNA but we also find classes with addition or cleavage variations at the five prime end. To conclude, we provide an effective tool for the identification and the

visualization of miRNAs and their isomiRs in high-throughput small RNA sequencing experiments. Moreover, by using it on several public datasets we revealed that isomiRs are non-randomly distributed across different cell lines or tissue types.
**Contact email:** luigra@gmail.com

**Using replicates to evaluate ChIP-seq peaks**

Jalili V(1), Matteucci M(1), Masseroli M(1), Morelli M(2)

*(1) Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano, Piazza Leonardo da Vinci 32, 20133 Milano, Italy (2) Center for Genomic Science of IIT@SEMM, Istituto Italiano di Tecnologia (IIT), Via Adamello 16, 20139 Milan, Italy*

**Motivation:** The analysis of ChIP-Seq samples outputs a number of enriched regions (ER), each indicating a protein-DNA interaction or a specific chromatin modification. ERs (or "peaks") are called when the read distribution is significantly different from the background, and its corresponding significance measure (p-value) is below a user-defined threshold. In order to avoid a large number of false positives, commonly used thresholds are often very stringent; this generates many false negatives (undiscovered, genuine interaction sites, with read distribution considerably above background level). Biological replicates (i.e., multiple different but biologically equivalent samples, grown/treated under the same conditions) are used to assess the variations of the biological effect studied, and the differences among samples are attributed to true biological variability. Technical replicates (i.e., multiple samples taken from the very same biological system) are used to measure the reproducibility of a specific experiment and its analysis, and differences are attributed to technical issues in measurement. In both cases we expect largely overlapping ER patterns across replicates, with technical replicates much more homogeneous than biological ones. Leveraging on this feature, we propose to use replicates for a comprehensive evaluation of ERs across multiple samples.
**Methods:** For each sample Sj,, where j varies between 1 and the total number of replicates J, we define two disjoint sets of ERs: the stringent set Rjs, containing all the ERs with p-value below a stringent threshold Ts (stringent evidence), and Rjw, containing all the ERs with a p-value below a permissive threshold Tw, but above or equal to Ts (weak evidence). The stringent threshold Ts corresponds to the usual threshold for single sample analysis. Leveraging the presence of biological or technical replicates, we combine the evidence of the weak ERs, and assess whether it gives rise to a p-value equivalent to that of a stringent peak. For this, given that samples under the same conditions are independent, we evaluate if the combined evidence rejects the same overall null hypothesis ("the ER is not a true binding site"). We determine intersections among ERs on multiple samples. If one region on a sample intersects with multiple regions on another sample, for that sample we consider only the intersecting region with the most stringent p-value. For each set of (stringent and weak) intersecting ERs, we obtain a single test statistic, assuming that "all null hypothesis coming from different independent intersecting ER's are true" is the null hypothesis of the data fusion, and considering "at least one of the alternative hypothesis of independent intersecting ERs is true" as alternative hypothesis of the combined analysis. Multiple evidence is integrated through the Fisher's combined probability test, which estimates a single-test statistic X2 of multiple evidences, which follows a chi square distribution with 2k degrees of freedom, where k denotes the number of ERs being combined. To evaluate X2, we define a threshold gamma. If X2 > gamma we reject the null hypothesis of the combined test, and therefore reject the null hypothesis of each individual test (i.e., intersecting ERs are part of the background), thus validating all the intersecting peaks, even if they fall in the weak set, and would therefore be rejected in a single-sample analysis.
**Results:** The method was tested on a collection of samples (including both technical and biological replicates) from the publicly available ENCODE datasets. The method validates

weak ERs with p-values close to Ts if they intersect with stringent ERs and their combined evidence is strong enough. ERs with p-values close to Tw are more possibly a portion of background signal rather than a binding site. However, if such ERs intersects over a large enough number of replicates, the chances of being a true binding site increase strongly and the ERs may be "rescued". The new collection of ERs was validated through Gene Ontology and motif analysis with good results.

**Contact email:** vahid.jalili@polimi.it

**3USS: a web-server for detecting alternative 3'UTRs from RNA-seq experiments.**

Le Pera L(1), Mazzapioda M(2), Tramontano A(2,3)

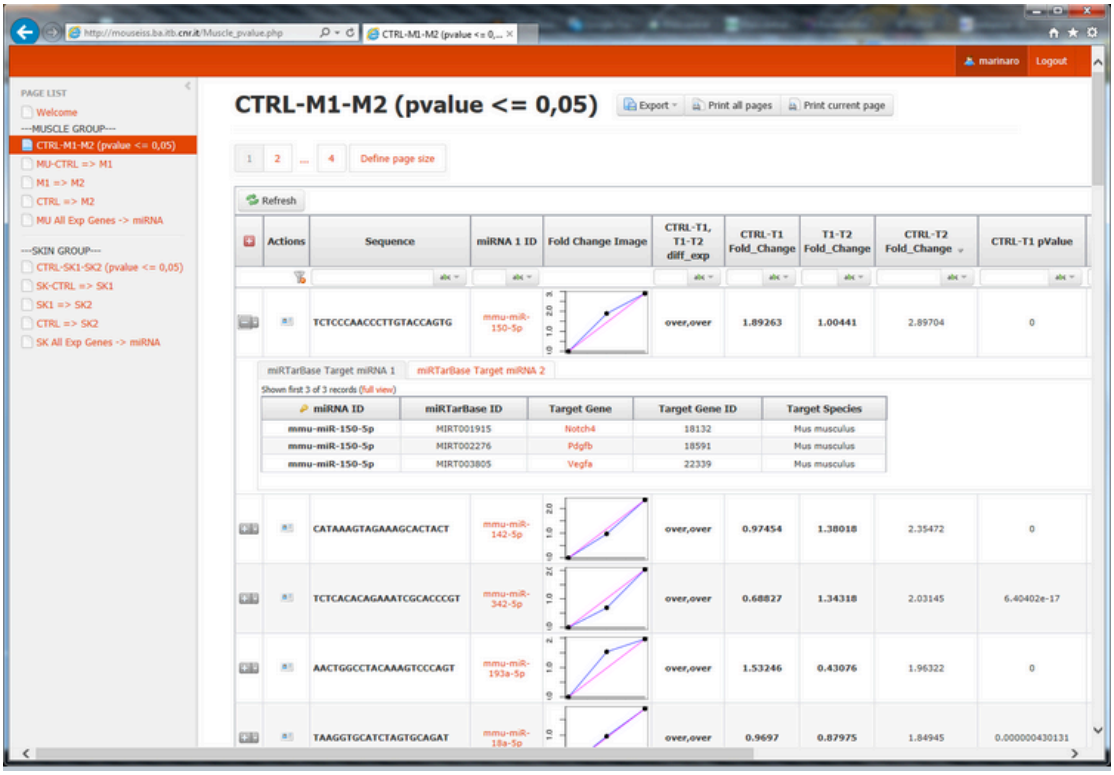
*(1)Center for Life Nano Science@Sapienza,Istituto Italiano di Tecnologia,Viale Regina Elena 291, 00161,Roma (2)Department of Physics,Sapienza University,Rome, 00185,Italy (3)Istituto Pasteur—Fondazione Cenci Bolognetti, Sapienza University,Rome,00185,Italy*

**Motivation:** Alternative polyadenylation has been identified as a widespread mechanism in eukaryotic cells and as an important step of post-transcriptional regulation. Protein-coding genes with multiple alternative polyadenylation sites can generate various lengths of their mRNA 3'UTR sequences, with the loss or gain of regulatory elements affecting stability, degradation, subcellular localization and translation. Next-generation sequencing technology (unlike previous approaches) is now able to detect variations with respect to already annotated transcripts. In standard RNA-seq experiments and data analysis protocols, in fact, the obtained reconstructed transcripts can result as already known or putative novel transcript forms. Most of the reconstructed transcripts entirely match the annotated ones, some others can differ for short or long parts (one or more exons). For this reason, it could be very promising a tool able to automatically detect transcripts with alternative (also putative novel) 3' untranslated regions, shorter or longer in specific biological conditions.
**Methods:** We implemented a web-server called 3USS (3' Utr Sequence Seeker) to help researchers to easily retrieve 3'UTR genomic coordinates and nucleotide sequences from the protein-coding transcripts reconstructed by RNA-seq data analysis protocols. The server accepts as input the GTF transcript file generated by procedures such as Cufflinks or Scripture (the most used tools performing the transcriptome assembly). Subsequently it compares the transcripts to the ones annotated in UCSC, NCBI, or Ensembl repositories, detecting those with diverse 3'UTR length. The data obtained from two RNA-seq experiments can also be uploaded and compared.
**Results:** 3USS is the unique web-server able to identify RNA-seq reconstructed transcripts with 3'UTR shorter or longer than the annotated one. It returns the following Results: i) the list of the transcripts, along with the length differences; ii) the putative novel 3'UTR genomic coordinates and a multi-fasta file of the nucleotide sequences. Furthermore, it gives the possibility to compare data of two experiments, determining which transcripts share putative novel 3'UTRs in specific biological samples. These findings can be very useful to further investigate the alternative polyadenylation process, the functional role of which remains poorly understood.
**Contact email:** Anna.Tramontano@uniroma1.it

**ncARENA: an integrated resource for non-coding RNAs functional annotation**

Licciulli F, Consiglio A, De Caro G, D'Elia D, Gisel A, Grillo G, Tulipano A, Liuni S
*CNR - Istituto di Tecnologie Biomediche, Bari, Italy*

**Motivation:** Non-coding RNAs (ncRNAs) are key regulators in several and important cellular processes (e.g., transcription, chromosome replication, RNA processing and modification, protein degradation) and represent the majority of information stored in organism genomes. Indeed, recent reports of the ENCODE project underline that while 80% of the human genome is transcribed, only 2% is protein coding, suggesting that the vast majority of the genome is transcribed as non-protein-coding RNA. Thanks to the advent of Next-Generation High Throughput Sequencing (HT-NGS), large screens of genomes are now useful for the study of ncRNAs, but a challenging task for bioinformaticians is the development of tools for an easy and effective analysis of these data. In this context it is of crucial importance the availability of a reliable and comprehensive data resource, including all known non-coding RNA sequences and cross-referenced resources, for their classification and functional annotation. In addition, suitable statistical tools are necessary for the expression profiling of data, when comparing samples in different conditions and time courses. We have developed a bioinformatics platform for the annotation and classification of ncRNAs from different HT-NGS platforms based on a data-warehouse approach. Here we present recent improvements of this platform, that have been focused on the architectural design and implementation of tools and pipelines for the annotation of NGS data and the analysis of differentially expressed ncRNAs in human and mouse.
**Methods:** The data-warehouse has been implemented with Infobright (www.infobright.org), while the source databases have been integrated by using the open-source Pentaho Data Integration tool. The data-warehouse contains our non-redundant ncRNA reference database (ncRNAdb) that has been recently re-built by integrating ncRNA gene lists from MGI (Mouse Genome Informatics) and HGNC (Human Genome Nomenclature Committee) with sequences and biotype annotations from VEGA (Vertebrate Genome Annotation), ENSEMBL and RefSeq. Additional resources are: i) miRBase for miRNA mapping and classification in family and gene clusters; ii) experimental validated miRNA target interactions databases (miRTarBAse, TarBase, miRecords); iii) Sequence Ontology (SO) and Gene Ontology (GO) for sequence functional annotation; v) NCBI BioSystems Database for pathway annotation. The web interface has been developed by using the JAVA platform (JSP, Hibernate, Apache Tomcat). The data analysis pipeline currently includes tools for miRNAs identification (mirDeep2), differential expression profiling (Fisher's Exact test, FDR with R) and perl scripts for reads adapter removal and other pre/post-processing steps.
**Results:** Currently the ncRNAdb contains a total of 37,828 sequences classified in 23 biotypes associated to Sequence Ontology terms. As for functional annotation, the data-warehouse contains, among others, information about 410,581 biochemical pathways and 121,579 experimentally validated miRNA target interactions. As test case, we have used Illumina small RNA-seq data produced for expression profiling of smallRNAs in the immune response induced by rabies vaccines in M. musculus. This experiment includes three time courses and three technical replicates from two different tissues. ncARENA includes a query and retrieval system that allows to extract and visualize data by using different search criteria (e.g., ncRNA gene, fold change, RPKM, p-values, over/under-expression trend). Results are enriched with additional information such as target genes and pathways for each miRNA. The Figure shows an example of query results visualization with graphics and tables.
**Contact email:** flavio.licciulli@ba.itb.cnr.it


CTRL-M1-M2 (pvalue <= 0,05)
Export
Print all pages
Print current page
Refresh
Actions
Sequence
miRNA 1 ID
Fold Change Image
CTRL-T1, T1-T2 diff_exp
CTRL-T1 Fold_Change
T1-T2 Fold_Change
CTRL-T2 Fold_Change
CTRL-T1 pValue
TCTCCCAACCCTTGTACCAGTG
mmu-miR-150-5p
over,over
1.89263
1.00441
2.89704
0
miRTarBase Target miRNA 1
miRTarBase Target miRNA 2
miRNA ID
miRTarBase ID
Target Gene
Target Gene ID
Target Species
mmu-miR-150-5p
MIRT001915
Notch4
18132
Mus musculus
mmu-miR-150-5p
MIRT002276
Pdgfb
18591
Mus musculus
mmu-miR-150-5p
MIRT003805
Vegfa
22339
Mus musculus
CATAAAGTAGAAAGCACTACT
mmu-miR-142-5p
over,over
0.97454
1.38018
2.35472
0
TCTCACACAGAAATCGCACCCGT
mmu-miR-342-5p
over,over
0.68827
1.34318
2.03145
6.40402e-17
AACTGGCCTACAAAGTCCCAGT
mmu-miR-193a-5p
over,over
1.53246
0.43076
1.96322
0
TAAGGTGCATCTAGTGCAGAT
mmu-miR-18a-5p
over,over
0.9697
0.87975
1.84945
0.000000430131

**COSMOS: NGS Analysis in the Cloud**

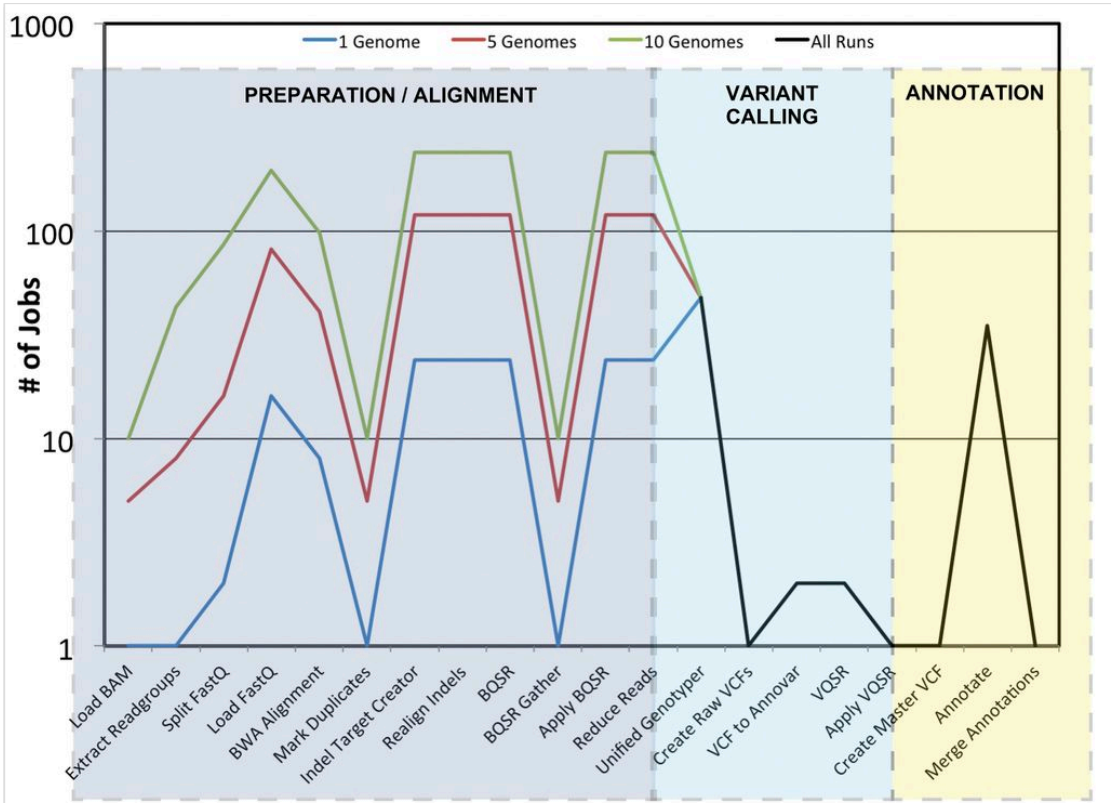
Ettore Rizzo (1), Jared B. Hawkins Ph.D. (2), Yassine Souilmi (3), Jae-Yoon Jung Ph.D. (2), Riccardo Bellazzi Ph.D. (1), Dennis P. Wall Ph.D (4), Peter Tonellato Ph.D (2).

*(1) University of Pavia, Italy, (2) Harvard Medical School, Boston, MA, (3) Faculty of Sciences of Rabat, Morocco, (4) Stanford Medical School, Palo Alto, CA*

**Motivation:** Modern sequencing platforms are capable of sequencing approximately 5,000 megabases a day and programs such as the 1000 Genomes project are routinely generating data on the petabyte scale. The current challenge lies in the analysis and interpretation of this data, which has become the new rate-limiting step. Providing a solution to this problem that is both timely and cost-effective will be of great scientific importance and a major technological breakthrough. To address this need, we have developed COSMOS, a scalable and parallelizable workflow manager capable of running on the cloud (Amazon Web Services–AWS) that has the potential to reduce the cost of the analysis of whole genomic data over 10-fold, from ~1,000€ to under 100€. COSMOS is able to lower the cost of analyzing genomic data two ways: 1) implementing a highly parallelizable workflow that can be run quickly and efficiently on a large compute cluster, and 2) taking advantage of AWS spot-instance pricing to reduce the cost per hour.
**Methods:** COSMOS is a workflow optimization manager for which we have developed three NGS-analysis workflows that analyze both germline genomic data and matched normal-tumor genomic data. The first workflow implements the GATK best practice protocol developed by the Broad Institute, which is widely accepted as the industry standard. The second workflow is being used to analyze “ancient” genomic data. The focus of this study is the implementation of a third COSMOS workflow which executes GATK but also performs additional important analysis on somatic variants and Copy Number Variant detection using Mutect and SVDetect. These sophisticated workflows perform a thorough sequence analysis, including quality control, alignment, SNP, indel variant calling, copy number variation detection and a robust and complete set of automated annotation using a custom extension of ANNOVAR. COSMOS also provides a dynamic web interface making it possible to conduct real-time monitoring of the workflow, state and job dependencies and resource use of each job. After COSMOS loads the workflow, a “workflow” parser, breaks up each stage of the workflow into multiple jobs that are then executed in parallel (Figure). Jobs are distributed from a master node to worker nodes using a standard job manager. To provide complete control of the clouded cluster, we use standard AWS EC2 nodes and install the open source cluster management software, StarCluter, that allows to launch and shutdown cluster nodes without user intervention and automatically installs a job manager (Sun Grid Engine) on each worker node. Finally, we have incorporated a COSMOS plugin to install an HPC cluster file sharing system (GlusterFS) in all nodes.
**Results:** We are testing COSMOS to process and analyze successively larger sets of genomic data to optimized the software for speed, accuracy, automation, and computational cost. In this way, COSMOS is becoming a scalable, optimal solution to the increasing demand for effective computational systems in bioinformatics. Figure demonstrates how COSMOS breaks up stages of the NGS-pipeline into multiple jobs for different sized datasets (shown here for 1, 5 and 10 genomes). This technique allows for efficient parallelization of the early stages of the pipeline. Later stages are run in serial to maximize accuracy of the analysis. In addition, we have testing time performances for one exome using a 5 node cluster with each node a “cc2.8xlarge” AWS instance (32 cores, 60 Gb Ram). The whole analysis took less than

3 hours of AWS “wall” time and (more importantly) cost less than 50€. Our work recently resulted in a grant from AWS to demonstrate the scalability of COSMOS and we are analyzing several cohorts: 1) 50 Epilepsy patient exomes; 2) 50 matched tumor-control exomes from Chronic Myelomonocytic Leukemia patients; and 3) a cohort of 50 genomes. We have set a daily goal of fully analyzing the output of the HiSeq 2500 recently bought by the Policlinico San Matteo (≈ 80 exomes).

**Contact email:** ettore.rizzo@unipv.it



**The RNA-Seq application to characterize the gene expression and isoform signatures of ten genes associated with cancer survival in non-involved lung adenocarcinoma tissue**

Spinelli R(1) and Galvan A(2), Piazza R(1), Pirola A(1), Dragani T.A.(2) and Gambacorti-Passerini C(1,3)

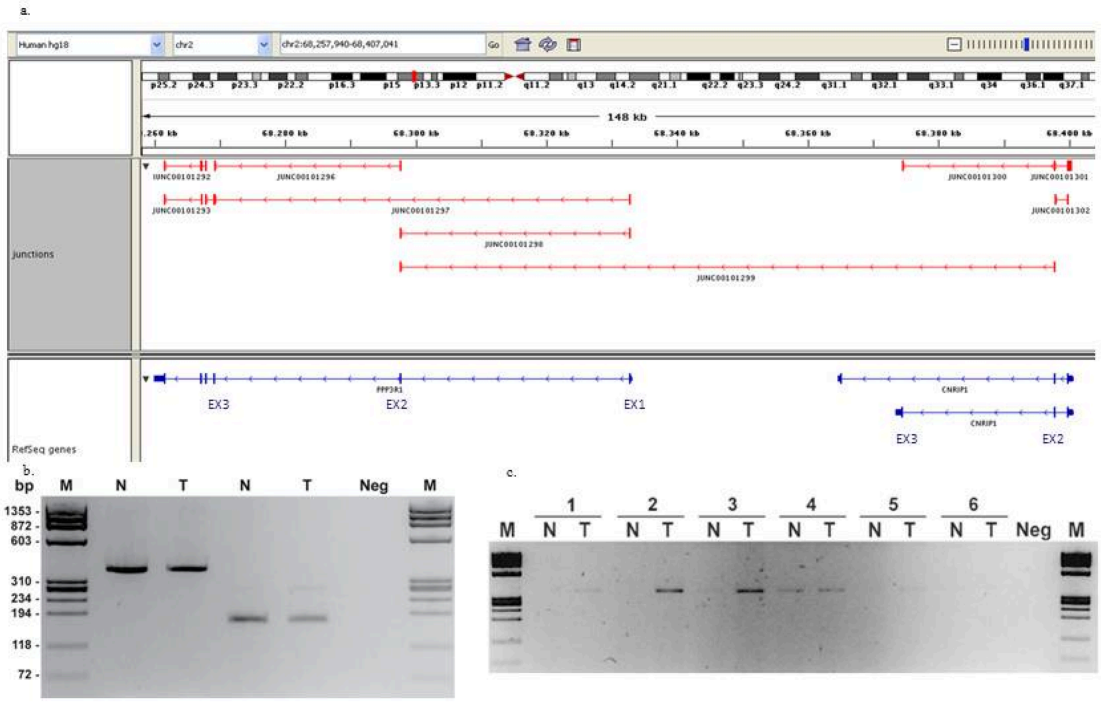
*(1) Dipartimento di Scienze della Salute, Università Milano-Bicocca, Monza, Italy. (2) Fondazione IRCCS Istituto Nazionale dei Tumori, Milan, Italy. (3) Unità di Ricerca Clinica ed Ematologica, Ospedale San Gerardo, Monza, Italy.*

**Motivation:** Lung adenocarcinoma patients of similar clinical stage and undergoing the same treatments often have marked inter-individual variations in prognosis. The clinical discrepancies may be due to the genetic background modulating an individual's predisposition to fighting cancer. We supposed that the lung microenvironment, as reflected by its expression profile, may affect lung adenocarcinoma patients' survival. The high-throughput RNA-Seq is a very powerful technology for transcriptomics studies and a great opportunity to identify significant alterations in gene expression profiles. RNA-Seq could capture almost all of the expressed transcripts without relying on prior information with low background noise, high sensitivity and allowing the identification of novel splicing variants and fusion genes. In order to characterize the gene and isoform expression patterns in non-involved lung adenocarcinoma tissue of 10 candidate genes obtained by microarray survival analysis (Galvan A. et al., 2013), the RNA-Seq analysis was carried out in 12 patients randomly selected from a cohort of patients analyzed by microarray survival analysis (Galvan A.et al., 2013).
**Methods:** Starting from 2µg of total RNA, extracted from non-involved lung tissue of 12 lung adenocarcinoma patients. The libraries were generated using Illumina TruSeqTM RNA Sample Preparation Kit with fragment size of 400-500 bp and sequenced by Illumina GAIIX with 76 bp paired-end reads. Image processing and base calling were performed by RTA (vs1.9.35). Qseq files were de-indexed and converted to the Sanger-FastQ file format, using in-house scripts. FastQ sequences were aligned to the hg18 by TopHat (vs.1.2.0) (Trapnell, et al., 2009) with default parameters. The RNA-Seq reads were mapped using the gene and splice junction models as provided in the annotation GTF file (Ensembl vs54). TopHat aligns the RNA-Seq reads through the genome using Bowtie (Langmead B, et al. 2009) and then maps the initially unmappable reads to the known splice junctions sequences supplied by the GTF file. A splice junction map of each gene and each isoform was inferred from TopHat and checked using Ensembl database (vs71, GRCh37). The gene expression profile was estimated by SAMMate (vs2.6.1, Xu, et al., 2011) using the GTF file and the default parameters. The gene expression values for paired-end data were measured as FPKM (Mortazavi A., et al. 2008), which is a normalized measure of the exonic read density and the concentration of transcript.
**Results:** By using this approach we identified the complex expression pattern of the 10 candidate genes potentially associated with cancer survival in lung adenocarcinoma patients. The genes have from 4 to 20 annotated transcripts according to the Ensembl database (vs71). The RNA-Seq analysis confirmed that all the genes are expressed in lung tissue showing that they are present as a main isoform in all samples and from 0 to 6 minor isoforms in a subset of cases. Surprisingly, RNA-Seq was able to detect a novel isoform of CNTNAP1 in 6 samples not yet reported in public databases. It was characterized by the alternative splicing of the main isoform by skipping the exon 10 and resulting in a novel out-of-frame transcript. Moreover, in 3 patients a read-trough between PPP3R1 and the flanking

CNRIP1 was detected. In one sample from exon 2 to exon 2 of PPP3R1 and CNRIP1 (Fig. 1a). No evidence of this event was previously found in Ensembl. The read-through was validated experimentally using the PCR amplification of mRNA fragments around the fusion site from paired non-involved (normal) and lung adenocarcinoma tissue samples (Fig. 1b). This analysis in 40 additional tumor and normal paired samples showed that the PPP3R1-CNRIP1 gene fusion was present in normal and tumoral tissues but with a tendency to higher expression levels in tumor samples than in non-involved tissue (Fig. 1c). Further studies in a larger dataset are needed to assess the transcriptional results obtained in this study.

**Contact email:** roberta.spinelli@unimib.it

**Supplementary information:**Fig 1. Legend. a. RNA-Seq read-through between PPP3R1 and CNRIP1 genes displayed by the Integrative Genomics Viewer (Robinson, J.T. et al., 2011). Splicing map showed 8 junction reads (JUNC00101299) between ex2-ex2 of PPP3R1 and CNRIP1 genes. b. PCR amplification of PPP3R1–CNRIP1 fusion fragments detected by RNA-Seq. PCR amplification in non-involved (N) or lung adenocarcinoma (T) tissue, using PCR primers amplifying either a long fragment (first two bands from the left) or a short fragment (third and fourth bands from the left). M, DNA molecular weight marker, ϕX174 DNA-Hae III Digest; Neg, negative control using water instead of DNA as PCR template. c. PCR amplification of gene fusion fragments in 6 representative pairs of non-involved lung (N) and lung adenocarcinoma (T) tissues with different expression levels. The gene fusion band is clearly detected in patients 1 to 5, but not 6.



**Adaptive smoothing algorithm for rare fusion events detection in exon-like array data**

Mazza T(1), Bosotti R(2)

*(1)Bioinformatics Unit,IRCCS Ospedale Casa Sollievo della Sofferenza, Roma (2)Genomics Group, Biotechnology Dep., Nerviano Medical Sciences, Nerviano (MI)*

**Motivation:** Critical events such as translocations, deletions or chromosomal inversions are known to contribute to the development of cancer. Among them, gene fusions arising from genetic rearrangements of two distinct genes can result in constitutively activated chimeric proteins. Several fusion genes involve kinases, which are a family of proteins that participate in several key cellular functions that are disregulated in cancer. Just to cite a few, this is the case of the EML4-ALK fusion protein in lung adenocarcinomas, TPM3-TRKA in papillary thyroid carcinomas and lung adenocarcinomas and KIF5B-RET in lung adenocarcinomas. For these tyrosine kinase receptors, the fusion partner promoter drives the overexpression of the chimeric gene, encoding a dimerization domain that induces abnormal activation of the fused kinase catalytic domain [1, for a review]. Therefore, unbalanced expression of 5' and 3' gene regions, such as selective expression of kinase intracellular domain in the absence of the extracellular portion, may be used as readout of genomic rearrangements. In literature, gene expression arrays [2], SNP arrays [3] and, more recently, RNAseq sequencing data [4, for a review] have been analyzed with the aim of identifying potentially rearranged genes. With respect to gene expression arrays, Human Exon 1.0 ST Arrays (Affymetrix) provide accurate information on the different gene splicing variants, producing per each exon a signal representing the average of four probes, along the whole sequence. Public databases such as GEO (Gene Expression Omnibus) provide this type of data for 1588 samples. We exploited the potential of Human Exon 1.0 ST Array to search for unbalanced exon expression, with the aim to identify potentially rearranged genes by rare fusion events.
**Methods:** We designed an algorithm that searches outlier expression levels of contiguous exons in a limited number of samples among those analyzed. First, it annotates exon-like arrays data by custom CDF files taken from the "Brainarray web portal" [5]. Candidate rearranged transcripts are those with at least one exon with outlier expression value among all samples. Then, expression profiles of the nominees are checked for trend coherency, in order to guarantee a significant difference of expression levels between the 5' and 3' ends. The whole procedure relies on an adaptive segmentation sub-routine, which aims at identifying the two partners making the fused gene. The algorithm outputs a list of candidate transcripts with anomalous expression of a subset of contiguous exons.
**Results:** Preliminarily, the method was validated on publicly available Human Exon 1.0 ST Array data from prostate tumor tissue samples, half of which were reported to carry the TMPRSS2-ERG rearrangement [6]. The algorithm correctly identified outlier exon expression levels in TMPRSS2-ERG rearrangement-positive tumor samples. This encouraging result justifies its extensive test on datasets generated on other tumor types, from which we reserve to collect overall statistics of efficiency and efficacy. To our knowledge, this is the first algorithm that makes use of Human Exon 1.0 ST Array with the aim to identify outlier exon expression levels of all the transcripts of the human kinome, through a cohort of samples. The algorithm can be directly used with the newly launched Affymetrix Human Transcriptome Array (HTA 2.0), which contains more than six million probes deeply covering coding transcripts, non-coding transcripts and exon-exon splice junctions. REFERENCES: 1. Medves S and Demoulin JB, J Cell Mol Med.,16(2):237-48 (2012) 2. Tomlins SA et al. Science,

310(5748):644-8 (2005) 3. Thieme and Groth, Bioinformatics, 29(6):671-7 (2013) 4. Carrara M et al. Biomed Res Int., 2013:340620 (2013) 5. http://brainarray.mbni.med.umich.edu/brainarray 6. http://www.ncbi.nlm.nih.gov/geo/
**Contact email:** roberta.bosotti@gmail.com

**Benchmarking alternative splicing detection methods**

Carrara M(1), Loom J(2), Cordero F(3), Beccuti M(3), Donatelli S(3), Zolezzi F(2), Calogero RA(1)

*(1)Department of Molecular Biotechnology and Health Sciences, University of Torino, Via Nizza 52, 10126 Torino, Italy (2)Singapore Immunology Network (SIgN), Agency for Science, Technology and Research (A*STAR), Singapore (3)Department of Computer Sciences, University of Torino, C.so Svizzera 185, 10149 Torino, Italy*

**Motivation:** RNAseq provides tremendous power in the area of biomarkers discovery and disease stratification. However, statistical detection of alternative splicing (AltDE) is a particularly challenging task since it has to take in account moderate fold change differences to be detected within a complex environment of different isoforms.
**Methods:** AltDE can be investigated using transcripts, exons and exon-exon junction approaches. The evaluation of the performance of those approaches is actually limited by the lack of benchmark datasets. To overcome this issue we constructed a spike-in RNAseq experiments in which we spike-in both synthetic and real data.
**Results:** The results obtained using this dataset highlight the strong limit of the transcript-based AltDE detection, i.e CuffDiff 1/2. Instead exon-level AltDE is very sensitive, i.e. DEXseq, although the presence of highly expressed non-differentially expressed isoforms quenches AltDE. Exon-exon junction based approaches seems to perform reasonably if used in an isoform-specific contest.
**Contact email:** raffaele.calogero@gmail.com

**Library prep effects on RNAseq downstream analyses**

Loom J(2), Cordero F(3), Carrara M(1), Beccuti M(3), Donatelli S(3), Calogero RA(1), Zolezzi F(2)

*(1)Department of Molecular Biotechnology and Health Sciences, University of Torino, Via Nizza 52, 10126 Torino, Italy (2)Singapore Immunology Network (SIgN), Agency for Science, Technology and Research (A*STAR), Singapore (3)Department of Computer Sciences, University of Torino, C.so Svizzera 185, 10149 Torino, Italy*

**Motivation:** RNAseq provides tremendous power in the area of biomarkers discovery and disease stratification. However, the technology is not mature, yet. A lot of new applications and protocols are appearing but very little is know on the effects they have on data generation.
**Methods:** To better understand this issue we run multiple RNAseq experiments on the same RNA sample using different sample prep procedures (Fig. 1). The resulting data were investigated both at gene and transcript level to grasp the effects produced by different sample preps on coding genes detection.
**Results:** Our data indicate that sample preps affect transcript representation both at the level of detection, i.e. FPKM, and at the level of coverage. Low input protocols are strongly biased at transcript/exon level
**Contact email:** raffaele.calogero@unito.it

Fig. 1

| Vendor | Kit | Assay | # PF reads |
|---|---|---|---|
| Illumina | TruSeq RNA | **mRNA-Seq** | **239,990,688** |
| Epicenter | ScriptSeq | **mRNA-Seq** | **124,118,496** |
| Illumina | TruSeq Stranded total RNA | **Total RNA-Seq** | **155,696,736** |
| Illumina | TruSeq Stranded mRNA | **Stranded mRNA-Seq** | **205,921,656** |
| Nugen | Ovation v1/2, Rapid DR | **Low input mRNA-Seq** | **478,361,952** |
| | | | **1,204,089,528** |

**A meta-analysis of transcriptomic data on Hepatocellular carcinoma to discover new putative biomarkers**

Sakshi (1), Colonna G (1), Castello G (2), Costantini S (2)

*(1) Department of Biochemistry, Biophysics and General Pathology, Second University of Naples, Naples, Italy (2) INT Pascale – Cancer Research Centre of Mercogliano, Mercogliano, Italy*

**Motivation:** Hepatocellular carcinoma (HCC) is a major health problem worldwide being the fifth most common malignancy in men and the eighth in women; the third most common cause of cancer-related death in the world. In particular, Southern Italy has the highest rates of HCC in Europe. Moreover early diagnosis is uncommon and medical treatments are inadeguate. For these reasons it is necessary to define new putative markers for diagnosis, prognosis and treatment of hepatocellular carcinoma. Recently we performed the evaluation of gene expression of human hepatoma cells (HepG2) in comparison to normal human hepatocytes by microarray studies and defined the signature of down- and up-regulated genes in these cells [1]. In this work we carried out the comparison between transcriptomic data obtained in our laboratory with those reported in public databases on HCC in order to identify new biomarkers of the different processes that can lead to HCC development in presence or absence of other risk factors such as hepatitis C virus (HCV), type 2 diabetes (T2D), cirrhosis (LC) etc.
**Methods:** Firstly we collected the data derived from our experiment conducted by HumanWG-6 BeadChip array that permits to evaluate 48,000 transcripts (Whole Genome). Data were normalized and selected using p value < 0.01, diff score > 13 and Fold > 2 [1]. Now we are collecting data related to HCC cells and tissues reported in ArrayExpress and in The Cancer Genome Atlas (TCGA) that are databases of functional genomics experiments that can be queried and downloaded. Hierarchical clustering of genes is performed by Cluster program. Information on the biological functions of the genes that are significantly regulated ($P<0.05$) is determined using the Gene Ontology Enrichment tool in DAVID. Pathway analysis is made by DAVID and IPA programs.
**Results:** Hierarchical cluster analysis showed the differential expression of genes in HepG2 cells respect to human hepatocytes, used as healthy controls. 2646 genes were significantly down-regulated in HepG2 cells respect to hepatocytes, whereas an additional 3586 genes were significantly up-regulated [1]. We have collected transcriptomic data related to patients with HCV, LC, T2D and HCC and to healthy controls. In this moment we have compared the list of collected genes and have identified the differentially expressed genes in the different disease conditions. We are performing functional analysis on selected genes that result involved in different metabolic pathways and biological functions such as translational, ribosomal structure, RNA processing and modification, lipid metabolism, amino acid transport etc. Moreover we are focusing our attention on selenotranscriptome and genes involved in circadian rhythms.
**Contact email:** susan.costantini@unina2.it
**Supplementary information:**References [1] S. Costantini, G. Di Bernardo, M. Cammarota, G. Castello and G. Colonna. Gene Expression segnature of human hepatoma cells". Gene (2013) 518:335-345.

**Nucleosome loss facilitates the inflammatory response in macrophages**

De Toma I(1), Bianchi ME(1,2), Alessandra Agresti A (2)

*(1) University Vita-Salute San Raffaele, Milan, Italy (2) San Raffaele Scientific Institute, Division of Genetics and Cell Biology, Milan, Italy*

**Motivation:** HMGB1 is a small nuclear protein with two different functions: in the nucleus it bends DNA, helping DNA wrapping of around nucleosomes; when secreted, it both recruits inflammatory cells and induces them them to produce cytokines. HMGB1 secretion in inflammatory cells is preceded by its relocation to the cytoplasm, leaving cell nuclei partially or totally depleted of HMGB1. We previously showed that cells lacking HMGB1 contain 20% fewer nucleosomes and have transcription levels increased genomewide by 30%. We hypothesized that the depletion of nuclear HMGB1 plays a role in inflammation that is additional and/or complementary to that of extracellular HMGB1.
**Methods:** We analysed the transcriptional profile of wild type and Hmgb1–/– MEFs, as a proxy for cells that have lost HMGB1 from their nuclei. We also explored the transcriptome of wild type and Hmgb1–/– macrophages in their polarised M1 basal state and exposed to LPS/IFN-γ. In the same cells, histones and nuclear HMGB1 were quantified.
**Results:** We found that Hmgb1–/– mouse embryonic fibroblasts show a transcriptional profile associated with the response to stress and inflammation. Moreover, wild type macrophages that have lost HMGB1 because of LPS/IFN-γ exposure rapidly reduce their histone content as much as cells genetically lacking HMGB1. Importantly, unstimulated Hmgb1–/– M1 macrophages activate transcriptional pathways associated to cell migration and chemotaxis. We suggest that nucleosome loss is an early event that facilitates the transcriptional response of macrophages to inflammation. HMGB1's dual roles in the nucleus and in the extracellular space appear to complement each other.
**Contact email:** detoma.ilario@hsr.it

Poster: Management – MGM1

**GMO-Matrix: A tool for in silico prediction of GMO detection.**

Angers A(1), Petrillo M(1), Patak A(1), Kreysa J(1)

*(1) Molecular Biology and Genomics Unit, Joint Research Centre, European Commission, Ispra (VA), Italy*

**Motivation:** The Joint Research Centre (JRC) of the European Commission is host to the European Union Reference Laboratory for Genetically Modified Food and Feed (EU-RL GMFF). The core tasks of the EU-RL GMFF are the scientific assessment and validation of detection methods for GM Food and Feed as part of the EU authorisation procedure. In addition, availability of reliable information on methods for detection, identification, and quantification of GMOs is of fundamental importance for control and inspection measures and this activity represent an important part of the mission of the EU-RL GMFF.
**Methods:** The EU-RL GMFF maintains an EU database of polymerase chain reaction (PCR) - based reference methods for GMO analysis, called "GMOMETHODS", publicly available on its website (http://gmo-crl.jrc.ec.europa.eu/gmomethods/). A platform is currently under development to integrate the primers and probe sequence of these reference methods with the Central Core DNA Sequences Information System (CCSIS), the in-house molecular database containing GMO sequence data, either submitted to the EURL-GMFF by applicants or identified within public databases (EMBL, Genbank, Patents, etc.). The resulting platform is a web-based tool that performs in silico predictions of which detection method(s) can detect each GMO. For this, the program simulates the PCR reaction (allowing for gaps and mismatches) using e-PCR, a tool developed at NCBI and installed locally, and, when necessary, aligning the probe sequence using matcher, a sequence comparison application of the EMBOSS package. The platform is built on three main servers: a server dedicated to the database (PostgreSQL) that stores all the DNA sequences and related information, a server that hosts a Ruby on Rails web application that serves as the interface with the user, and a high-performance cluster with the bioinformatics scripts (written in PHP) which can by invoked by the web application for running all the computer-intensive tasks.
**Results:** The GMO Matrix application is currently deployed in the intranet of the JRC and is used to answer ad hoc requests, as well as for the development of new detection methods (identification of gaps, in silico prediction of specificity, etc.). The web interface allows simple queries, such as running the simulation of specific detection methods on specific GMO events, or more complex queries, such as the simultaneous search of positive detection methods on different GMO events, in order to simulate detection of GMO stacked events. The output is usually presented as a two-dimensional matrix, with the selected detection methods presented horizontally and the GMO events vertically. None of the results shown are stored in the database, as the computations are always performed on-the-fly to ensure they are up-to-date with current methods and events information.
**Contact email:** alexandre.angers@ec.europa.eu

Poster: Management - MGM2

**High Accuracy Multiple DNA Sequence Alignment Guided by Single/Multiple Protein Coding Domains**

Balech B(1), Vicario S(2), Donvito G(3), Monaco A(3), Notarangelo P(3), Pesole G(1,4)

*(1) Institute of Biomembranes and Bioenergetics (IBBE), National Research Council, Bari (Italy) (2) Institute of Biomedical Technologies (ITB), National Research Council, Bari (Italy (3)National Institute of Nuclear Physics (INFN), Bari (Italy) (4)Department of Biosciences, Biotechnology and Pharmacological Sciences, University of Bari "Aldo Moro", Bari (Italy)*

**Motivation:** Multiple sequence alignment (MSA) is the core item in biological sequence analysis, especially in phylogeny inference in which parameters are estimated from the MSA observed as a fixed matrix. Numerous algorithms have been developed to generate multiple alignments, such as Muscle (Edgar, 2004), Mafft (Katoh et al., 2002), ClustalO (Sievers et al., 2011), PROMALS (Pei & Grishin, 2007). All these algorithms take the whole sequence length information, while none of them partitions DNA sequence according to its protein coding domain/s and maps each fragment against its best protein domain match. In addition, few algorithms performs an amino acid back translation into DNA procedure, such as translatorX (Abascal et al., 2010) and tranalign (Emboss package). The main shortcoming of these algorithms is that the correct genetic code and frame should be predefined by the user and the merging procedure of the sub-back aligned multiple protein domains should be conducted by the user himself. In the framework of BioVel project (http://www.biovel.eu/), aiming to expose web services and workflows for biodiversity studies, we developed a phylogenetic service set in which a high accuracy multiple DNA protein coding sequence alignment workflow is implemented. This workflow estimates the most likely genetic code and frame to translate DNA sequences and at the same time partitions the data set into the corresponding protein domains found in it. Additionally, it back translates each amino acid domain into DNA alignment and concatenates all domains alignments to finally output a highly accurate protein guided DNA multiple alignment.
**Methods:** The workflow is built in Taverna workflow manager (Wolstencroft et al., 2013) and calls a RESTfull web service to perform the underlying computation and a Job Submission Tool (JST, Donvito et al., 2012) to upload and download the input and output files respectively. It is based on three main parts: i) the first, implemented in Python2.7 and Biopython1.59, translates DNA sequences following a user defined range of possible genetic codes and reading frames and cuts the translated sequence at each stop codon. The genetic code and the frame are tracked by adding their information to the sequence ID for later parsing protocol; ii) the second searches, using hmmsearch and hmmalign algorithms (HMMer3.0 package), the translated amino acid sequences against a local mirror of PFAM-A conserved domains database. It performs a MSA of either single or multiple protein domains coding sequences. Sequences not coding for the same domain/s of the majority of sequences and those not following domains succession (just in case of multiple domains coding sequences) are discarded; iii) The third, using a custom Python2.7 script, retrieves the relevant information regarding domain position falling within or on a complete sequence to align each sequence or fragment of it against its corresponding domain. Then, it assembles all fragments to produce a multiple domain DNA alignment, by back translating protein alignments. The back-translated DNA alignments are merged together to construct the whole multiple domains alignment. During merge, aligned DNA blocks are added from 5' to 3', and if sites do overlap across domain, the 3' overlapping domain section is discarded and registered in a separate file as well as sites that do not match any domain. The present

workflow takes as input a multiple DNA sequences data set coding for single or multiple protein domains and gives a main output which is the final DNA multiple alignment (fasta and stockholm formats) together with the following description files: 1. DomainsSuccession.txt: the succession of domains coded by the input sequences. 2. DomainReport.txt: the position of each domain in the final multiple alignment useful for partitioned phylogeny models. 3. SitesReport.txt: the position of each domain per sequence. 4. Excluded_Sequence_IDs.txt: This file is produced in case one or more sequence do not code for the same number of domains and/or in the same order. 5. File/s with suffix hmmAligned: Amino acid and DNA back-aligned files for each domain.

**Results:** We used this workflow to align five mitochondrial genomes of five different Drosophila species. The genomes were retrieved from NCBI genome database. We validated the domains succession output of the workflow together with the correct genetic code and the different frames per protein domain with the original annotation of each genome. Additional testing on big dimension data sets including repetitive protein domains are still on-going. Our results indicate the accuracy and precision of this workflow regarding the detection of protein coding domain, their correct order and the suitability of the multiple alignment to conduct a robust phylogeny inference.

**Contact email:** balechbachir@gmail.com

Poster: Management - MGM3

**InSilico merging techniques allowed the identification and validation of a novel gene signature predicting patients' outcome on a six hundred neuroblastoma patients' dataset**

Cangelosi D(1), Becherini P(1), Oberthuer A(2), Sementa AR(3), Conte M(4), Garaventa A(4), Muselli M(5), and Varesio L(1)

*(1)Laboratory of Molecular Biology, Giannina Gaslini Institute, Genoa (2)Department of Pediatric Oncology and Hematology, University of Cologne, Cologne (3)Departments of Pediatric Pathology, Giannina Gaslini Institute, Genoa (4)Department of Hematology–Oncology, Giannina Gaslini Institute, Genoa (5)Institute of Electronics, Computer and Telecommunication Engineering, Italian National Research Council, Genoa*

**Motivation:** Cancer patient's outcome is written, in part, in the tumor microenvironment described by gene expression profile. Neuroblastoma tumor hypoxia is related to tumor aggressiveness and can be measured by gene expression. Different studies often measure gene expression profiles utilizing different platforms. For this reason, lacking of a sufficient large dataset is one of the main recurrent problems in microarray gene expression analysis. The consistent merging of data coming from different platforms can give new biological insights and stronger statistical support to the analysis with respect to the past. We report on a new neuroblastoma hypoxia-based gene signature (NB-hop) designed for outcome prediction validated on a merged 636 neuroblastoma patients' dataset.
**Methods:** Gene expression and clinical data of 636 patients, accessible in the R2: microarray analysis visualization platform (http:// r2.amc.nl) as separate multi-platform datasets, were consistently merged in a single cross-platform cohort utilizing the 'InSilico' merging package recently implemented in R. Five qualitative and quantitative validation methods were utilized to inspect the merging results. Signature design and bioinformatics analysis is based on Attribute Driven Incremental Discretization of continuous variables, Logic Learning Machine models for classification and feature selection. Hypoxia status of the tumors was defined by unsupervised k-means clustering based on the NB-hop signature, gene expression values. Surival analysis was performed with Kaplan-Meier method and Multivariate Cox model. Log rank test statistics was utilized to assess significance. P values smaller than 0.001 were considered significant.
**Results:** We merged three distinct multi-platform datasets into a single cohort of 636 patients. The consistency of the merging was demonstrated by different validation methods. We defined a new 7 genes signature (NB-hop) measuring tumor hypoxia but tailored to patient's outcome prediction. Multivariate Cox analysis demonstrated that NB-hop was an independent risk factor for NB patients. Unsupervised K-means clustering divided the patients in high (174) and low (462) tumor hypoxia. Kaplan-Meier analysis showed a significant low probability of survival in patients with highly hypoxic tumors ( $P< 0.00001$). Selected groups of high risk patients were further stratified by NB-hop in statistically significant groups. NB-hop was highly accurate (87%) in predicting NB patients' outcome. Furthermore, outcome prediction analysis identified a homogeneous poor outcome group defined by Stage 4, age > 12months, NB-hop high, that could be a prototype of the High Risk category. In conclusion, 'InSilico' merging techniques allowed the identification and validation of a novel gene signature predicting patients' outcome accurately on a consistent six hundred neuroblastoma patients' dataset. NB-hop is an independent risk factor that identify poor outcome patients benefitting from hypoxia targeted therapies.

Poster: Management - MGM4

**INTEGRATED ANALYSIS OF DNA COPY NUMBER AND GENE EXPRESSION DATA IN LUNG CANCER MODELS OF RESISTANCE TO TARGETED THERAPY**

Fustaino V(1,2), Presutti D(1), Cardinali B(1), Colombo T(3), Papoff G(1), Santini S(1), Lalli C(1), Giannini G(4), Brandi R(5), Arisi I(5), D'Onofrio M(5), Felici G(2) and Ruberti G(1)

*(1) Istituto di Biologia Cellulare e Neurobiologia, Consiglio Nazionale delle Ricerche (CNR) Monterotondo (RM), Italy (2) Istituto di Analisi dei Sistemi ed Informatica "Antonio Ruberti", CNR, Roma, Italy (3) Istituto per le Applicazioni del Calcolo, CNR, Roma, Italy (4) Department of Molecular Medicine, Università degli studi "La Sapienza", Roma, Italy (5) Genomics Facilty, European Brain Research Institute, Rita Levi-Montalcini, Roma, Italy*

**Motivation:** Tyrosine Kinase inhibitors (TKIs) constitute the most promising frontier of the last decade in the field of cancer treatment. However, the development of resistance mechanisms often makes the tumour insensitive to TKI-targeted therapy. Identification of the factors responsible for the emergence of resistance remains a major clinical challenge for cancer treatment. Non Small Cell Lung Cancer (NSCLC) with mutations activating the Epidermal Growth Factor Receptor (EGFR) provides a suitable model for studying TKI resistance mechanisms. The majority of patients, after an initial positive response to EGFR-TKI drugs (e.g. Erlotinib), develops resistance to treatment for the onset of EGFR secondary mutations and/or the activation of alternative signaling pathways. The present work aims to identify genes and signaling pathways involved in the development of TKI-resistance. To this end, we developed cellular models of NSCLC and designed cellular, molecular and bioinformatics analyses to correlate genotypes to phenotypes.
**Methods:** We generated six cell lines resistant to EGFR-TKI treatment by exposing NSCLC cell lines, harboring EGFR activating mutations in Exon 19, to Erlotinib (ERL) for 5 months. The selected resistant lines were then characterized at the cellular and molecular level to identify phenotypes associated with ERL-resistance. Briefly, immunofluorescence and confocal microscopy analysis as well as cell viability, wound healing and anchorage-independent cell growth assays were used to study cell morphology, cell growth, cell migration, and tumorigenic features. Sequence analysis, quantitative PCR and Western blot were used to characterize mutations, gene copy number variations (CNVs) and gene expression profiles of selected receptor tyrosine kinases (namely EGFR, KRAS, MET, AXL, AKT) that have been previously reported to be mutated or dysregulated in NSCLC. Next, CNVs and gene expression data were assayed for both parental and resistant derivative cell lines at genome-scale level by using microarray technology. We used R (cran.r-project.org/) and Bioconductor (www.bioconductor.org/) for data processing and analysis. The IGV genome browser (www.broadinstitute.org/igv/) and Gorilla web tool (cbl-gorilla.cs.technion.ac.il/) will be used, respectively, for interactive exploration and integrative analysis of the different data levels and for gene functional enrichment analysis.
**Results:** Preliminary cellular and molecular characterization of the NSCLC cell lines highlighted several features associated to an ERL-resistant phenotype, including altered cell migration and cell adhesion, Epithelial-to-Mesenchimal-Transition (EMT) features and metabolic dysregulation. In-depth analysis is still on-going. Moreover, comparison of array data between resistant and sensitive cells array data identified several chromosomal regions that undergone amplification or deletion, as well as lists of thousands of genes aberrantly expressed in selected drug-resistant cell lines, and hundreds of genes aberrantly expressed in all resistant cell lines. Interestingly, preliminary functional enrichment analysis of genes,

found consistently altered in all resistant cell lines, indicates their involvement in key biological processes. Further bioinformatics studies are in progress.
**Contact email:** v.fustaino@ibc.cnr.it

Poster: Management – MGM5

**OMEGA (Offline Mutation Easy Genetic Annotator), a software analysis pipeline designed for thorough and easy annotation of NGS output**

Mortola F(1), Zoppoli G(2), Izzo M(1), Bagnasco L(2), Carminati E(2), Garuti A(2), Fato MM(1)

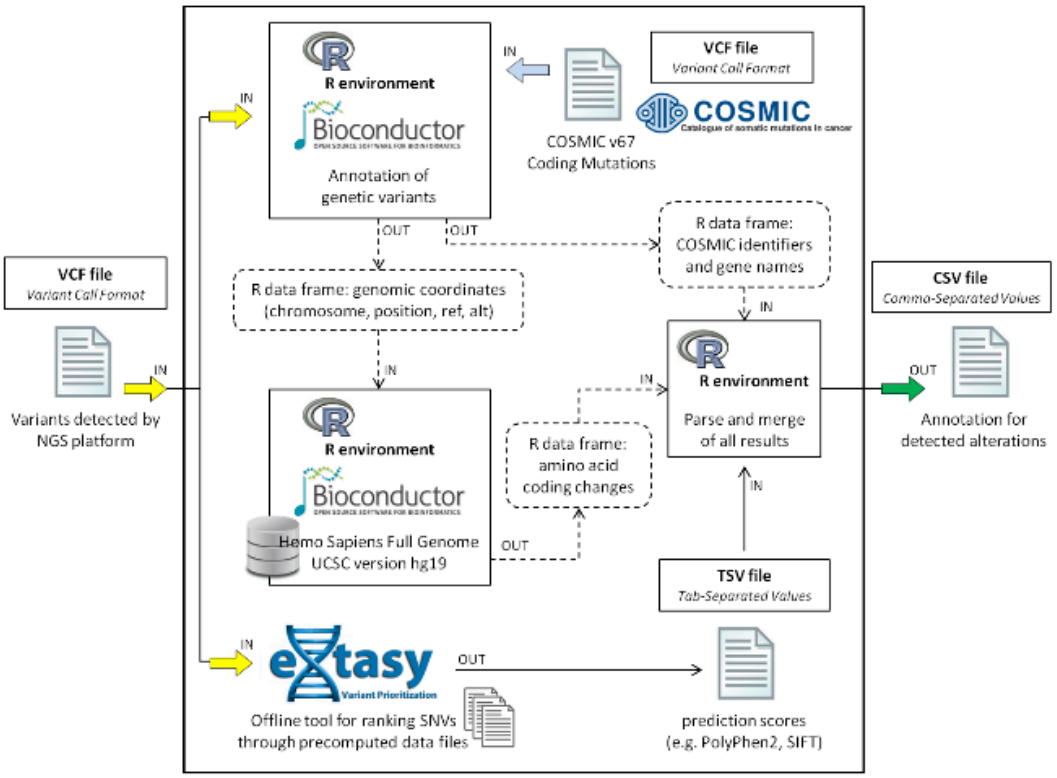
*(1) Department of Informatics, Bioengineering, Robotics and System Engineering (DIBRIS), University of Genoa, Italy (2) Department of Internal Medicine (DiMI), University of Genoa, Italy*

**Motivation:** Next-generation sequencing (NGS) has undergone tremendous developments in both basic and clinical research over the last few years. On the one hand, NGS devices have increased their throughput by orders of magnitude and, on the other, sequencing costs have dropped steadily. Recently, benchtop-size, price-affordable devices such as the Personal Genome AnalyzerTM by Life TechnologiesTM and the MiSeq® Benchtop Sequencer by Illumina® have entered the market, allowing small/medium size laboratories to generate NGS data over very short time frames. All these events have led to an initially unexpected bottleneck, which is neither the sequencing cost or its speed, but rather data analysis and interpretation. Primary (i.e. flow-processing and base calling) and secondary analyses (i.e. alignment to reference genome) are often taken care for by software developed by sequencing companies, but tertiary analysis (i.e. annotation and interpretation of results) is still a task mostly left to the NGS user. Especially with the progressive transitioning of NGS to clinical genomics, scientists without a dedicated background in bioinformatics, like biologists and physicians, find themselves to deal with obscure data tables. These spreadsheets are not interpretable without a great effort of annotation and prioritization of the most relevant findings, and it is therefore not surprising that commercial programs have flourished under such circumstances. For such reasons, we have developed OMEGA (Offline Mutation Easy Genetic Annotator), a free, user-friendly software analysis pipeline designed for thorough and easy annotation of NGS output.
**Methods:** OMEGA 0.9.1 leverages several freeware tools and biological datasets for genomic data annotation. Variant Call Format (VCF) 4.1 files are used as input data: since VCF is a specific file format used in bioinformatics as a post-alignment output, dedicated scripts written in R language allow to parse the records from its fields. The first step consists in extracting both the coordinates of single nucleotide variants (SNVs) or insertions/deletions from the VCF file produced by the NGS platform, and qualitative parameters for every tracked variation. Detected alterations are then searched in the COSMIC catalogue and annotated accordingly. R-Bioconductor software packages related to human full genome (UCSC version hg19) have been chosen to obtain information about gene nomenclature and amino acid coding changes for non-synonymous variants. The user-provided VCF file is also processed by a novel open source analysis tool, eXtasy-1.0, that returns several deleteriousness prediction scores upon specification of a human phenotype ontology (HPO) term required by the user (e.g. “breast cancer” or “neoplasm of the breast”). An R script finally merges all results in a single table and writes it within a CSV file, which is very simple to manage and to export in a user-friendly framework for final interpretation.
**Results:** We achieved our primary goal by returning a user-friendly interpretation file of basic NGS aligned data regarding nucleotide variants in human DNA, within a simple but complete genome annotation for every variant detected by the NGS platform. The final output includes information such as genomic coordinates, gene names, amino acids variants, COSMIC identifiers, and prediction scores. Moreover, the qualitative data provided by the NGS platform-specific software in the output VCF is retained, providing an

assessment of the reliability of the analysis at a glance. Our software pipeline, OMEGA, is still in its beta version (currently 0.9.1). We are currently expanding OMEGA to include more complete parameters such as SNP IDs, so our work is still in progress. In addition, we are currently defining the graphic-user interface, with the final goal to provide both an offline (standalone) and a browser-based version for our software.

**Contact email:** fra.mortola@gmail.com

**Supplementary information:**http://www.bioconductor.org/ http://homes.esat.kuleuven.be/~bioiuser/eXtasy/ http://cancer.sanger.ac.uk/cancergenome/projects/cosmic/ fra.mortola@gmail.com zoppoli@gmail.com (Zoppoli G et al PNAS 2012)



Poster: Management – MGM6

**GMO-Scan: a tool for fast detection of GMO sequence elements.**

Petrillo M(1), Angers A(1), Patak A(1), Kreysa J(1)

*(1) Molecular Biology and Genomics Unit, Joint Research Centre, European Commission, Ispra (VA), Italy*

**Motivation:** The Joint Research Centre (JRC) of the European Commission is host to the European Union Reference Laboratory for Genetically Modified Food and Feed (EU-RL GMFF). The core tasks of the EU-RL GMFF are the scientific assessment and validation of detection methods for GM Food and Feed as part of the EU authorisation procedure. In addition, the EU-RL GMFF maintains the Central Core DNA Sequences Information System (CCSIS), an in-house molecular EU database containing GMO sequences and the respective sequence annotation data, either submitted to the EURL-GMFF by applicants or identified within public databases (EMBL, Genbank, Patents, etc.). Those sequences very often share specific sub-regions as deputed to the same function. The availability of sequence information is fundamental both for the validation of detection methods and for the identification of unauthorised GMOs. A web-based application has been developed for fast detection of potential GMO-related sequences, based on the CCSIS sequence information and Megablast.
**Methods:** Sequences of the CCSIS database are extensively annotated and the whole set of metadata is stored into a relational database. Briefly, each GMO event is composed of several different GMO elements, i.e. specific sub-regions of the transgenic cassette, like promoters, terminators or coding sequences. Elements are in turn classified and grouped, i.e. elements sharing the same origin and/or specific sub-regions are assigned to a specific element tag. An automatic script-based procedure has been developed, that for each of the element tag, retrieves the sequences of element group member, align them by running ClustalW and from the alignment a consensus is generated. All the consensus sequences are then transferred as multi-fasta file to a local file server repository, available for other servers automatic retrieving. The application is a web-based tool that allows a similarity comparison of a given DNA sequence with the set of GMO elements' consensuses by Megablast, a tool developed at NCBI and installed locally. The platform is built on three main servers: a server that generates the GMO elements' consensuses multi-fasta file as described above, a server that hosts a Ruby on Rails web application that acts as the interface with the user, and a high-performance server with the bioinformatics scripts (written in PHP) which, when invoked by the web application, runs Megablast and returns the result. This result is then parsed by the web application and displayed as graph in the Blast output style. On the high-performance computer server a script, running at cron time, takes care of updating the database blast index, retrieving the sequences from the local file server repository.
**Results:** The GMO-Scan application is currently deployed in the intranet of the JRC and is used both to assist screening of sequences derived from suspected GMO sequences and to quickly annotate new incoming sequences to the EU-RL GMFF. In the future, we are planning to use the consensus of the GMO elements as query on the patent database, in order to create a patent subset of GMO-related sequences that can be used as Blast database set within the same application.
**Contact email:** mauro.petrillo@jrc.ec.europa.eu

Poster: Management - MGM7

**adLIMS: a customized open source software that allows bridging clinical and basic molecular research studies**

Calabria A(1), Spinozzi G(1), Benedicenti F(1), Tenderini E(1), Biasco L(2), Montini E(1)

*(1) HSR-TIGET - The San Raffaele Telethon Institute for Gene Therapy, Safety of gene therapy and insertional mutagenesis Unit; San Raffaele Scientific Institute, Division of Regenerative medicine, Stem cells, and Gene therapy - Milan, Italy (2) HSR-TIGET - The San Raffaele Telethon Institute for Gene Therapy; San Raffaele Scientific Institute, Division of Regenerative medicine, Stem cells, and Gene therapy - Milan, Italy*

**Motivation:** Many biological laboratories dealing with genomic samples are facing the problem of sample tracking, both for pure laboratory management and efficiency, and for internal policies, such as Good Laboratory Practices (GLP). Our laboratory exploits PCR techniques and next-generation sequencing (NGS) methods, to perform high-throughput integration site monitoring in different clinical trials and scientific projects, based on the delivery of therapeutic genes by viral vectors integrating into the genome of target cells. We process around 1500 samples/year resulting in hundreds of millions of sequencing reads, requiring automation and posing new challenges in data storage, monitoring of sample process and computational tools for analyses. Thus we need to standardize data management and tracking systems, built on a scalable, flexible structure with an easily accessible and web based interface, what is usually called Laboratory Information Management System (LIMS).
**Methods:** We first collected end-users requirements, composed by desired functionalities of the systems and graphical user interfaces (GUI). Then we evaluated available tools that could address our requirements, spanning from pure LIMS to CMS (Content Management Systems) up to enterprise information systems. Native LIMSs are often very expensive and/or could lack the flexibility and scalability needed to adjust seamlessly to the frequently changing protocols employed. We derived similar conclusions after analysing some of the most used and developed CMS, such as Plone, Drupal or Bika, LabKey. We chose ADempiere ERP (Enterprise Resource Planning under GPL license written in Java J2EE, with Model-View-Controller design pattern and database-driven logic) given its technological advances, such as the high usability (web and mobile interfaces) and modularity that grants high scalability. In order to translate requirements into functionalities and GUI into ADempiere, we designed an extension of the database, applying the required functionalities and user's policies resulting in a scalable system that natively supports Java fat client and web interface. For each end-user interaction and functionality, we designed a custom workflow with dedicated GUIs. As GUI's design principle to reduce users' errors, we decided to maximize the use of itemized form elements (dropdown menus, checkboxes, etc.) that only authorized users can edit.
**Results:** We implemented adLIMS currently in use in our laboratories and our end-users validated it verifying functionalities and GUIs through test cases for PCRs samples and pre-sequencing data. We observed a clear improvement in terms of efficiency, data accessibility and data quality by using adLIMS and an increased simplification on sample reporting and tracking with respect to traditional data storage methods. adLIMS not only keeps track of samples and their products, avoiding user's typos, errors, or wrong barcode assignment, but also eases dealing with laboratory automation, allowing sample data multiplexing and parallel processing. Here we reported two scenarios related to two workflows with different scopes and actors, thus requiring different user's policies and sample activities. The first

workflow concerns the interaction with clinicians and patient's samples. Once a clinician harvested a patient's sample (i.e. peripheral blood), she/he reported sample data into adLIMS that easily supported and guided the user to select and input the sample information with ad hoc forms, from the project ID to the patient ID and finally to the new sample creation with data of cell populations sorted from peripheral blood. The clinician user profile is authorized to modify a specific project and patient. In the second scenario a technician is required to process a patient cell population, for which she/he has got authorizations, (1) performing a PCR, (2) storing the gel images, (3) preparing a sequencing pool, and (4) exporting a report. The later activities are directly supported by the adLIMS automation that allows both the creation of parallel processes often required in case of laboratory automation (PCRs liquid handler with 96 wells/plate), and a direct definition of a sequencing pool in which all samples are automatically barcoded. The final step of report generation from the selected pool is directly data-driven allowing exporting results in a wide range of file formats (PDF, XLS, CSV, etc.) through the Jasper Report library. adLIMS can be adapted to other laboratories with limited efforts, potentially without any coding skills by defining the business and view logics (workflows and GUI) and creating the data structures (PostgreSQL or Oracle database schema and tables). Then ADempiere automatically generates interfaces and data views. Moreover adLIMS can natively be extended to incorporate ERP solutions, such as CRM, supply chain management, billing and accounting, integrated features that are critical for many laboratory facilities.

**Contact email:** spinozzi.giulio@hsr.it

Poster: Management – MGM8

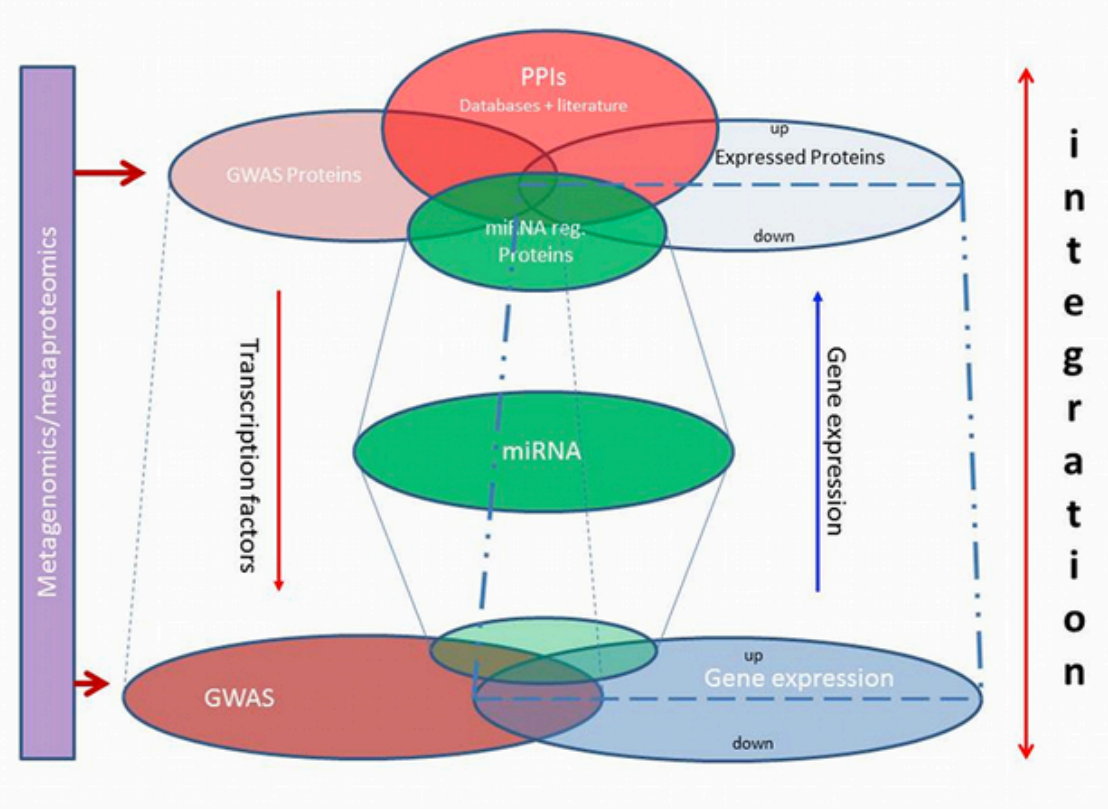
**Multi-omic Data Integration for Rheumatoid Arthritis**

Tieri P (1,2) , Zhou XY (2), Zhu (2), Nardini C (2)

*(1) IAC Istituto per le Applicazioni del Calcolo, CNR National Research Council, Rome, Italy. (2) Group of Clinical Genomic Networks, Key Laboratory of Computational Biology, CAS-MPG Partner Institute for Computational Biology, Shanghai Institutes for Biological Sciences, Shanghai, PR China.*

**Motivation:** The etiology of rheumatoid arthritis (RA), an autoimmune chronic inflammatory disease [1], has long been challenging medical research due to its genetic complexity and multiple environmental influences. It is now clear that capturing the essence of the disease to provide appropriate cure requires a wide-spectrum systemic approach in order to take under a single, unified view as much as possible of the amount of the molecular knowledge available. This can nowadays be achieved by integrating heterogeneous high-throughput data sets, via so called multi-omic approaches [2-4]. Moving from the analysis of such RA omic-integrated molecular landscape and using network approaches it is possible to grasp relationships that would be otherwise hidden (system's emergent properties) and key architectural elements, which better clarify the ground of the disease. Ultimately, the integration of disease susceptibility genes, gene expressions, protein interaction networks as well as other available omic layers together can lead to the identification of a reduced number of marker molecules can be experimentally tested as drug targets [5].
**Methods:** Multi-omic data integration represents a well known challenge in modern computational biology. Development of efficient integrative methodologies will be critical in the future, and efforts have been already spent in this direction [6, 7]. However, given the issues related to databases curation and uniformation [8], we believe manual approaches are crucial at least at some level of the analysis refinement, in particular to take into account various levels of association between a disease and its molecular counterpart. The map presented here assembles up-to-date genomic, epigenomic, transcriptomic, post-transcriptomic, proteomic and host-microbiome data related to RA, and integrates such information at the functional level of protein-protein interactions (PPIs). It has to be stressed that several studies already employed PPI networks to find relationship and to interpret gene and protein seed sets [9-11]. These added layers build a novel complex map, which permit to gain, via network analysis, truly novel insights in the form of systemic disease targets and/or markers.
**Results:** This approach allowed to prioritize a set of proteins based on previous knowledge and in-depth integrative network analysis, providing a new reference framework for future experimental approach to RA. The novelty of the present work lies in: i) the condensation into a single, analytical picture the molecular information available to date on RA; ii) the provision of an easily consultable and extendable reference map for researchers and clinicians in the field; iiI) the contribute to the general debate about data integration by offering details on our methodology, and to the area of complex inflammatory diseases, by providing specific examples of data choice and operational results.
**Contact email:** p.tieri@iac.cnr.it
**Supplementary information:**1. Gregersen, P.K., Genetics of rheumatoid arthritis: confronting complexity. Arthritis Res, 1999. 1(1): p. 37-44. 2. Joyce, A.R. and B. Palsson, The model organism as a system: integrating 'omics' data sets. Nat Rev Mol Cell Biol, 2006. 7(3): p. 198-210. 3. Palsson, B. and K. Zengler, The challenges of integrating multi-omic data sets. Nat Chem Biol, 2010. 6(11): p. 787-9. 4. Louie, B., et al., Data integration and genomic

medicine. J Biomed Inform, 2007. 40(1): p. 5-16. 5. Silverman, E.K. and J. Loscalzo, Network medicine approaches to the genetics of complex diseases. Discov Med, 2012. 14(75): p. 143-52. 6. Searls, D.B., Data integration: challenges for drug discovery. Nat Rev Drug Discov, 2005. 4(1): p. 45-58. 7. Bebek, G., et al., Network biology methods integrating biological data for translational science. Brief Bioinform, 2012. 13(4): p. 446-59. 8. Tieri, P. and C. Nardini, Signalling pathway database usability: lessons learned. Mol Biosyst, 2013. 9(10): p. 2401-7. 9. Dittrich, M.T., et al., Identifying functional modules in protein-protein interaction networks: an integrated exact approach. Bioinformatics, 2008. 24(13): p. i223-31. 10. Jin, G., et al., The knowledge-integrated network biomarkers discovery for major adverse cardiac events. J Proteome Res, 2008. 7(9): p. 4013-21. 11. Kim, T.Y., H.U. Kim, and S.Y. Lee, Data integration and analysis of biological networks. Curr Opin Biotechnol, 2010. 21(1): p. 78-84.



Poster: Metagenomics - MTG1

**BioMaS: a web service for Bioinformatic analysis of Metagenomic AmpliconS**

Fosso B(1,2), Giacinto Donvito(3), Alfonso Monaco(3),Pasquale Notarangelo(3), Monica Santamaria(2), Graziano Pesole(1,2)

*(1) Department of Biosciences, Biotechnology and Pharmacological Sciences, University of Bari, Bari. (2) Institute of Biomembranes and Bioenergetics, CNR, via Amendola 165/A, 70126 Bari, Italy. (3) Institute national of Nuclear Physics, Via Orabona, 4, 70126 Bari, Italy.*

**Motivation:** Substantial advances in microbiology, molecular evolution and biodiversity have been carried out in recent years thanks to Metagenomics which allows to directly access to composition and functions of mixed microbial communities in any environmental niche. Currently, a very serious bottleneck is represented by the availability of convincing systems which allows a friendly and comprehensive large-scale bioinformatic analysis of metagenomic reads produced by Next Generation Sequencing technologies. By developing BioMaS (Bioinformatic analysis of Metagenomic AmpliconS) we would like to equip the biomolecular researchers involved in taxonomic studies of environmental microbial communities with a simple and versatile workflow, comprehensive of all the fundamental bioinformatic analysis steps, from raw sequence data arrangement to final taxonomic identification, to be used in target-based Metagenomics experiments. Here we present the implementation of BioMaS into a public web service available at http://webgateway.ba.infn.it:9999.
**Methods:** BioMaS has been developed as a modular pipeline based on third-party tools and ad hoc developed Python and Bash scripts. It has been implemented as a web-service on the INFN/UNIBA computing resources developed in the PON ReCaS, a project for the infrastructure upgrading. High level SaaS services are applied to provide to the end user the ability to use the BioMaS components that are already opportunely configured and optimized to run on the infrastructure. The interface of access to these services is based on the concept of web services. SOAP and REST protocols are available to access both to the computational resources of the farm available in Bari and to the EGI geographically distributed infrastructure.
**Results:** In its current version, BioMaS allows the analysis of both bacterial and fungal environments and two alternative path can be performed in order to process data obtained by Roche 454 or Illumina platforms. The access to the web-service is allowed through a free registration, by using a valid email address. In order to start the BioMaS computation user needs to indicate the sequencing platform, submit the sequence file (sff file for 454 or paired-end fastq files for Illumina platform) and associate a label (base name) to the starting job. The process progress can be monitored by simply reporting the same email address used for the registration in the "monitoring jobs" section. At the end of the analysis the user will receive a mail including the indication to download a compressed folder containing the produced data: a html report showing general information about the performed analysis and six interactive pie-charts summarizing the inferred taxonomy at the different rank (from species to phylum). Moreover a high-resolution taxonomical complexity of the studied microbiome is represented by a graphical tree accessible through a link in the report.
**Contact email:** bruno.fosso@gmail.com

Poster: Metagenomics - MTG2

**Identification of new hydrolases from deep sea metagenomes**

Placido A (1, 2), Tran H (2), Ceci LR (1), Horner DS (3), Chiara M (3), Chernikova TN (2), Golyshina OV (2), Pesole G (1, 4) and Golyshin P (2)

*(1) Institute of Biomembranes and Bioenergetics, National Research Council, Bari, Italy; (2) Environmental Centre Wales, Bangor University, Wales, UK; (3) Department of Biosciences, Milan University, Italy; (4) Department of Biosciences, Biotechnologies and Biopharmaceutics, Bari University, Italy.*

**Motivation:** In the last few years, extremophile microbes, mainly archea and bacteria, are becoming a source of novel and unique enzymes for a broad range of potential applications in biotechnology industry. Extremophiles live under chemical and physical extreme conditions. Their enzymes catalyze reactions in conditions which inhibit or denature their non-extreme counterparts. Hydrolases represent a class of enzymes involved in fundamental cellular biochemical processes. Among them, lipases, esterases and cellulases are the most requested enzymes from industry due to their relevant commercial applications in food processing, textile, detergents, dairy, pharmaceuticals and generation of sustainable energy. In the present work metagenomic DNA from extremophile prokaryotic cells of deep sea sediments (Eolie islands, Italy) have been isolated, cloned in fosmid vector, screened for activity, sequenced by NGS platforms and assembled and annotated by bioinformatic approaches to identify ORFs encoding for novel lipases, esterases and cellulases.
**Methods:** 0
**Results:** The environmental DNA has been isolated and cloned into pCC2FOS vector. Activity screening of recombinant fosmids has been carried out onto agar plates containing either tributyrin or carboxymethylcellulase as substrates. Seven positive clones have been detected. Fosmid inserts of 40-45 Kb have been sequenced by Miseq Illumina platform. A total of twenty ORFs encoding for putative lipases, esterases and cellulases have been identified by using the MetaGeneMark software, available at the GeneMark web site (http://opal.biology.gatech.edu/). Identified orfs have been subsequently cloned into expression vectors for future characterization.
**Contact email:** a.placido@bangor.ac.uk

Poster: Molecular Evolution – ME1

**A customizable gene selection and clustering software for phylogenomics and duplication/recombination analyses**

Comandatore F(1), Gaiarsa S(1), Bandi C(1), Sassera D(1)

*(1) DIVET Università degli Studi di Milano, Milano*

**Motivation:** Genomic data represent the most important source of information to unveil the phylogenetic relationships among prokaryotic lineages, through the use of phylogenomic approaches. The selection of a robust gene dataset is one of the most important steps of phylogenomic analysis. Ideally, each gene selected to be used in a phylogenomic analysis should provide evolutionary signal congruent with the evolution of the studied organism. Thus, the evolutionary signal provided by these selected genes should be affected by noise as little as possible. Recombination and duplication events are the main source of evolutionary signal noise in gene sequences. The recombined genes present one or more regions of the sequence that are imported from others prokaryotes possibly providing noisy phylogenetic signals. Duplicated genes should be avoided in phylogenomic analyses due to the difficulty of generating orthologous clusters and because they often present high levels of mutational rates. Albeit problematic for phylogenomic analyses, recombined and/or duplicated genes can be of interest due to their possibility to evolve fast, possibly developing novel functions and adaptations. In order to address these issues, we developed a novel, robust and customizable bioinformatics software to perform gene selection and clustering for phylogenomic and duplication/recombination analyses.
**Methods:** The proposed software work-flow presents five main steps: 1 Input formatting If the input is a collection of .fna files the software performs an orf calling process with the Prodigal software on each input file (obtaining the corresponding .ffn and .faa files). If the input is a collection of .ffn files, the corresponding .faa files are obtained. 2 Duplication analysis (on each .faa) The software processes each .faa file to detect clusters of duplicated genes with Single Best Hit Blast algorithm. The user can set the software to decide how to process each of the detected duplicated sequences clusters. Two settings can be used: a) sequences are aligned with Muscle and merged in a single consensus sequence to produce a new .faa file containing all the not duplicated gene sequences and all the generated consensus sequences (Merge setting); b) all the sequences belonging to a cluster of duplication are removed from the .faa input file (Remove setting). The user can also decide to skip this duplication analysis step. 3 Orthologues clusters detection The software detects pairs of orthologues genes among all the .faa files with Double Best Hit algorithm. The result is organized in a graph, with a node for each gene and an arch for each orthologues gene pair. The obtained Orthologues graph is then subjected to clustering analysis. Two orthologues clustering algorithms can be used: a) Markov Cluster Algorithm (MCL), one of the most frequently used algorithm in orthologues gene clustering analysis (e.g. implement in the OrthoMCL software), it applies a stochastic approach to solve the clustering problem; b) A deterministic algorithm, which detects all the unique cliques in the graph: 1) the biggest clique in the graph is detected (D = clique dimension); 2) the nodes belonging to these cliques of dimension D are removed from the graph and the corresponding genes are organised in orthologues clusters; 3) D is decreased of one unit and the point 2 and 3 are repeated until D>0. If the software was set to merge duplicated sequences during the duplication analysis (step 2), the user can set the software to split each consensus (merged) sequence within each orthologues cluster re-obtaining all the original duplicated sequences. The merge setting in step2 allows to obtain gene clusters containing duplicated genes,

reducing the noise introduced by these duplicated sequences in orthologues gene clustering analysis. 4 Recombination analysis The software implements the PhiPack software to detect orthologues clusters with signals of recombination. The user can set the software to remove these recombined orthologues gene clusters or select them. The user can also decide to skip this recombination analysis step. 5 Concatenation If the software was used with the "remove setting" at step 2, the orthologous clusters obtained from the previous steps and shared among all organisms will be used to produce a concatenate useful for phylogenomic analyses. In order to do so the software uses the Muscle and Gblocks softwares.
**Results:** We developed a user-friendly and customizable tool that allows to perform gene selection and clusterization with different approaches. This software can be useful for phylogenomic analyses as well as for detection of orthologues clusters and of recombined and/or duplicated genes.
**Contact email:** francesco.comandatore@unimi.it

Poster: Molecular Evolution – ME2

**Functional Divergence and Evolutionary Turnover in Mammalian Phosphoproteomes**

Freschi L(1,2,3), Osseni M(1,2,3), Landry CR(1,2,3)

*(1) Département de Biologie, Université Laval, Québec, Canada (2) Institut de Biologie Intégrative et des Systèmes (IBIS), Université Laval, Québec, Canada (3) PROTEO, The Quebec Research Network on Protein Function, Structure and Engineering, Université Laval, Québec, Canada*

**Motivation:** Protein phosphorylation is a key mechanism to regulate protein functions. However, the contribution of this protein modification to species divergence is still largely unknown. Here, we studied the evolution of mammalian phosphoregulation by comparing the human and mouse phosphoproteomes. These two phosphoproteomes are the ones for which we have the greatest amount of phosphoproteomics data between closely related species. Further, mouse is the best model system currently available to study human biology and diseases. By studying how human and mouse diverge in their phosphoregulation we can make a step forward in our understanding of how mouse biology translates to human biology.
**Methods:** 0
**Results:** We found a general conservation of phosphorylation sites between the human and mouse phosphoproteomes. Indeed, 84% of the 88,413 positions that are phosphorylated in one species or the other are conserved at the residue level (conserved sites). Twenty percent of these conserved sites are phosphorylated in both human and mouse, suggesting that purifying selection is preserving phosphoregulation at these positions. The other 80% of conserved sites are differentially phosphorylated in the two species. This difference in phosphoregulation can result from false-negative identifications due to incomplete experimental coverage, false-positive identifications and non-functional sites. However, we showed that for at least 5% of conserved sites the divergence in phosphoregulation is unlikely to be the result of these confounding factors. These sites represent novel candidate sites that have the potential to explain differences between human and mouse signalling networks that do not depend on the divergence of orthologous residues. Furthermore, we identified 20,146 sites where to a phosphorylation site in one species corresponds a non-phosphorylatable residue in the other one. These cases represent clear differences in protein regulation. Finally, recent studies suggest that phosphorylation sites can shift position during evolution, leading to configurations in which pairs of divergent phosphorylation sites are functionally redundant. We identified more than 100 putative such cases, suggesting that divergence in amino acid does not necessarily imply functional divergence when comparing phosphoproteomes. Overall, our study provides an advanced comparison of mammalian phosphoproteomes and a framework for the study of their contribution to phenotypic evolution.
**Contact email:** luca.freschi@bio.ulaval.ca

Poster: Molecular Evolution - ME3

**Transcription factors of the MocR family: in silico comparative analysis**

Milano T(1), Pascarella S(2).

*(1)Department of Anatomy, Histology, Forensic Medicine and Orthopaedics, Sapienza University of Rome, Rome (2)Department of Biochemical Sciences, Sapienza University of Rome, Rome*

**Motivation:** The MocR/GabR subfamily of chimeric proteins is widespread in bacteria and regulates a variety of biological processes. All members of this subfamily are typically composed of a winged helix-turn-helix (HTH) DNA-binding domain and a putative aminotransferase domain (AAT), belonging to the superfamily of the pyridoxal-5'-phosphate (PLP) enzymes of fold type I. The HTH-containing regions of these proteins vary from 60 to 120 residues in lenght and are similar to the winged HTH regions of bacterial transcriptional regulators of the broad GntR family. Little is known about the function of the AAT domain and about its ability to retain some catalytic activity. Moreover, the transcriptional regulatory pathways in which the MocR proteins might be involved are not known. So far, only a few MocR regulators have been characterized experimentally. One of these, PdxR is involved in the regulation of the divergently oriented pdxST operon, coding for the subunits of pyridoxal-5'-phosphate synthase in Corynebacterium glutamicum. Our research projects are aimed at the characterization of this regulator family through the following sub-tasks: - give an overview of the distribution of MocR factors among the different taxonomical bacterial divisions; - try to assess the degree of conservation of the main structural features of the AAT domain and compare them to those of the freestanding PLP enzymes; - predict which genes are controlled by these regulators; - identify, in the intergenic regions, the transcription factor binding sites (TFBSs).
**Methods:** An ad-hoc Profile-HMM was built from a multiple sequence alignment of MocR proteins. HMM searches were carried out on the complete proteomes of each phylum reported in the taxonomy division of UniProt databank to collect most of the members of the MocR family. A Python script was used to extract information about the location, into the genome, of gene coding MocR and their neighbors genes. All sequences of neighboring genes were compared against the Pfam database to locate known domains. Pfam domain frequency was calculated as the fraction of the proteins having that particular Pfam domain. This frequency was calculated for each Pfam domain for all proteins in the dataset. Based on these frequencies and knowledge available in the literature, a subset of convergently transcribed pairs was selected. Pattern search and comparative genomics methods were used to detect direct and inverted repeat sequences in the intergenic regions between MocR and neighbors genes. Homology modelling techniques were applied to predict the three-dimensional structure of representative MocR members based on the B. subtilis GabR with PDB code 4MGR. The most interesting regulators will be selected for experimental characterization.
**Results:** Our results indicate that presence of MocR regulators is rather heterogeneous, even within the same bacterial phylum. In general, the genomic distribution is, as expected, highly correlated to the size of the genome. The degree of conservation of the most significant residues potentially interacting with PLP at the active site of the fold-type I domain was analyzed. The residues predicted to interact with PLP are mostly conserved, altough several phyla display stronger conservation than other. Given the low overall degree of similarity, MocR proteins appear to undergo strong selective pressure in order to maintain the residues essential for cofactor binding. This may suggest that this function is

important for the physiological role the regulators play in the cell. A large variability is also observed for the arrangement of neighboring genes, and this reflects the wide range of processes in which they are involved. However, we found a significant frequency of Pfam domains related to the proteins coded by pdxS and pdxT genes. Further analyses were carried out only on this subset of genes and we have identified at least a directed repeat motif of six nucleotides in the intergenic regions between MocR proteins and pdxST operon.
**Contact email:** teresa.milano@uniroma1.it

Poster: Molecular Evolution – ME4

**Conservation of Meningococcal Antigens in the Genus Neisseria**

Muzzi A(1), Mora M(1), Pizza M(1), Rappuoli R(1) and Donati C(2)

*(1) Novartis Vaccines and Diagnostics Research Center, Via Fiorentina 1, 53100 Siena, Italy (2) Centre for Research and Innovation, Fondazione Edmund Mach, Via E. Mach 1, 38010 San Michele all'Adige, Trento, Italy*

**Motivation:** The upper respiratory tract of healthy individuals is a complex ecosystem colonized by many bacterial species. Among these, there are representatives of the genus Neisseria, including Neisseria meningitidis, a major cause of bacterial meningitis and sepsis. Given the close relationship between commensal and pathogenic species, a protein-based vaccine against N. meningitidis has the potential to impact the other commensal species of Neisseria. For this reason, we have studied the distribution and evolutionary history of the antigen components of a recombinant vaccine, 4CMenB, that recently received approval in Europe under the commercial name of Bexsero®. We found that fHbp, NHBA, and NadA can be found in some of the human commensal species and that the evolution of these antigens has been essentially shaped by the high rate of genetic exchange that occurs between strains of neisseriae that cocolonize the same environment.
**Methods:** 0
**Results:** Three meningococcal proteins, factor H-binding protein (fHbp), neisserial heparin-binding antigen (NHBA), and N. meningitidis adhesin A (NadA), have been described as antigens protective against N. meningitidis of serogroup B, and they have been employed as vaccine components in preclinical and clinical studies. In the vaccine formulation, fHbp and NHBA were fused to the GNA2091 and GNA1030 proteins, respectively, to enhance protein stability and immunogenicity. To determine the possible impact of vaccination on commensal neisseriae, we determined the presence, distribution, and conservation of these antigens in the available genome sequences of the genus Neisseria, finding that fHbp, NHBA, and NadA were conserved only in species colonizing humans, while GNA1030 and GNA2091 were conserved in many human and nonhuman neisseriae. Sequence analysis showed that homologous recombination contributed to shape the evolution and distribution of both NHBA and fHbp, three major variants of which have been defined. fHbp variant 3 was probably the ancestral form of meningococcal fHbp, while fHbp variant 1 from N. cinerea was introduced into N. meningitidis by a recombination event. fHbp variant 2 was the result of a recombination event inserting a stretch of 483 bp from variant 1 into the variant 3 background. These data indicate that a high rate of exchange of genetic material between neisseriae that colonize the human upper respiratory tract exists.
**Contact email:** alessandro.muzzi@novartis.com

Poster: Molecular Evolution – ME5

**GeoKS: an automatic tool for estimating burnin and assessing convergence in MCMC Phylogenetic Inferences.**

Saverio Vicario(1), Simone Battagliero(2), Giuseppe Puglia(3), Teresa Maria Creanza(4), Giancarlo Tria(4), Gaetano Scioscia(4) and Pietro Leo(4)

*(1). Consiglio Nazionale delle Ricerche – Istituto di Tecnologie Biomediche – Sede di Bari, via Amendola 122/D, 70126 Bari, Italy. E-mail: saverio.vicario@ba.itb.cnr.it. (2). Exprivia S.p.A., viale A. Olivetti 11/A, 70056 Molfetta, Italy: since Nov 9th, 2010; IBM Italia S.p.A., GBS BAO Advanced Analytics Services and MBLab, IBM Italia S.p.A., via Pietro Leonida Laforgia 14, 70125 Bari, Italy: until Nov 8th, 2010. E-mail: s.battagliero@googlemail.com. (3). Department of Biological, Geological, and Environmental Sciences, University of Catania, via A. Longo 19, 95125 Catania, Italy: since July 01, 2010; IBM Italia S.p.A., GBS BAO Advanced Analytics Services and MBLab, IBM Italia S.p.A., via Pietro Leonida Laforgia 14, 70125 Bari, Italy. E-mail: giuseppe.puglia@unict.it. (4). IBM Italia S.p.A., GBS BAO Advanced Analytics Services and MBLab, IBM Italia S.p.A., via Pietro Leonida Laforgia 14, 70125 Bari, Italy. E-mail:, t.creanza@libero.it, giancarlo.tria@gmail.com, pietro_leo@it.ibm.com.*

**Motivation:** An open problem of Bayesian methods for phylogenetic inference is the burn-in detection and convergence assessment. In fact, several solutions were proposed (Nylander et al. 2008, Beiko 2006) but none give objective criterion to accept or reject convergence of the Markov Chain Monte Carlo used to perform phylogenetic inference. This puts several limitations in the use of Bayesian phylogenetic inference within an automatic pipeline of analysis. Further, it's still very difficult to diagnose which aspect, poor mixing or too few generations, caused the lack of the convergence, given the difficulty to apply all the statistical tools developed for general MCMC integration (that involves scalar parameters) to phylogenetic inference (where the main parameter to be estimated is the topology).
**Methods:** We propose an automatic method for addressing burn-in and convergence assessment, grounded on the theoretical framework given by the space of phylogenetic trees (given by topology and branch lengths). This space gives the possibility to calculate distances between trees but also to estimate a mean tree, that unlike a consensus tree includes also branch lengths. By means of tree distances and the mean tree, we estimate the autocorrelation across a tree series, which could be used to compare the efficiency of mixing across different runs and models. Further, with this statistics we are able to estimate the effective sample size of the sampled trees and to calibrate a Kolmogorov-Smirnov test that detects differences between independent runs of the MCMC integration. To increase the detection power, before the convergence test we discard from each run an initial number of generation, given by our burn-in detection algorithm.
**Results:** We tested the power of the convergence test by using two different strategies: 1) by artificially increasing the distances between two sets of trees 2) by observing the effective number of trees necessary in order to detect convergence on several slices of increasing size of a run that reached convergence thanks to a very large number of generations. In the latter tests we observed the empirical rule, which appears to be dataset-independent, that 300 effective trees should be obtained before convergence can be detected with no bias. The algorithm is implemented in a software written in C available as a web application (http://mblabproject.it/geoks/) or as a step of a larger pipeline. The web application is able to calculate the geodesic distance, the mean tree, the burn-in and the p-value of convergence; in the second deployment, the tool is included in a series of workflows for phylogenetic inference with the program MrBayes

(http://www.myexperiment.org/packs/371.html) that can be played on the portal of the FP7 project BioVeL (http://www.biovel.eu/). This web service implements only the burn-in estimation and convergence test procedures, but is wrapped in a Python script that implements a Fisher procedure to combine p-values from several pairwise comparison of runs: this allows to generalize the approach to more than two runs.
**Contact email:** saverio.vicario@ba.itb.cnr.it

Poster: Other – OT1

**Mapping the hydropathy of amino acids based on their local solvation structure**

Milanetti E(1), Raimondo D(3), Bonella S(3), Ciccotti G(3), Tramontano A(2,3)

*(1) Center for Life Nano Science @Sapienza, Istituto Italiano di Tecnologia, Sapienza University of Rome, P.le A. Moro 5, 00185, Italy (2) Institut Pasteur—Fondazione Cenci Bolognetti, Sapienza University, Rome, P.le A. Moro 5, 00185, Italy (3) Department of Physics, Sapienza University, Rome, P.le A. Moro 5, 00185, Italy*

**Motivation:** An accurate determination of the hydropathy of amino acids side chains in peptides and proteins is important in a number of key biological problems including protein folding, protein/protein and peptide/protein interaction. However, there is no general consensus on hydropathy quantitative measure. many hydrophobicity scales exist, often exhibiting quite different rankings. Furthermore, both experimental and theoretical methods have a number of limitations due to a series of factors, for example the high times and costs of the experiments or the choice of the dataset. Aim of the project is to characterize the hydropathy properties of the 20 natural amino acids using theoretical and computational method, focusing the study on the local solvation structure around the solute.
**Methods:** To make progress towards a systematic classification, we analyzed amino acids' hydropathy based on the orientation of water molecules at a given distance from the solute as computed from molecular dynamics simulations. In this way it is possible to study the conditional probability distribution for each amino acid in order to obtain information on its hydrophathy characteristics. To validate our hydropathicity analysis for the 20 natual of amino acid, we compared our theoretical scale with the most relevant scales available in literature by mean cluster and correlation analysis. We applied our scale to predict transmembrane regions of proteins and compared the results with the predictions based on the Translocon hydrophobicity scale, one of the most reiable experimental scale avaiable in literature, derived for transmembrane helical proteins. The accuracy of the prediction was quantified via two indicators, the two-class classification accuracy criterion (Q2) and the Segment OVerlap value (SOV).
**Results:** We have shown that very satisfying results can be obtained by simulating the actual environment of the water distribution around the amino acid, which has also the great advantage of permitting to selectively consider the contribution of its various moieties. We tested the accuracy of our scale by comparing its values with those provided by the many available scales and, interestingly, although our results do not depend on complex and time consuming experiments, they correlate very well with recently experimentally derived scales, thus opening the road to the accurate investigation of the effect of amino acid modifications as well as to the measure of the hydrophobicity of non natural amino acids, present for example in most antibiotics. The scale is also effective in the prediction of the location of transmembrane segments in proteins. This method can be extended to more complex systems and be used to accurately compute the hydrophobicity of regions of biomolecules, such as binding sites or interfaces.
**Contact email:** edoardo.milanetti@gmail.com

Poster: Other – OT2

**ELIXIR - European Life Science Infrastructure for Biological information: the Italian Node**

Graziano Pesole (1,2) and Anna Tramontano (3)

*(1) Institute of Biomembranes and Bioenergetics (IBBE), National Research Council, Bari (Italy) (2) Department of Biosciences, Biotechnology and Pharmacological Sciences, University of Bari "Aldo Moro", Bari (Italy) (3) Department of Department of Physics, Sapienza University of Rome, Rome (Italy)*

**Motivation:** ELIXIR (http://www.elixir-europe.org/) is a pan-European infrastructure for life science research support, in particular aimed for handling biological information. Due to new technologies in life sciences, the rate of production of biological data is intensely increasing. The collection, curation, storage, archiving, integration and deployment of large amounts of biomolecular data is a challenge that cannot be handled by a single organisation or by one country alone, but requires international coordination. ELIXIR is built on existing data resources and services. It follows a hub-and-nodes model, with a single Hub located in a new building alongside EMBL-EBI in Hinxton, Cambridge, UK and a growing number of Nodes located at centres of excellence throughout Europe. The goal of this distributed structure is to ensure that data are kept safe and made easily accessible. To that end, ELIXIR should be sustainably funded by European member nations as well as by regional and international agencies. At present, 17 member states and EMBL have signed a Memorandum of Understanding (MoU) to participate in ELIXIR. This is a non-binding agreement, which commits Member States to work towards signing the ELIXIR Consortium Agreement (ECA). On 18 December 2013, ELIXIR became a permanent legal entity following the ratification of the ELIXIR Consortium Agreement by the UK, Sweden, Switzerland, the Czech Republic, Estonia and EMBL. The remaining ELIXIR MoU countries will work towards ratifying the ECA throughout 2014. The Italian node is a virtual node configured as a Joint Research Unit (JRU) called Elixir-ITA that will coordinate the delivery of existing bioinformatics services at the national level and will provide many leading services to the ELIXIR infrastructure. Elixir-ITA is led by National Research Council (CNR) of Italy and brings together 12 partners including several universities as well as leading high-performance computing partners such as CINECA, CRS4, GARR and INFN. The ELIXIR Italy Node is going to establish a robust procedure, based on an open call and a peer review system, to allow additional participants to join with their relevant resources. This JRU will coordinate the delivery of existing bioinformatics services at the national level and will provide many leading services to the ELIXIR infrastructure. The ELIXIR Italy Node also has strong local connections with the Lifewatch Research Infrastructure, which is an e-infrastructure for ecosystems and biodiversity research.

**Contact email:** g.pesole@ibbe.cnr.it

**Supplementary information:**This abstract is not a research abstract but is just a communication notice we think useful to include in the Abstract Book.

Poster: Other – OT3

**Users' requests and desired features of the MIRRI-IS platform**

Romano P(1), Bunk B(2), Klindworth A(3), Robert V(4), Smith D(5), Vasilenko A(6), Glöckner FO(3)

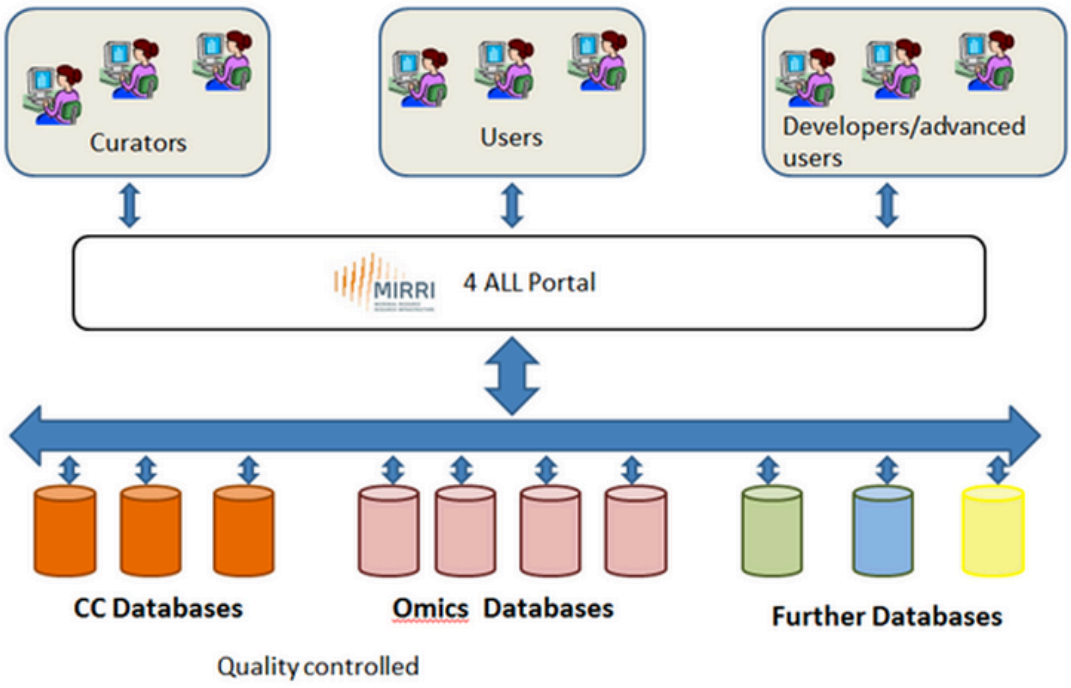
*(1) IRCCS AOU San Martino IST, Genova, Italy (2) Leibniz-Institut DSMZ, Braunschweig, Germany (3) Jacobs University Bremen, Bremen, Germany (4) CBS-KNAW Fungal Biodiversity Centre, Utrecht, The Netherlands (5) CAB International, Egham, United Kingdom (6) Russian Academy of Sciences, Pushchino, Russia*

**Motivation:** The Microbial Resource Research Infrastructure (MIRRI) started its preparatory phase in 2012 in the context of the European Strategic Forum for Research Infrastructure (ESFRI) programme aiming at providing microbial resources, associated data, taxonomic methods, and expertise to serve users' needs (1). One of its main design challenges relates to the most appropriate definition of the characteristics and features of the MIRRI Information System (MIRRI-IS) platform. The MIRRI "Data Resources Management" workpackage activity serves to improve the quantity, quality, interoperability, and usage of data associated with biological material. Here, we present the results of first surveys on users' requests and desired features for the MIRRI-IS.
**Methods:** A first collation of information needs was achieved through two surveys that were submitted to members of the European Culture Collection Organization (ECCO) and to their customers and users. Further discussions on users' requests and desired features for the MIRRI-IS platform was conducted in Athens, during a Satellite workshop of the XXXII Annual Meeting of the European Culture Collection Organization (ECCO), June 12-14, 2013, and in Beijing, at the Workshop on "Data management, bioinformatics, information system and networks of microbial resources", September 26, 2013, jointly organized by the World Data Center for Microorganisms (WDCM) and the CODATA Task Group on Advancing Informatics for Microbiology (TG-AIM).
**Results:** Users' requests mainly relate to improved and reliable collections' metadata and to improved interoperability both between collections databases and with some of the main molecular biology databases. These provide the essential basis that is needed to leverage the MIRRI platform for an effective integrated analysis aimed at allowing innovative user analysis and at developing "killer applications" in all fields that may profit from research on microbial resources, including health, food and environment. The search for type strains for some taxa and for rare and unique names of microorganisms (MOs), as well as tools for identification and correcting misspellings in the names are seen as basic needs, especially with reference to new taxa. The search for strains on the basis of specific information, such as ecological conditions at isolation, host, substrate, growth conditions, and, of course, biological activity is also requested. In addition proper bibliographical references for resulting MOs are needed. Combined lists of microbial cultures from several collections enabling selection of the most appropriate to their needs are essential. With reference to "external" databases, information on taxa diversity for microorganisms present in INSDC and search for bibliographical references on cultures of specific collection and comparison with information in other collections, are top priorities. Beyond the search for microorganisms based on their properties, it appears that "navigation" in a common information space for microbiology, bioinformatics, environment, agriculture, and medicine would especially be welcomed by users. On the basis of these premises, the criteria for the establishment of the MIRRI-IS should be: i) maintain, and possibly extend to more collections, the high data curation quality, ii) extend and improve interoperability, both within culture collections and

with external databases, and iii) build an open platform able to allow innovative downstream analysis. The next steps will include the definition of the MIRRI Minimum Data Set on the basis of the OECD and CABRI standards (2,3), to be developed incrementally, following the example of the MIxS standard and checklists (4), as Minimum Information about Biological Resources (MIaBRe), and of an appropriate interoperability language on the basis of the Microbiological Common Language (MCL) (5).

**Contact email:** paolo.romano@hsanmartino.it

**Supplementary information:**References 1. Schüngel M, Stackebrandt E, Bizet C, Smith D. MIRRI - The Microbial Resource Research Infrastructure: managing resources for the bio-economy. EMBnet.journal 2013, 19.1:5-8. 2. OECD Best Practice Guidelines for Biological Resource Centres, OECD Publishing, Paris. See: http://www.oecd.org/health/biotech/38777417.pdf (also available in French) 3. Romano P, Kracht M, Manniello MA, Stehehuis G, Fritze D. The role of informatics in the coordinated management of biological resources collections. Appl Bioinformatics 2005, 4(3):175-186. doi: 10.2165/00822942-200594030-00002 4. Yilmaz P, Kottmann R, Field D, Knight R, Cole JR et al. Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications. Nat. Biotechnol. 2011, 29:415-420. doi: 10.1038/nbt.1823 5. Verslyppe B, Kottmann R, De Smet W, De Baets B, De Vos P, Dawyndt P. Microbiological Common Language (MCL): a standard for electronic information exchange in the Microbial Commons. Research in Microbiology 2010, 161(6):439-445. doi: 10.1016/j.resmic.2010.02.005



Poster: Pharmacogenomics – PHG1

**TuNDRA: a Tripartite Network based Drug Repositioning Algorithm**

Alaimo S(1), Pulvirenti A(2)∗, Giugno R(2), Ferro A(2)

*(1) Department of Mathematics and Computer Science, University of Catania. (2) Department of Clinical and Molecular Biomedicine, University of Catania. ∗ Corresponding author*

**Motivation:** Developing new drugs is an extremely expensive, time-consuming, and risky process. Current knowledge of existing drugs is limited, and new discoveries concerning possible new applications or unknown side effects are very common. Consequently, the development of algorithmic techniques for the repositioning of existing drugs may be a way to make this process more rational, yielding a better understanding of the underlying mechanisms. In the favorable cases this may greatly reduce the failure risk together with the new drugs development cost and time.
**Methods:** 0
**Results:** Here we propose a method called TuNDRA for the repositioning of existing drugs based on an extension of the recommendation technique described in Alaimo et al. 2013 for drug-target interaction prediction. Following Lee et al. 2013, we can represent our knowledge through a tripartite network, whose nodes can be drugs, targets, or diseases. Interactions in the network associate each drug with a disease through its targets. Our algorithm, starting from such a network, computes a weight for each possible pair of diseases, measuring how reliable is to claim that a drug, which is related the first disease, can be also associated with the second one. This weight is then used to compute recommendations and assign new diseases to each drug. These recommendations are then filtered producing a smaller set from which novel biological insight can be discovered. This process is applied to each drug by computing a measure of correlation between the aforementioned diseases and the known ones. Next we select a subset of diseases minimizing the probability of obtaining by chance a better correlation than the predicted one. Finally we can use these filtered predictions to guide the experimental activity reducing time and cost of the development process.
**Contact email:** apulvirenti@dmi.unict.it
**Supplementary information:**Salvatore Alaimo, Alfredo Pulvirenti, Rosalba Giugno, and Alfredo Ferro. “Drug-target interaction prediction through domain-tuned network-based inference.” Bioinformatics 2013 29: 2004-2008. Lee, Hee, et al. "Rational drug repositioning guided by an integrated pharmacological network of protein, disease and drug." BMC systems biology 6.1 (2012): 80.

Poster: Pharmacogenomics – PHG2

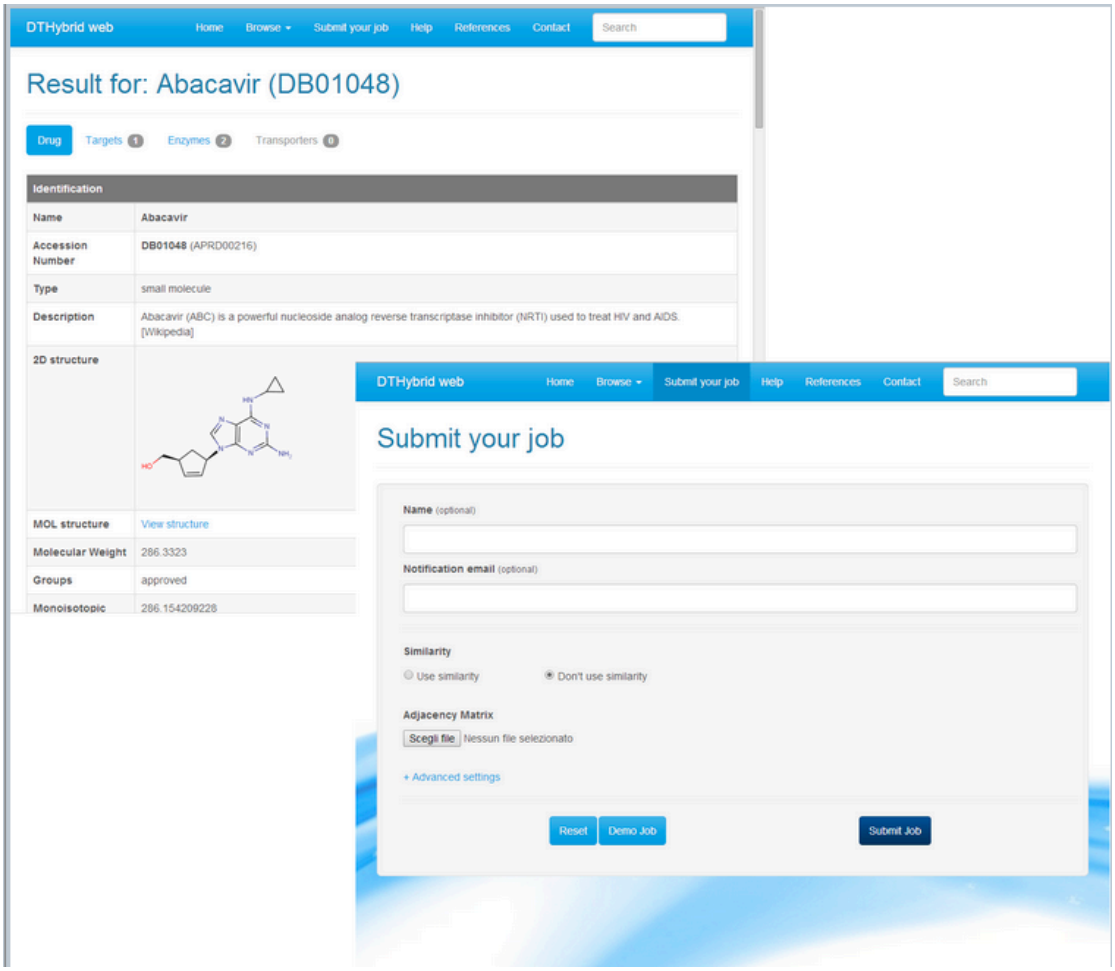
**DT-Hybrid Web: a web-based application for Drug-Target interaction prediction through domain-tuned network-based inference**

Alaimo S(1), Bonnici V(3), Cancemi D(1), Ferro A(2) , Giugno R(2)*+, Pulvirenti A(2)*+

*(1) Department of Mathematics and Computer Science, University of Catania. (2) Department of Clinical and Molecular Biomedicine, University of Catania. (3) Department of Computer Science, University of Verona. * Corresponding author + Equal Contributors*

**Motivation:** The identification of drug–target interaction (DTI) represents a costly and time-consuming step in drug discovery and design. Computational methods capable of predicting reliable DTI play an important role in the field. Algorithms may aim to design new therapies based on a combination of approved drugs. Recently, recommendation methods relying on network-based inference in connection with knowledge coming from the specific domain have been proposed. However no knowledge base offering the drug-target prediction service through web interfaces has been reported.
**Methods:** Here we propose a web-based interface to the DT-Hybrid algorithm (Alaimo et al., 2013), which applies a recommendation technique based on bipartite network projection implementing resources transfer within the network. Given a bipartite graph whose nodes denote drugs and targets, a two-phase resource transfer is associated with one of its projections. Initially the resource is transferred from target nodes to drug nodes, and subsequently the resource is transferred back to the target ones. This technique combined with domain-specific knowledge expressing drugs and targets similarity is used to compute recommendations for each drug. Our web interface allows the users: (i) to browse all the predictions made by the algorithm; (ii) to upload their custom data on which a prediction using DT-Hybrid will be computed; (iii) to predict combinations of drugs whose targets are at an optimal distance from the candidates disease genes. Our interface is periodically synchronized with DrugBank and updated accordingly. The website is free, open to all users, and available at the address http://alpha.dmi.unict.it/dtweb/.
**Results:** Our web interface allows users to search and visualize information on drugs and targets eventually providing their own data to compute a list of predictions. As shown in Figure 1, the user can visualize information about the characteristics of each drug, a list of predicted and validated targets, associated enzymes and transporters. A table containing key information and GO classification allows the users to perform their own analysis on our data. A special interface for data submission allows the execution of a version of the DT-Hybrid algorithm which predicts new targets for each drug providing a p-value expressing the reliability of each prediction. It is also possible to specify a list of genes tracking down all the drugs that may have an indirect influence on them based on their position in some pathway. Starting from the list of genes, we compute all possible pairs of targets (reported in drugbank or predicted) which are 3-4 steps away from the input genes. Such 3-4 distance value in the literature is claimed to reduce drug side effects. However this default parameter can be modified by the user. Finally, all drugs targeting such pairs are extracted and returned as combination prediction.
**Contact email:** apulvirenti@dmi.unict.it,giugno@dmi.unict.it
**Supplementary information:**Info: Alaimo S et al. “Drug-target interaction prediction through domain-tuned network-based inference.” Bioinformatics 2013 29: 2004-2008. Cheng F, et al. “Prediction of drug-target interactions and drug repositioning via network-based inference.” PLoS Comput. Biol. 2012;8:e1002503. Knox C et al. DrugBank 3.0: a comprehensive resource

for \'omics\' research on drugs. Nucleic Acids Res. 2011 Jan;39(Database issue):D1035-41. PMID: 21059682



Poster: Structure and Function – SF1

**Local backbone geometry plays a key role in determining the conformational preferences of amino acids**

Balasco N (1,2), Esposito L (1), De Simone A (3), Vitagliano L (1)

*(1) Institute of Biostructures and Bioimaging, CNR, via Mezzocannone 16, I- 80134 Napoli, Italy (2) Second University of Napoli, Via Vivaldi 43 - 81100 Caserta, Italy (3) Division of Molecular Biosciences, Imperial College South Kensington Campus London SW7 2AZ, UK*

**Motivation:** The classic experiments of Anfinsen have established that the three-dimensional structure of proteins is essentially dictated by their amino acid sequence. The so-called protein folding problem deals with the definition of the mechanism of how the amino acid sequence determines the protein structure. The solution of the protein folding problem would have a huge impact on our understanding of protein structure and function. Among the many facets of this problem, the definition of the structural basis of the conformational preferences of the genetically encoded amino acids represents one of the most elusive aspects. Indeed, although a large number of computational and experimental investigations have highlighted that the different protein residues are endowed with distinct conformational propensities, none of the current hypotheses is able to satisfactorily explain these preferences. Obviously, a complete understanding of amino acid propensities is crucial to decipher the folding code. In order to gain insights into this intricate issue, we here determined and compared amino acid propensity scales for different ($\phi$,$\psi$) regions of the Ramachandran plot and for different secondary structure elements. Furthermore, we also evaluated the role of local geometry in determining these conformational preferences.
**Methods:** The Ramachandran plot was initially divided in ($\phi$,$\psi$) boxes with dimensions of (30°,30°). Propensities of each residue for a given ($\phi$,$\psi$) or a given secondary structure element (E, G, H) were calculated by using the Chou-Fasman approach on a dataset of 4731 non-redundant (sequence identity ≤ 25%) protein chains. The chains were selected from protein structures solved at resolution better than 2.2 Å and refined to an R-factor lower than 0.20. The statistical analyses were carried out for residues whose average backbone B-factor (atomic displacement parameter) was lower than 1.3 times the average backbone B-factor of their own chain. Similarities between propensity scales of different ($\phi$,$\psi$) regions were evaluated by linear regression analyses in terms of the correlation coefficient *r*. The significance of the correlation coefficients between different scales was established with the so-called *null hypothesis*, i.e. the hypothesis that the variables are not correlated, by using Student's *t* distribution.
**Results:** One of the most striking and unexpected findings emerged from this study is that distant ($\phi$,$\psi$) regions of the Ramachandran plot occasionally exhibit significantly similar propensity scales. On the other hand, contiguous regions of the Ramachandran plot present anti-correlated propensities. These results are corroborated by the analysis of the propensity scales for different secondary structure elements. These comparisons unveiled some previously undetected (anti)correlations between different scales for repetitive structural motifs such as helices and beta-sheets. In order to provide an interpretative background to these results, we evaluated the role that the local variability of protein backbone geometry plays in this context. Our analysis indicates that (dis)similarities of propensity scales between different regions of the Ramachandran plot are coupled with (dis)similarities in the local geometry. In recent years, many investigations have highlighted the influence of the conformation on the protein local geometry. The present findings

reverse this concept by showing that the local geometric requirements have a significant impact on the preference of individual amino acids for specific conformational states.

**Contact email:** nicole.balasco@unina2.it

Poster: Structure and Function – SF2

**In silico structural genomics of Streptococcus mutans proteins**

Bizai M (1), Polticelli F (1,2)

*(1) Department of Sciences, University of Roma Tre, Rome, Italy (2) National Institute of Nuclear Physics, Roma Tre section, Rome, Italy*

**Motivation:** The structural genomics of pathogens is a field of investigation at the interface between biology and medicine that has received greater impulse in recent years due to advances in genomics and techniques for determining the three-dimensional structure of proteins. However, the experimental determination of protein structure is an extremely demanding task in terms of human and economic resources and therefore the number of known structures is still a fraction of the potential drug targets. Our attention was focused on Streptococcus mutans that is the leading cause of dental caries being the most cariogenic of all the oral streptococci. Moreover, a function has been assigned to only about 63% of its entire proteome, while the remaining proteins have no assigned function yet. With this in mind, the aim of this work was to predict the structure and the biological function of these latter proteins through a computational approach. The final aim of this work is to identify proteins which are potentially critical for the survival of the bacterium to be used as targets for the development of novel antibacterial drugs.
**Methods:** Ab initio molecular modelling strategies were exploited to study the structural features of Streptococcus mutans proteins whose biological function is still unknown. Structural models of these proteins were built using the pipeline implemented in the protein structure prediction server I-TASSER and the ab initio protein structure prediction program Rosetta ab initio. For each of the proteins analysed, a function was assigned on the basis of structural similarity with characterized proteins using DALI and sequence/structure information using ProFunc.
**Results:** Compact and well-ordered structural models were obtained for 71 proteins belonging to Streptococcus mutans proteome. Among these 33 display a significant structural similarity with proteins having a known structure and function. These proteins were sorted in different groups according to their predicted function. The “Major Facilitator Superfamily” of transporters is the group with the largest number of proteins (13 proteins), followed by “DNA-binding proteins” (5 proteins). Furthermore, three proteins are predicted to be ribosomal proteins, while two more proteins are predicted to be involved in lipid binding. Examples will be presented in which the approach used allows to assign a function with high reliability to the proteins analysed in the absence of any evolutionary information.
**Contact email:** fabio.polticelli@uniroma3.it

Poster: Structure and Function – SF3

**Ab initio protein structure and function prediction of Pseudomonas aeruginosa periplasmic proteome**

Caprari S(1), Barbabella G(1), Frangipani E(1), Imperi F(2), Casalino, M(1), Visca P(1), Polticelli F(1,3)

*(1)Department of Sciences, University Roma Tre, Roma (2)Department of Biology and Biotechnology, University La Sapienza, Roma (3)National Institute of Nuclear Physics, Roma Tre Section, Roma*

**Motivation:** Pseudomonas aeruginosa is an ubiquitous Gram-negative bacterium causing a wide range of chronic and acute infections. A huge versatility and adaptability to both the hospital environment and the human hosts, as well as a high intrinsic resistance to antibiotics and disinfectants, make Pseudomonas aeruginosa a pathogen with a relevant biomedical interest. Consequently, there is an urgent need for the development of novel and efficient drugs anti Pseudomonas aeruginosa. As a gram-negative bacterium, Pseudomonas aeruginosa shows a well-defined periplasmic space sandwiched between the outer and the inner membranes. Many processes needed for the bacterial survival and pathogenesis take place into the periplasmic compartment. Previous proteomic analyses aimed to characterize the entire repertoire of periplasmic proteins of Pseudomonas aeruginosa strain PAO1, revealed that more than 30% of these proteins have no assigned functions yet based on the sequence similarity with characterized proteins. These proteins are thus classified as "hypothetical". Therefore, the aim of the present work was to predict, by using a computational approach, the structure and biological function of these proteins in order to identify new molecular targets which might be used for the development of new anti-Pseudomonas drugs. Furthermore, experimental analyses on two hypothetical proteins were also performed in order to validate the conclusions drawn by the "in silico"approach.
**Methods:** Ab initio molecular modelling strategies were exploited to study the structural features of Pseudomonas aeruginosa periplasmic proteins whose biological function is still unknown or just putative. In particular, structural models of the hypothetical periplasmic proteins were built using the protein structure prediction program I-TASSER. In addition, the structural homology with proteins with known function was assessed using the DALI server.
**Results:** For a number of proteins, structural models were obtained which show a well-structured fold and a significant structural similarity with proteins having a known structure and function. In particular, all the selected periplasmic proteins (94 in total) were sorted in 5 main groups according to the results obtained from the structure and function prediction. The groups "periplasmic binding proteins" and "enzymes" are those in which the largest number of periplasmic proteins fell (34 and 17 proteins, respectively). Seven periplasmic proteins fell neither in the "periplasmic binding proteins" nor in the "enzymes" groups, since they are predicted to carry out different functions, including the uptake and transfer of molecules such as lipids or metals. Therefore, they were grouped in a generic "other functions" class. Furthermore, the group "particular cases" includes 4 proteins on which the computational analyses produced unusual results. For instance, some of the proteins belonging to this group display regions with structural similarities with other known domains but the remaining part of the structure predicted is disordered. Lastly, for 32 periplasmic proteins no significant results were obtained by using the "in silico" approach.
**Contact email:** fabio.polticelli@uniroma3.it

Poster: Structure and Function – SF4

**Towards the identification of the allosteric Phe-binding site in the phenylalanine hydroxylase**

Carluccio C(1), Fraternali F(2), Salvatore F(1,3), Fornili A(2), Zagari A(1)

*(1) CEINGE-Biotecnologie Avanzate, S.c. a r.l., Napoli, Italy (2) Randall Division of Cell and Molecular Biophysics, King's College London, London, U.K. (3) SDN-Istituto di Ricerca Diagnostica e Nucleare, Napoli, Italy*

**Motivation:** The hyperphenylalaninemias (HPAs) are autosomal recessively inherited disorders that can lead to severe mental retardation in humans, if left untreated. In fact, due to the relevance of these disorders, many countries include a test for HPAs in neonatal screening programmes. In this classical inborn error of metabolism, the gene primarily affected is the phenylalanine hydroxylase (PAH) gene, which results in a protein with reduced enzyme activity [1]. PAH catalyses the conversion of the essential amino acid phenylalanine into tyrosine. PAH is a tetramer and each monomer consists of an N-terminal regulatory domain (RD), a catalytic domain (CD) and a tetramerization domain. PAH is highly regulated in order to maintain the correct levels of Phe in the blood [2]. The major regulatory mechanisms of PAH include activation by the substrate, Phe [2]. It acts as an allosteric activator and modulates the activity of the enzyme by positive cooperativity with a mechanism still unclear. There is a debate regarding the ability of PAH to bind Phe at an allosteric site in the RD, besides to the active site. In a previous Molecular Dynamics (MD) study on the isolated monomeric RD, we put forward the hypothesis that Phe could bind to a putative hydrophobic region at the dimeric interface between the RD and the CD of two adjacent subunits [3]. In this context, we aimed to shed light on the open questions on the regulation mechanisms through the in silico identification of the allosteric Phe-binding site in the wild-type enzyme. Furthermore, taking into account that the mutation p.G46S results in mild/severe HPA [4], we selected to study the human G46S mutated protein. Indeed the position 46 lies within the putative allosteric Phe-binding site and the mutant G46S exhibits a reduced activation by Phe [4]. Therefore we aimed to elucidate the structural effects of this natural disease-causing mutation on the allosteric Phe-binding.
**Methods:** We performed molecular docking studies combined with MD simulations on three dimeric PAH forms (human, rat and human G46S mutant forms). We generated in silico models of the dimer of the human wt enzyme and G46S mutant, with the program Modeller9v8, using as templates the structures of the dimeric rat PAH (PDB code 1PHZ, sequence 33-116) and of the dimeric human PAH (PDB code 1PAH, sequence 117-424). The generated models of the human enzymes and the rat PAH structure (PDB code 1PHZ, sequence 33-424) were subjected to MD simulations in solution for 30ns, using the GROMACS package. Subsequently, Phe docking was performed on the minimum energy structures extracted from the MD simulations of the unbound forms, using the program MOE (Molecular Operating Environment). The best energy pose of each PAH form was refined by performing 10ns-long MD simulations. The simulation trajectories were analyzed with GROMACS analysis tools.
**Results:** The analysis of the trajectories provided dynamical and structural insight on the unbound dimeric forms, on the complexes with Phe and on the effects of the G46S mutation. In particular, during MD simulations the trajectories of the selected Phe-bound wild-type forms reached a plateau within a root mean square deviation of 1.5 Å from the initial structures calculated over all Cα atoms, indicating their stability and reliability. Although the Phe position slightly differs in the various analyzed systems, the binding site

over the simulations remained localized at the dimeric RD/CD interface, confirming our starting hypothesis on the Phe-binding region at the dimer interface. Instead, for the G46S mutant, during the MD simulations, the dissociation of Phe from the complex occurred. This result suggests that the mutation can affect the correct binding of the allosteric Phe. Defects in the regulation of the G46S mutant can contribute to explain the disease-causing nature of the mutant. In conclusion, this work paves the way for more detailed computational investigations and sets the basis for further experimental works.

**Contact email:** carla.carluccio@unina.it

**Supplementary information:**Acknowledgments The work was performed under the HPC-EUROPA2 project (project number: 228398) with the support of the European Commission - Capacities Area - Research Infrastructures. References 1. Cerreto M, Cavaliere P, Carluccio C, Amato F, Zagari A, Daniele A and Salvatore F (2011) Biochem et Biophys Acta 1812: 1435-1445. 2. Hufton S, Jennings I, Cotton R (1995) Biochem J 311: 353-366. 3. Carluccio C, Fraternali F, Salvatore F, Fornili A, Zagari A (2013) PLoS ONE 8 (11): e79482. 4. Pey A, Stricher F, Serrano L, Martinez A (2007) Am J Hum Genet 81: 1006-24.

Poster: Structure and Function – SF5

**Protein threading modeling approaches for describing the structural conformation of the Calcium-sensing receptor cytoplasmic C-terminal portion.**

Chiappori F(1), Milanesi L(1), Merelli I(1)

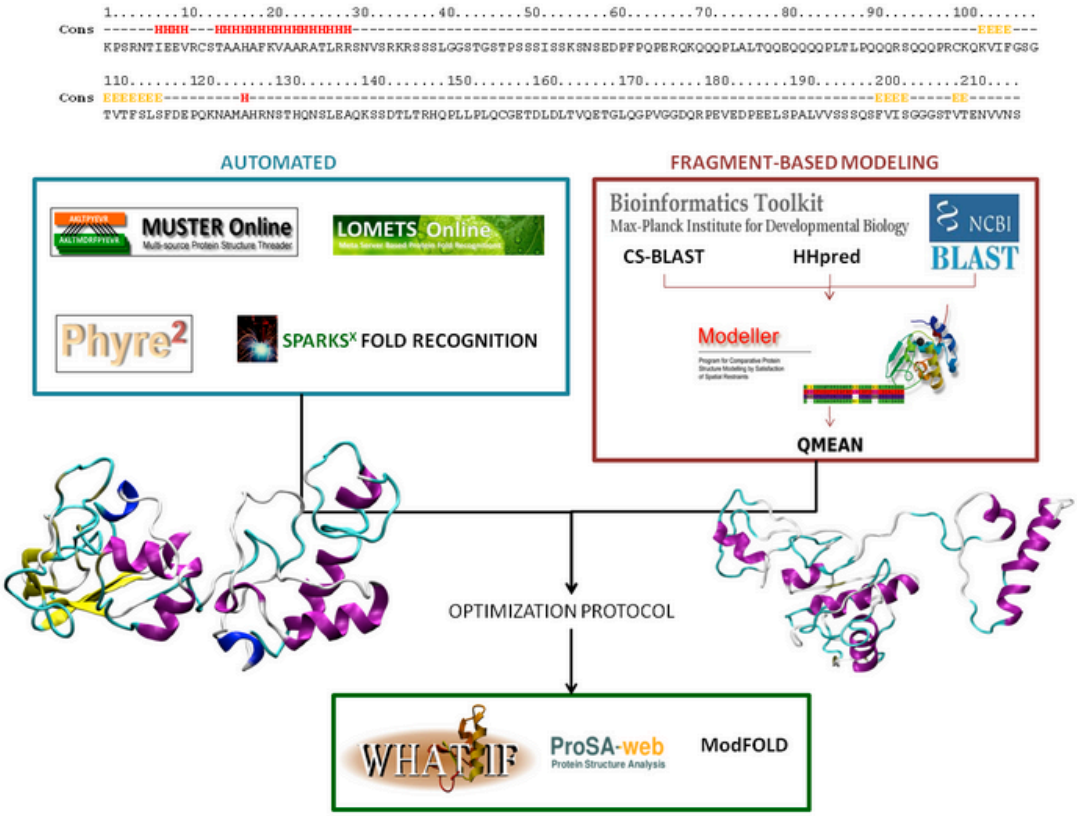
*(1) Istituto di Tecnologie Biomediche - CNR, Milano*

**Motivation:** Calcium-sensing receptor (CaSR) detects changes in extracellular calcium concentrations, regulates parathyroid hormone (PTH) secretion and renal tubular calcium re-absorption to maintain serum calcium levels. It is a 1078-amino acid glycoprotein with a predicted topology of a large extracellular domain, a seven-transmembrane-spanning region, and an intracellular tail. Several non-conservative SNPs were found in the CASR gene, related to familiar benign hypocalciuric hypercalcemia (FBHH), and neonatal severe primary hyperparathyroidism (NSHPT), or autosomal dominant hypocalcemia and type 5 Bartter syndrome. These polymorphisms in the cytoplasmic portion constitutively activate or inactivate the receptor inducing the pathological states. Therefore it is interesting to evaluate the effects of these substitutions on protein structure. No structure of CASR intracellular protein portion is available in the PDB and no homologs structures are available. For this reason, threading methodology was employed to obtain the 3D structure of CASR cytoplasmic protein tail.
**Methods:** 0
**Results:** The intracellular portion of CASR consists of 216 residues. First of all, secondary structures were predicted with different web servers partially included in Gensilico (1): HMMSTR, sspro4, sspal, proteus, SPARROW, sable, prof, nnssp, netsurfp, ssp, pssfinder, raptorxss, psspred, spineX and soprano, and in PSIPRED (2): Jnet, jhmm, jpssm,. A consensus was retrieved and employed to filter structure predictions (top of Figure1). Two different strategies were followed to build the 3D structure, an “automated approach” and a “fragment-based modeling” (Figure1). In the “automated approach”, several threading/fold recognition servers (MUSTER (3), LOMETS (4), Phyre2 (5) and SPARKS-X (6))were employed to achieve a model of the whole structure. In all obtained models, secondary structures were visually inspected and only models displaying secondary structures in the corrected positions were selected. Finally, only one model of 216 residues was retrieved, TS5 build on 3DP1 obtained from SPARKS-X. The “fragment-based modeling” approach consists in a manually search for homology, using local alignments tools such as PSI-BLAST (7), CS-BLAST (8), and HHPRED (9). Only alignments longer than 30 residues were considered. Then we tried to superpose these alignments along with the protein sequence. Moreover, the presence of previously predicted secondary structures in the aligned portion of the X-ray structures was verified. Finally, 9 structures were aligned as templates. Modeller 9.11 (10) was used to build the model with model_mult.py script; five models were obtained and were evaluated with evaluate_model.py script of Modeller and with QMEAN server (11) of Swiss-Model suite. The fourth model of 207 residues was choose. Both models undergone to an optimization protocol including two minimization steps: 1000 steps of steepest descendent followed by 1000 steps of Broyden-Fletcher-Goldfarb-Shanno approach. Then, a 30ps NVT simulation of equilibration was performed, where the system was gradually heated from 0 to 300K with 2.5ps of annealing timestep, followed by another equilibration of 100ps using an NPT simulation. Finally, 30 ns of MD simulation to relax the whole systems was performed. The representative conformations (Figure1) obtained from both the starting models were evaluated to choose the best model using What-if “structure validation”(12),

ProSA (13) and ModFold4.0 (14). The best conformation will be further employed to evaluate the effect of the SNPs on protein structure.

**Contact email:** federica.chiappori@itb.cnr.it

**Supplementary information:**(1) https://www.genesilico.pl/meta2/ (2) http://bioinf.cs.ucl.ac.uk/psipred/ (3) http://zhanglab.ccmb.med.umich.edu/MUSTER/ (4) http://zhanglab.ccmb.med.umich.edu/LOMETS/ (5) http://www.sbg.bio.ic.ac.uk/phyre2/html/page.cgi?id=index (6) http://sparks-lab.org/yueyang/server/SPARKS-X/ (7) http://blast.ncbi.nlm.nih.gov/Blast.cgi (8) http://toolkit.tuebingen.mpg.de/cs_blast (9) http://toolkit.tuebingen.mpg.de/hhpred (10) http://salilab.org/modeller/ (11) http://swissmodel.expasy.org/qmean/cgi/index.cgi (12) http://swift.cmbi.ru.nl/servers/html/index.html (13) https://prosa.services.came.sbg.ac.at/prosa.php (14) http://www.reading.ac.uk/bioinf/ModFOLD/ModFOLD_form_4_0.html



Poster: Structure and Function – SF6

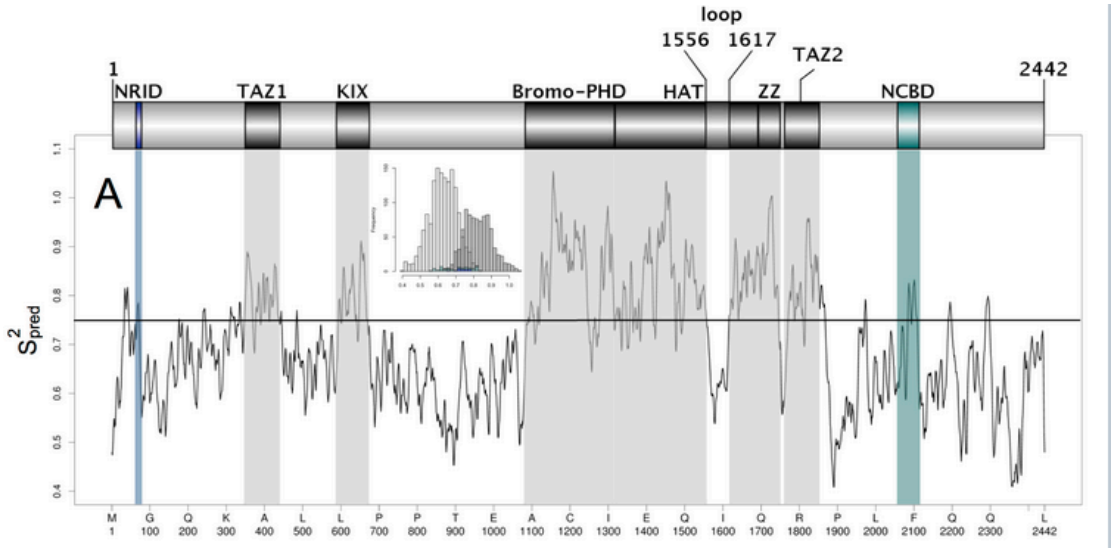
**DynaMine: from protein sequence to dynamics and disorder**

Cilia E(1,2), Pancsa R(3,4), Tompa P(2,3,4), Lenaerts T(1,2,5) and Vranken W F(2,3,4)

*(1) Computer Science Department, Université Libre de Bruxelles (ULB), Brussels, Belgium (2) Interuniversity Institute of Bioinformatics Brussels (IB2), ULB-VUB, Brussels, Belgium (3) Structural Biology Brussels, Vrije Universiteit Brussel (VUB), Brussels, Belgium (4) Department of Structural Biology, VIB, Brussels, Belgium (5) Computer Science Department, Vrije Universiteit Brussel, Brussels, Belgium*

**Motivation:** Dynamics are an essential of protein function. Cases in point are intrinsically disordered proteins, which can have key biological functions despite adopting an ensemble of conformations without a consistent three-dimensional structure. Understanding dynamics and disorder poses significant challenges, mainly because accurate protein dynamics information remains difficult to obtain.
**Methods:** We have recently shown how statistical analysis of NMR data of proteins in solution can give quantitative insight into the relationship between amino acid sequence and backbone dynamics and developed, DynaMine, a fast and accurate predictor of protein of residue-level backbone dynamics starting from protein sequence information alone [1]. The predictor consists of a linear regression model trained on backbone N-H S2 order parameter values for 210880 residues in 1952 proteins; these S2 values were estimated with the Random Coil Index [2] from a carefully assembled dataset of NMR chemical shift data extracted from the BioMagResBank (BMRB) [3]. S2 order parameters represent how restricted the movement of an atomic bond vector is with respect to the molecular reference frame, with physical values varying between 1 for fully restricted (rigid conformation) and 0 for fully random movement (highly dynamic). DynaMine takes as input a protein sequence and produces a profile of per-residue predicted S2 values (S2pred) as in Fig. 1. Each S2pred is predicted based on the local sequence environment provided by the 25 residues preceding and following the target residue in the sequence.
**Results:** The predictor has been first validated through both 10-fold cross-validation experiments and validation on unseen data (on multiple independent sets). Then we have shown on a set of well-studied proteins covering a broad range of structural and functional properties that DynaMine has great potential in distinguishing regions of different structural organization, such as folded domains, disordered linkers, molten globules and pre-structured binding motifs (e.g. Fig. 1). Additionally, we challenged the relevance of the predictions in relation to protein disorder [4]: DynaMine identifies disordered protein regions with an accuracy comparable to the most sophisticated existing predictors, without relying on any prior disorder annotation, therefore providing an independent link between dynamics and structural disorder. DynaMine is very fast, produces excellent results despite the simple linear prediction methodology applied, and gives a continuous and subtle picture of how amino acid residues behave with respect to their backbone rigidity and, by extension, to residue order and disorder. We therefore developed a web-server incorporating this novel predictor, which is available at http://dynamine.ibsquare.be. Through the DynaMine webserver we want to provide molecular biologists with an efficient and easy to use tool for estimating the dynamical characteristics of any protein of interest even in the absence of experimental observations. 1. Cilia, E., Pancsa, R., Tompa, P., Lenaerts, T. & Vranken, W. F. From protein sequence to dynamics and disorder with DynaMine. Nat Commun 4, 2741 (2013). 2. Berjanskii, M. V. & Wishart, D. S. The RCI server: rapid and accurate calculation of protein flexibility using chemical shifts. Nucleic Acids Res

35, W531-537 (2007). 3. Ulrich, E. et al. BioMagResBank. Nucleic Acids Res 36, D402-408 (2008). 4. Tompa, P. Trends in biochemical sciences 27, 527-533 (2002).
**Contact email:** ecilia@ulb.ac.be
**Supplementary information:**This work has been published in: Cilia, E., Pancsa, R., Tompa, P., Lenaerts, T. & Vranken, W. F. From protein sequence to dynamics and disorder with DynaMine. Nat Commun 4, 2741 (2013).



Poster: Structure and Function – SF7

**Structure-function relationships of the human selenoprotein M (SelM), found over-expressed in hepatocellular carcinoma, and of its six mutants.**

Guariniello S(1), Colonna G(1), Raucci R(1), Costantini M(2), Castello G(3), Costantini S(3)

*(1) Biochemistry, Biophysics and General Pathology Department, Second University of Naples, Naples, Italy (2) Stazione Zoologica "A. Dohrn", Naples, Italy (3) Istituto Nazionale Per Lo Studio E La Cura Dei Tumori "Fondazione Giovanni Pascale", IRCCS, Italy*

**Motivation:** Selenoprotein M (SelM) is a protein with central α/β domain and 145 amino acids long, of which the first 23 represent a signal peptide. Structure-based multiple sequence alignments suggested that SelM is member of the thioredoxin family composed of a two-layer α/β sandwich with a mixed four-stranded β-sheet and a pair of α-helices that are packed against one side of the β-sheet. Moreover, it has a short N-terminal extension that precedes strand β1 and a flexible C-terminal extension after helix α3. Recent studies conducted in our laboratory showed that SelM is over-expressed in two hepatoma cell lines (HepG2 and Huh7) compared to normal hepatocytes [1-2]. Therefore, aim of this study has been to analyze the structure–function relationships of SelM. Hence, firstly we studied the evolutionary history of this protein by phylogenetic analysis and GC content of genes from various species. Then, we modeled the three-dimensional structure of the human SelM evaluating its energetic stability by molecular dynamics simulations. Finally, we modeled six mutants to obtain structural information helpful for structure-based drug design.
**Methods:** A phylogenetic study was done by Clustalw program to evaluate the SelM presence in different organisms and the evolutionary pressure measuring the GC content in all SelM genes. The structures of SelM and of its mutants were modeled by MODELLER9v11 program using the murine protein as template. The best models were selected using Prosa program to check its energetic quality , and Ramachandran Plot to evaluate its stereochemical quality. The models were subjected to 20 ns of molecular dynamics simulations using the GROMACS program. The obtained trajectories were analyzed in terms of RMSD, RMSF, radius of gyration, H-Bonds and principal component analysis (PCA) to evalutate the SelM fluctuations during the dynamics and to understand the explored conformational spaces. The Cytoscape program was used to analyze the interactions between residues using two NetworkAnalyzer and RINalyzer plugin.
**Results:** We found 48 SelM sequences among which 21 mammalians, 4 birds, one reptile, 2 amphibious, 3 fishes, one nematode, 5 brown and green seaweeds, 10 arthropodes and one sponge. The related phylogenetic tree showed well clustered organisms according to evolutionary scale, where the chordata are clustered and separated by arthropoda and by chromalveolata (brown algae)–and clorophyta (green algae). In overall our studies evidenced that: i) SelM had accumulated mutation specific to each phyla, ii) an increase of GC content appeared along the phylogenetic tree according the GC heterogeneity of their genomes, iii) the SelM 3D-model showed an α/β structure with some loop regions and a disordered C-terminal region according to the disorder propensity prediction and the local flexibility evaluation, iv) the structural organization of SelM was stable during the molecular dynamics simulation except for the C-terminal portion as confirmed by the PCA analysis and covariance matrix, v) the SelM 3D-model was stabilized after MD by H-bonds, π-cation interactions, salt bridges and π-stacking interactions and presented three well conserved residues F59, L82 and L84, and vi) six mutants of human SelM replacing the F59, L82 and L84, with A or G residues, explored a larger conformational space in respect to the human model, showing a significant decrease in their surrounding interactions with the loss of their

structural characteristics as hub residues, and suggesting an higher structural flexibility and the critical role played from these three residues in the structural stability of human SelM.
**Contact email:** susan.costantini@unina2.it
**Supplementary information:**References [1] S. Costantini, G. Di Bernardo, M. Cammarota, G. Castello and G. Colonna "Gene Expression segnature of human hepatoma cells" Gene (2013), 518: 335–345 [2] S. Guariniello, G. Colonna, R. Raucci, M. Costantini, G. Di Bernardo, F. Bergantino, G. Castello and S. Costantini: "Structure-function relationships and evolutionary history of the human selenoprotein M (SelM) found over-expressed in hepatocellular carcinoma" BBA (2014) in press

Poster: Structure and Function – SF8

**Conformational study of membrane CXCR3 chemokine receptor by molecular dynamics simulations in a lipid bilayer and network analysis**

Raucci R(1), Costantini S(2), Castello G(2), Colonna G(1)

*(1) Department of Biochemistry, Biophysics and General Pathology, Second University of Naples, Naples, Italy (2) Istituto Nazionale Per Lo Studio E La Cura Dei Tumori "Fondazione Giovanni Pascale", IRCCS, Italy*

**Motivation:** The family of chemokines is known to play an important role in chronic inflammatory processes leading to cancer and in metastasis development. Their receptors are membrane proteinsresponsible for transmitting information as GPCR receptors, and mediating chemotaxis and carcinogenesis. They contain seven helices (7TM) that span the membrane bilayer and are connected together by three extracellular loops and three intracellular cytoplasmic loops. Moreover, they are characterized by a long N-terminal extracellular segment responsible for the binding with chemokines and a long C-terminal segment specific for signaling.. Previous studies in our laboratory have shown that the two terminal regions have physico-chemical properties typical of intrinsically disordered domains and have high flexibility. Another peculiar feature of this family of cytokines is their pleiotropy. In recent years we have focused attention on the CXCR3 receptor, which is known to bind three chemokines. CXCL9, CXCL10 and CXCL11, found to be overexpressed in sera of patients with hepatocellular carcinoma [1-4]. Therefore we have created the model of the receptor CXCR3 by comparative modeling and subjected it to molecular dynamics studies in a system composed by a lipid bilayer and water molecules in order to evaluate its physico-chemical and structural features, and to study conformational changes occurred after the simulations.
**Methods:** We modeled the receptor using as reference the crystallographic structure of the CXCR4 receptor for the trans-membrane part, and the bovine rhodopsin for the N-terminal region because the experimental structure of CXCR4 was missing of the N-terminal region. The obtained model was evaluated in terms of stereochemical and energetic quality by Ramachandarn Plot and Prosa software. The molecular dynamics were performed using NAMD software for 100 nanoseconds in a system composed by a lipid bilayer and water molecules, and analyzed with Gromacs routines to evaluate RMSD, RMSF, radius of gyration and secondary structure and for PCA (Principal Component Analysis). The analysis of Residue Interaction Network (RIN) was performed by RIN plugin in Cytoscape software.
**Results:** CXCR3 reaches to equilibrium after a few nanoseconds in terms of RMSD. In particular, the trans-membrane helices result enough conserved during the simulation. As visible from analysis of RMSF and radius of gyration and by secondary structure evolution, the N-terminal region is the most flexible and tends to explore many conformations in the extracellular space, even if keeps always a compactness due to a network of H-bonds that is formed during the simulation. Usually, the binding of a ligand from outside the cell induces a conformational change in the 7TM receptor that can be detected inside the cell. In particular, our studies regarding the evolution of conformational changes, made by the binding of chemokine, indicate that helix-helix interaction contributes to the functional tertiary structure of CXCR3 necessary for receptor folding and stability, ligand binding and ligand-induced conformational changes for G protein coupling. Further studies will regard the modeling of the complex between CXCR3 and its natural ligand, CXCL11, by molecular docking strategies and molecular dynamic simulations to study conformational changes occurred after binding between the receptor and CXCL11.

**Contact email:** raf.raucci@gmail.com
**Supplementary information:**References: [1] S. Costantini, F. Capone, E. Guerriero, P. Maio, G. Colonna and G. Castello: “Serum cytokine levels as putative prognostic markers in the progression of chronic HCV hepatitis leading to cirrhosis” Eur. Cyt. Netw. (2010) 21 (4), 251-6 [2] T. Trotta, S Costantini, G. Colonna: “Modelling of the membrane receptor CXCR3 and its complexes with CXCL9,CXCL10 and CXCL11 chemokines: Putative target for new drug design” Molecular Immunology (2009) ,47, 332-339 [3] P. Palladino, L. Portella, G. Colonna, R. Raucci, G. Saviano, F. Rossi, M. Napolitano, S. Scala, G. Castello, S. Costantini: The N-terminal peptide of CXCL11 as structural template for CXCR3 antagonist: synthesis, conformational analysis and binding studies. Chemical Biology & Drug Design (2012), 80(2): 254-265 [4] S. Costantini, R. Raucci, T. De Vero, G. Castello, G. Colonna: “Common structural interactions between the receptors CXCR3, CXCR4 and CXCR7 complexed with their natural ligands, CXCL11 and CXCL12, by a modeling approach” Cytokine (2013) 64(1):316-21

Poster: Structure and Function – SF9

**In Silico Food Science: Lesson Learning from Medicinal Chemistry**

Cozzini P. Molecular Modelling Lab, Department of Food Chemistry, University of Parma

*Molecular Modelling Lab, Department of Food Chemistry, University of Parma*

**Motivation:** In silico techniques are a well known and applied instruments in Medicinal Chemistry to discover new lead compounds and to decipher mechanisms of binding. Molecular Modelling allows the prediction of protein–ligand interactions thus it can be used to predict the interactions of millions of compounds from data bases with a selected receptor as target for a specific disease. The aim is to find the best lead compounds to be used as a base to develop a right drug, with a reduction of false positives to reach a "pharmaceutical lead" in a discrete chemical space with a good chemical selectivity. The question we ask is: is it possible to apply these well known computational approach to the same problem we face up in food science? EFSA, the European Food Safety Agency, suggested to develop methods to minimize the number of in vivo and in vitro tests to reduce money and time and for ethical issues arising from experiments with animals. The unique answer we have is: yes, we can! It is always chemistry! The difference is on the aims. In food science the main aims are in order to discover new possible toxins, molecules that can act as new endocrine disruptors, to decipher mechanism of binding of know toxins and/or flavonoids and his metabolites. Thus we can talk about in silico toxicology, in silico risk assessment, in silico food safety, etc., but in this case we have to reduce false negatives to reach a "toxicological lead". Moreover chemical selectivity is not needed because we are in a heterogenous chemical space but we need food molecules data bases, e.g. a food additives data base. And last but not least, in many cases, targets are the same as for medicinal chemistry, gaining from the knowledge stored by more than 30 years of computational methods.
**Methods:** Here we present a molecular modeling method (showed in graphical abstract) we developed for medicinal chemistry and applied to food science problems to discover new xenoestrogens among food additives, new chemicals acting as possible xenoestrogens, or to understand the mechanism of binding for mycotoxins to nuclear receptors. This in silico method is based on a virtual screening procedure (using a new software, FLAP, Fingerprint for Ligands and Proteins) followed by a docking simulation of the selected compounds (using GOLD docking software) and a scoring/rescoring procedure to evaluate the energy of the interactions based on an empirical scoring function implemented into HINT, a software package. Moreover a 3D food additives database has been implemented to be used for virtual screening.
**Results:** Several applications of this metod will be illustrated for Zearalenone and derivatives towards Estrogen Receptors, ITX (inks) versus Androgen receptors, Urolithins and derivatives binding ER and Alternariol binding Topoisomerase I.
**Contact email:** pietro.cozzini@unipr.it

Poster: Structure and Function – SF10

**Modelling the conformational transitions in CDK2**

D'Abramo M(1), Besker N(2), Amadei A(2)
*1) SuperComputing Applications and Innovation Department, CINECA Roma, via dei Tizii, Rome 6 00185, Italy 2) Dipartimento di Scienze e Tecnologie Chimiche, Universita' di Roma "Tor Vergata", via della Ricerca Scientifica, Rome 1 00185, Italy*

**Motivation:** In the last few decades, the historical vision of proteins - described as a single and well defined structure - has been strongly questioned. It is now recognized that proteins work because they are dynamical entities able to explore several different conformations. The possibility to access different structures makes it possible for proteins to perform very different functions, such as enzymatic catalysis, allosteric transitions or ligand transport and binding. Structural studies have shown that transitions between different conformations of macromolecules and their complexes are important in biological processes. From the experimental viewpoint, the structures of populated states can be characterized in great detail using X-ray diffraction and NMR spectroscopy. The rates of population exchange can also be probed by a variety of spectroscopic techniques, for example NMR spectroscopy or time-resolved optical spectroscopy. Nonetheless, the characterization of the transitions between stable states deserves particular attention. In fact, a description of the transitions should aid in understanding the mechanism and possibly allows to select and interfere in a well defined "step" of the conformational transitions (for example by designing agonist ligands for receptors). However, because the system is predominantly found in one of the stable states and not in the transition region, direct experimental investigations of the transitions between them are difficult. In this respect, molecular simulations can potentially help to characterize in some details such kind of processes: the fast grow of the computational power and the development of more and more accurate force-fields allowed molecular dynamics simulations to contribute to the understanding of several of biophysical and biochemical processes at an atomic-molecular level of details. However, the large majority of such processes occurs on a time scale (milli-second to second) inaccessible to standard all-atom molecular dynamics simulations with transferable force fields. Then, it is not surprisingly that the quest for theoretical methodologies able to overcome such limits represents a very intensive area of research.
**Methods:** We present here the application of an enhanced sampling technique – the essential dynamics sampling (EDS) to the description of the Cyclin-dependent Kinase 2 (CDK2) open-closed and closed-open transitions. Such a technique is based on the possibility to guide the sampling towards a target and to reject movements (in the conformational space) which are not in the desired direction. With respect to other enhancing sampling methodologies, the advantages of the EDS technique are that the system moves along its own essential eigenvectors as provided by free (unbiased) molecular dynamics (MD) and no ad hoc terms are introduced in the Hamiltonian. Moreover, the contraction procedure guarantees – by construction – that the system goes from an initial state to a final state, being only slightly constrained.
**Results:** The method applied to the CDK2 show that such a procedure is robust, computationally efficient and does not require any a priori knowledge of the regions involved in the transition but requires only a description of the starting and final conformational states. At structural level, we found that both the opening and closure follow common transition paths in the essential subspace, which can be used to highlight the structural determinants involved in such processes.
**Contact email:** mdabramo.res@gmail.com

Poster: Structure and Function – SF11

**Structural properties of protein families as markers of functional features**

Del Prete Eugenio (1), Dotolo Serena (1), Marabotti Anna(2), Facchiano Angelo (1)

*(1)CNR - Istituto di Scienze dell'Alimentazione, Avellino (AV) (2)Università degli Studi di Salerno - Dip. Chimica e Biologia, Fisciano (SA) (*) Del Prete E. and Dotolo S. contributed equally to this work.*

**Motivation:** The conservation of the protein architecture in homologous proteins has been evidenced during years of investigation about structure. We are interested in investigating how the subtle structural differences existing among homologous proteins may be responsible of the functional properties modulation such as activity and thermostability. In addition, conformational changes are also observed for the same molecule, under different conditions as environmental parameters or presence of ligands. Therefore, we are exploring novel methods to investigate protein structure, and to analyze conformational features and differences, in order to find their relationships to functional properties.
**Methods:** We selected a few protein families based on our previous experience on their structural and functional properties. Structural parameters investigated for these proteins are: secondary structure, hydrogen bonds, accessible surface area, volume, phi - psi - omega angles, packing defects, charged residues, salt bridges, and others. Statistics and graphics have been performed with R packages in RStudio IDE. Some valuable results from Pearson correlation matrix have been validated with a Student's t-distribution test at a significance level of 5% (p-value).
**Results:** Best relationships among parameters have been investigated in detail. The correctness of this approach was borne out by relationships in agreement with geometric properties or expected by protein known structural properties. In addition, we found unknown relationships, which will be object of further studies, in order to consider them as putative markers related to the peculiar structure-function relationships for each family.
**Contact email:** eugenio.delprete@isa.cnr.it
**Supplementary information:**Acknowledgements This work is partially supported by the Flagship InterOmics Project (PB.P05, funded and supported by the Italian Ministry of Education, University and Research and Italian National Research Council organizations).

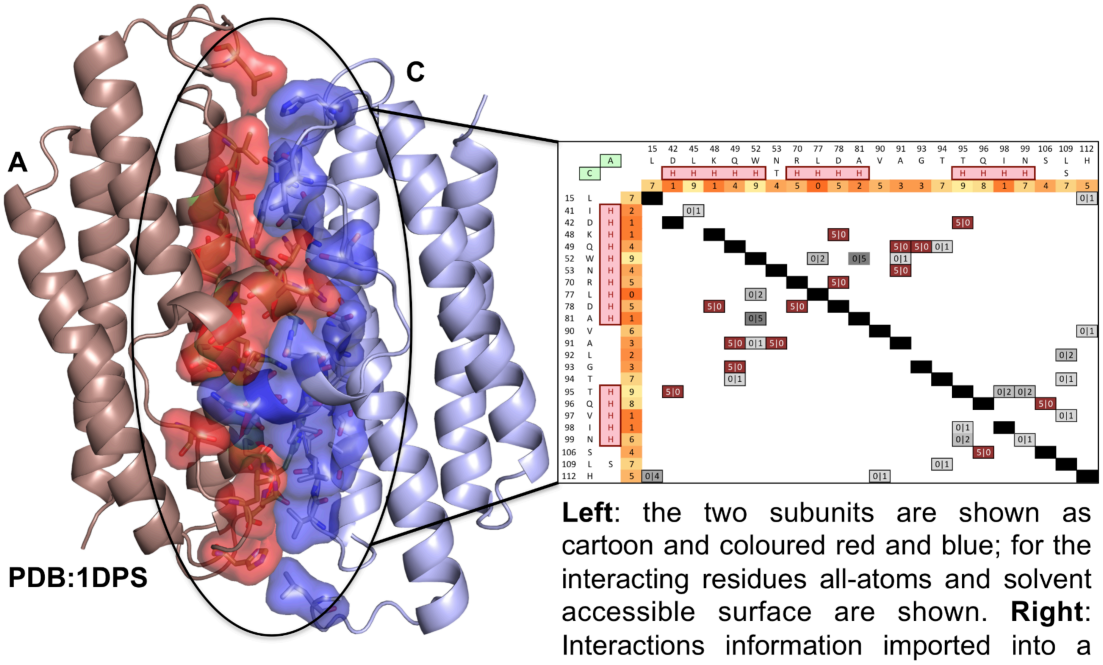
Poster: Structure and Function – SF12

**A simple tool for macromolecule interfaces analysis**

Di Micco P(1), Morea V(2)

*(1) Department of Biochemical Sciences "A. Rossi Fanelli", University of Rome "Sapienza", Rome (2) Institute of Molecular Biology and Pathology, CNR - National Research Council of Italy, Rome*

**Motivation:** Several kinds of interactions have been observed to be involved in the stabilization of tertiary and quaternary structures and in protein function. The role and importance of the most common inter-atomic interactions, namely those involving apolar residues or hydrogen bonds, are very well known. The analysis of macromolecule interfaces is a necessary requirement to understand the principles of molecular recognition; provide a valid aid to distinguish between transient and permanent complexes; contribute to our understanding of sequence-structure relationships, structural basis of protein stability and protein evolution; and for possible biomedical and biotechnological applications. Different valuable tools to analyse interfaces between biological macromolecules have been developed and are available from the WWW. Most of these tools are specialized in the analysis of protein-protein interfaces, whereas the analysis of nucleic acid interfaces is less common. However, to gain many of the information commonly required for interface analysis it is generally necessary to query different servers, and the data they provide is not always easy to obtain or handle. As an example, contacts are generally described at a residue rather than atom level, and information concerning non-polar contacts is either completely absent or provided as images (e.g., 2D contact maps), or other formats that cannot be parsed. Our aim was to create a fast and interactive program to analyse interfaces involving protein and/or nucleic acid molecules whose output could be easily handled and parsed by the user.
**Results:** To reach our aim we implemented MatrixInt, a collection of programs that takes as an input a file containing atomic coordinates of at least two protein or nucleic acid chains in PDB format. These can be either two or more subunits within oligomeric proteins; two different proteins; or protein-nucleic acid complexes. The program calculates structural information such as solvent accessibility, secondary structure, hydrogen bonds and hydrophobic contacts, the latter calculated at an atomic rather than residue level. This information is used to infer the strength of interactions occurring at macromolecule interfaces. The output is provided in different formats, all of which can be edited by the user or imported by commonly used programs. As an example, MatrixInt provides 2D contact maps in a format that can be imported into spreadsheet applications and modified by the user to insert additional information or remove information that is not required. Additionally, to aid the user in the visualization and evaluation of the interactions, PyMol scripts are provided that highlight the involved residues in the three-dimensional structures of the macromolecule partners. The Figure shows the interacting residues in two subunits within PDB file 1dps. The web server will be available soon.
**Contact email:** patrizio.dimicco@uniroma1.it




**Left**: the two subunits are shown as cartoon and coloured red and blue; for the interacting residues all-atoms and solvent accessible surface are shown. **Right**: Interactions information imported into a spreadsheet program.

Poster: Structure and Function – SF13

**Anti-cancer drug imatinib binding to the allosteric core of human serum albumin: a molecular docking study.**

Di Muzio E(1), Polticelli F(1,2), Fanali G(3), Fasano M(3), Ascenzi P(4,5)

*(1) Department of Sciences, University Roma Tre, Roma (2) National Institute of Nuclear Physics, Roma Tre Section, Roma (3) Department of Theoretical and Applied Sciences, University of Insubria, Busto Arsizio (4) Interdepartmental Laboratory for Electron Microscopy, University Roma Tre, Roma (5) National Institute of Biostructures and Biosystems, Roma*

**Motivation:** Human serum albumin (HSA) is the most abundant protein in blood plasma. One of the most important functions of HSA is the transport of physiological ligands and exogenous molecules. HSA affects the pharmacokinetics of many drugs, is responsible for metabolic modification of some ligands, renders potential toxins harmless by transporting them to disposal sites and accounts for most of the antioxidant capacity of human serum. Interaction between drugs and plasma proteins is an important pharmacological parameter which has a strong influence on their pharmacodynamic behavior and affects the distribution and elimination of drugs. Imatinib is an inhibitor of the Bcr-Abl kinase activity and induces the cytogenetic remission in the majority of chronic myeloid leukemia patients. Moreover, it is used in the treatment of several cancers, including gastrointestinal stromal tumors. Experimental studies were previously carried out to investigate imatinib binding to human serum heme-albumin in the absence and presence of the ferric heme (heme-Fe(III)). In order to support the interpretation of these data, docking simulations and analysis of local structural similarity between the available HSA three-dimensional structures and a set of 15 other imatinib-binding proteins structures were carried out.
**Methods:** Docking simulations of imatinib on HSA were carried out using AutoDock Vina. For this analysis two different crystal structures were used: the ligand free HSA (PDB code: 1AO6) and the HSA-heme-Fe(III) (PDB code: 1N5U). The simulations were carried out both by keeping all proteins residues rigid and by allowing flexibility of the residues building up the walls of the FA1 to FA7 sites. Imatinib rotatable bonds were kept flexible in all the simulations. Local structural comparison between HSA and other imatinib binding proteins were performed using ASSIST, a software tool for local structural comparison developed in our lab. This program is based on a geometric hashing approach to find the largest subset of similar residues between an input protein structure and a set of structural motifs. A search in the PDB retrieved a total of 9 different imatinib binding proteins. The structural motifs to be compared with HSA structures were composed of the residues involved in the binding of imatinib in this 9 proteins, as determined by LigPlot+.
**Results:** This study was performed with the aim to characterize the binding mechanism of imatinib to HSA and allowed us both to identify the preferential binding sites of the drug, and to investigate the interaction at an atomic level. Docking simulations of imatinib binding to HSA indicated the preferential binding of the drug to the FA1 and FA7 sites. On the other hand, docking simulations of imatinib binding to HSA-heme-Fe(III) indicated the preferential binding of the drug to the FA7 and FA2 sites. Furthermore, ASSIST analysis of the local structural similarities between HSA and a set of imatinib binding proteins identified four different sets of residues in HSA which display chemical and geometric similarity with the imatinib binding pocket of the spleen tyrosine kinase Syk (PDB code: 1XBB). The best match is obtained with HSA residues Ile264, Ser287, Leu260, Glu230 and Leu219 located on the walls of the FA7 site, the calculated all-atoms RMSD value between these residues and the

corresponding Syk residues being 1.78 Å. To summarize, results obtained indicate the preferential binding of imatinib in FA1 and FA7 sites in ligand-free HSA, and in FA7 and FA2 sites in HSA-heme-Fe(III). These data confirm the results obtained by experimental studies on the heme and ligand-free HSA and provide an atomic level view of the imatinib-HSA(-heme-Fe(III)) complexes.

**Contact email:** fabio.polticelli@uniroma3.it

Poster: Structure and Function – SF14

**Intrinsically disordered ProDom and pFam domains: how large is the human unfoldome?**

Deiana A(1), Giansanti A(1,2)

*(1)Department of Physics, Sapienza University of Rome, Italy (2)INFN, Sezione di Roma1, Rome, Italy*

**Motivation:** Protein science has been pervaded by the concept of unfoldome in the last decade or so. Since estimates seem to depend on the protocol adopted, there is interest in finding criteria to set the boundaries of the realm of protein genes coding for proteins that are unfolded under physiological conditions. In this note we investigate the problem of how wide is the unfoldome in the human proteome (HP). We adopt a simple but restrictive criterion based on conserved disorder and we show that the number of predicted disordered proteins in HP is consistent with current estimates of the number of protein sequences that escape structural SCOP and CATH classification.
**Methods:** IUPred, fine-tuned to match sensitivity with selectivity, has been used to predict disordered residues. The overlap between predicted disordered segments with ProDom and pFam-A domains has been extensively investigated and a criterion that quantitatively takes into account the degree of conservation of disorder is derived from the observed statistics.
**Results:** ProDom and Pfam domains have different, somehow opposed propensities to accommodate disordered residues, with ProDom domains far more permeable to disorder than Pfam domains. This is coherent with the different criteria and protocols followed in defining those two databases. If one defines as disordered a proteins sequence that contains at least a domain that is disordered in at least 50% of its residues then the fraction of disordered proteins in the HP is comprised between 0.04 and 0.14. This estimates are discussed in terms of previous estimates of singletons (homologically unrelated protein sequences) and in terms of consistent estimates of conserved vs. not conserved disorder.
**Contact email:** andrea.giansanti@roma1.infn.it

Poster: Structure and Function – SF15

**Use of Homology Information in Replica Exchange Methods for De-novo Protein Structure Prediction**

Iacoangeli A.(1), Tramontano A.(1,2), Marcatili P.(3)

*(1) Department of Physics,La Sapienza University of Rome, Roma (2) Istituto Pasteur – Fondazione Cenci Bolognetti, Roma (3) Center for Biological Sequence Analysis Department of Systems Biology, Technical University of Denmark,Lyngby,DENMARK*

**Motivation:** The performance of ab-initio methods for De-novo protein structure prediction strongly depends on two components: an energy function that is expected to have a stable minimum near the native structure, and a method for finding this minimum through the protein conformational space. Monte Carlo Replica Exchange methods provide a major advance to the latter point, outperforming widely used methods such as the original Monte Carlo Simulated Annealing, in terms of sampling low-energy states. The Hamiltonian Replica Exchange is one of the most promising among these novel methods. The aim of our project is to use the homology information to develop a method for De-novo protein structure prediction. Our approach improves both the exploitation of the energy function potentialities and the ability of spacing the energy landscape when compared to the state of the art of this class of methods.
**Methods:** It is accepted that homologous proteins have similar native structures, thus their energy landscapes, and therefore their hamiltonians, must be similar nearby such conformation at least. This allowed us to develop, starting from a standard ab-initio protocol implemented using the Rosetta modelling suite, a Replica Exchange method in which the hamiltonian of homologous proteins are used to characterize the replicas. More in detail, the replicas are characterized with the sequences of the homologous proteins. These sequences are tailored and selected according to the target protein specifics. Homologs are selected with Blast then their sequences are aligned with MAFFT and tailored to match the native sequence size. These sequences are clustered according to their similarity with CD-HIT and cluster centroids are selected. The resulting sequences are used to characterized the 70% of the replicas involved in the simulations.
**Results:** In a pilot study on 40 different domains selected from the SCOP database, we observed that, thanks to the exchange of homology information during the Monte Carlo simulation, our method is able to both improve the exploration of the energy landscape and reach better solutions in terms of distance from the native structure. Our protein set consists of proteins, which differ for both type, including α, β, α+β and α/β proteins, and size, ranging from 55 to 208 amino acids long. In this study we assessed the performance of our method comparing the predicted structures with the ones obtained by both a classic Rosetta Monte Carlo method and a classic Hamiltonian Replica exchange method. Considering the lowest energy structures predicted for all proteins in the protein set, our method is able to reach significantly better structures in about 45% of cases, similar ones in about 38% of cases and worse ones in about 17% of cases. These preliminary results show that our novel approach can greatly improve the performance of ab-initio methods for protein structure prediction, both overcoming errors and approximation in the energy functions and allowing for a more efficient search of the conformational space.
**Contact email:** anna.tramontano@uniroma1.it , alfredo.iacoangeli@live.it

Poster: Structure and Function – SF16

**Searching for targets of mycotoxins in the context of autism spectrum disorders: an example of computational analysis**

Scafuri B(1), Facchiano A(2), Raggi ME(3), Marabotti A(1,3,4)

*(1) Department of Chemistry and Biology, University of Salerno, Fisciano (SA) (2) Institute of Food Science, CNR, Avellino (3) IRCCS "E.Medea" Ass. "La Nostra Famiglia", Bosisio Parini (LC) (4) Institute of Biomedical Technologies, CNR, Segrate, Milano*

**Motivation:** Autism spectrum disorder (ASD) is a disorder of neural development characterized by impaired social interaction and verbal and non-verbal communication, and by restricted, repetitive or stereotyped behavior. A notable increase in ASD incidence induces in thinking that increased exposure to environmental factors, in addition to a direct genetic component, can play a role in its etiology. In people with ASD, gastrointestinal disorders are commonly reported, and some studies have found that a gluten-free casein-free diet can improve the clinical features of these patients. The negative effect of gluten and/or casein could be attributed to a direct role played by food components, but also to food contaminants, which, binding different proteins in the body, could directly entail the development of the reaction that would expose entire organism, including the CNS, to negative effects of xenobiotics. Among food contaminants, mycotoxins play a significant role and several negative effects on human health are already known. The aim of this work to identify possible protein targets for mycotoxins and verify if the bond between these target and mycotoxins could be involved in the onset of ASD in body of genetically predisposed patients.
**Methods:** The idTarget server [1] has been used to search for protein targets of twelve mycotoxins (ochratoxin, gliotoxin, aflatoxin B1, aflatoxin B2, aflatoxin M1, aflatoxin M2, aflatoxicol, a-zearalanol, b-zeralanol, zearalenone, deoxynivalenol, patulin). idTarget is a web server for identifying biomolecular targets of small chemical molecules, performing an inverse docking approach among proteins whose structure is present in PDB. The results were analyzed in order to select human protein targets expressed in the brain or involved in brain diseases, common to two or more mycotoxins, for further characterization made by PDB, KEGG and Uniprot databases available on-line. For all the finally selected targets, direct docking with each mycotoxin has been carried out, using Autodock 4.2 [2], in order to identify the mycotoxins' binding site on each selected protein target and evaluate the strength of the interaction.
**Results:** For each mycotoxin for which a reverse docking search was made, idTarget returned thousands of possible protein targets. Of these, the first 106 possible targets, ordered by decreasing binding energy for each mycotoxin, were further evaluated. Each one of the 106 targets of each mycotoxin has first been compared to all the other 106 possible targets of the other 11 mycotoxins. From this comparison, 537 targets were found in common to two or more mycotoxins. These possible targets have been analyzed for the organism in which they are expressed, the molecular function, the tissue specificity and their possible involvement in brain disease. Targets expressed in the brain of the human being, or implicated in brain disease, or in common to five or more mycotoxins were selected for further characterization. After this selection, 128 possible targets have been obtained, of which the best 20 targets have been analyzed for their crystallographic structure. Those structures with higher quality were further selected, discarding or replacing those structures with gaps or multiple residues conformations, with other structures of the same target protein more suitable for docking. A final list of 18 targets was obtained and these proteins

were evaluate for their ability to bind mycotoxins using a traditional docking approach. The results of the direct docking show that 9 of the 18 selected protein interact with several mycotoxins with a predicted binding energy lower than -7 kcal/mol and with resulting clusters of poses bigger than 30. In particular, aflatoxins show many interactions with several proteins selected with criteria formerly described, suggesting their potential involvement in the pathogenesis of neurologic diseases, including ASD. The next step of this study will be the validation of these results with an in vitro experimental approach to test the direct binding of mycotoxins to each selected protein. Finally, these results will be integrated with other data collected on patients and controls, to test new hypotheses about the role of these environmental factors in this disease.

**Contact email:** amarabotti@unisa.it

**Supplementary information:**Acknowledgments: This work was made in the frameshift of the project GR-2009-1570296: “The relationship among food, mycotoxins, gastrointestinal disorders and autism: a multidisciplinary approach for the molecular investigation” funded by the Ministry of Health (program “Ricerca Finalizzata e Giovani Ricercatori 2009”). AF acknowledges the contribution of Flagship InterOmics Project (PB.P05, funded and supported by the Italian Ministry of Education, University and Research and Italian National Research Council organizations) References: [1] Wang J-C et al. Nucleic Acids Res. 2012, 40, W393-9. [2] Morris GM et al. J. Comp. Chem. 2009, 30, 2785-91.

Poster: Structure and Function – SF17

**PROTEIN THERMAL STABILITY PREDICTION WITHIN HOMOLOGOUS FAMILIES**

Fabrizio Pucci(1) Marianne Rooman(1)

*(1) Department of BioSystems, BioModeling & BioProcesses, Universite Libre de Bruxelles, Roosevelt Ave. 50, 1050 Bruxelles, Belgium.*

**Motivation:** The understanding of the thermal stability of proteins remains one of the key questions of protein science. It is not only important for theoretical reasons aimed at studying the adaptive strategies used by the organism to live in extreme environmental conditions, but it has also a wide range of applications for all the bioprocesses in which proteins are involved. Unfortunately it is still highly non-trivial to have precise predictions about thermal stability since the results are in general family-dependent and seem sometimes even contradictory. Here we develop and compare some tools in view of making the thermal stability prediction more systematic and quantitative. Seven different methods that predict the best descriptor of the thermal stability, namely the melting temperature Tm, were constructed and compared. They are all based on the protein sequence; some also use information about the protein structure, the Tm of some proteins belonging to the same homologous family, or the environmental temperature of the host organism.
**Methods:** We present different tools that we have built and analyzed in view of getting an accurate Tm prediction. These methods were applied to predict the melting temperature of 45 proteins belonging to 11 homologous families, of which the Tm has been measured experimentally. The first two methods are based on the derivation of the melting temperature Tm in terms of the environmental temperature of the host organism Tenv. It is well known that Tm and Tenv are correlated, since thermophilic organisms necessarily host thermostable proteins (even if the converse is not true). More precisely, the first method to compute Tm is based on a (Tm,Tenv) regression line obtained on a dataset S of 166 proteins with known Tm and Tenv, which includes the 45 proteins belonging to the 11 abovementioned families. The second method is family-dependent: the (Tm,Tenv) regression lines are computed inside the 11 homologous families. In the third method the Tm of a target protein is estimated as being equal to the Tm of the protein of the same family that shows the highest sequence identity. The fourth, fifth and sixth methods use as main tools statistical potentials derived from sets of protein structures. For that purpose, the total dataset S was divided in two different subsets consisting of either thermostable (Sth) or mesostable (Smes) proteins only. Interresidue distance potentials and backbone torsion angle potentials were computed from S, Sth and Smes following the inverse Boltzmann law. The folding free energy of a target protein was computed as a linear combination of the distance and torsion potentials; three folding free energies were thus obtained: DG, DGmes, DGth. Two Tm predictions were made from these folding free energies: the first considered Tm to be linearly correlated to DG and the second to DGmes-DGth. The coefficients of the linear combination of potentials were identified so as to minimize the difference between predicted and experimental Tm's. The first of these two methods actually computes the thermal stability from the thermodynamic stability. The second takes more properly the temperature dependence of the amino acid interactions into account, since in each of the ensembles Sth and Smes only proteins with given thermal properties are considered. The sixth method also exploits the folding free energies DG, DGmes and DGth. It is based on the estimation of the full protein stability curve DG(T) that best fits these three energy values, which are associated with three different temperatures, estimated as a function of the average Tm of the proteins in the three sets S, Smes and Sth.

After parameter identification to minimize the difference between estimated and experimental Tm's, the melting temperature was extracted as the temperature where DG(T) is zero. The last Tm prediction method is based on the amino acid sequence only, and is derived from the proteins in the set S. In particular, a predictor for the melting temperature in terms of the protein's amino acid composition was constructed, using simple parameter optimization.

**Results:** Using these 7 methods, we predicted the melting temperature of the set of 45 proteins belonging to 11 homologous families, with the jack knife cross validation technique. The standard deviation k between the experimental and predicted melting temperatures was computed and compared for the different prediction methods. Some methods show better performances than others: the value of k ranges roughly from about 10 °C to 20 °C in cross validation. The good and the feeble points of each method were analyzed, as well as future directions, aiming at constructing a unique, fast and accurate method on the basis of a combination of the best of these methods.

**Contact email:** fabriziopucci81@gmail.com

Poster: Structure and Function – SF18

**Structure prediction of antibody hypervariable H3 loops using random forest**

Messih MA(1), Lepore R(1), Marcatili P(3), Tramontano A(1,2)

*(1) Department of Physics, Sapienza University, 00185 Rome, Italy (2) Istituto Pasteur—Fondazione Cenci Bolognetti, Sapienza University, Rome, 00185, Italy (3) Center for Biological Sequence Analysis, Department of Systems Biology, Technical University of Denmark, Lyngby, Denmark*

**Motivation:** Antibodies are a class of Y-shaped proteins, consisting of two identical heavy and light chain pairs, that the immune system uses to identify and neutralize foreign pathogens such as bacteria and viruses. They have the remarkable ability to recognize foreign targets, called antigens, and bind to these with extraordinary affinity and specificity. The antigen binding site, located on the upper tips of the Y shape, is formed by six hypervariable loops also referred to as the complementary determining regions (CDRs). Three of the loops belong to the variable domain of the light chain, (L1, L2 and L3) and three to the variable domain of the heavy chain (H1, H2 and H3). The framework regions of antibodies are fairly well conserved, while the structural variations mostly occur in the CDR loops. The variability in these CDR loops, in terms of loop length and amino acid compositions, is the main reason of the antibody capability to bind many different antigens. Predicting the three-dimensional structure of these loops is essential for the computational design and re-engineering of novel antibodies with enhanced affinity and specificity. The canonical structure model allows high accuracy prediction of the conformation for five of the CDR loops, based on the presence of key residues in specific positions of the antibody sequence. On the other hand, only a partial canonical structure model exists for the H3 loop, which allows the prediction of the structure of the N residues closer to the framework. Accordingly the prediction accuracy of H3 loops is not equally satisfactory, an important drawback since the H3 loop is central in the binding site and therefore likely to be essential in determining the antibody-antigen interaction.

**Methods:** Given the central position of the H3 loop in the antigen binding site, there are several interactions with the other CDR loops, as well as with the framework, that could affect the conformation of H3. In line with this hypothesis, we developed a method that takes advantage of both sequence and structural related features (e.g. germline families, canonical structure, contact matrix, BLOSUM similarity matrix and structure alphabet substitution matrix). The selected features are used to train a Random Forest (RF) machine learning regression model that, given a target H3 loop, predict the 3D distance between the target loop and all other loops in the training dataset. The loop with the shortest predicted distance is used as template to build the model of the target H3. When the value of the predicted 3D distance is above a specific threshold, a re-ranking of the top scoring RF templates is performed based on a knowledge-based potential approach whose definition relies on the presence of short range contacts between H3 residues and their structural environment.

**Results:** We tested the method on a known benchmark (Rosetta Antibody dataset) consisting of 53 antibodies with H3 length varying from 4 to 22 residues. The method outperformed Rosetta Antibody (the most accurate available tool) for all loop length ranges, achieving an improved accuracy in 62% of cases and significantly improved accuracy (ΔRMSD >1Å) in 33%. Interestingly, the actual best template is correctly identified in one fourth of the cases and in 13% of cases the method was able to produce a model in the sub-angstrom accuracy range (RMSD < 1 Å). Comparatively, Rosetta Antibody achieved the same accuracy

range in one case only. In order to assess the performance of the model in a more realistic setting, we also tested the method on recently deposited antibody structures (added to PDB after October 15, 2012 (the date when we downloaded our training dataset)) and obtained equally satisfactory results (backbone RMSD = 2.5 ± 1.5 Å). Moreover, in the 14% of cases the selected model was in the sub-angstrom accuracy range. These results are rather encouraging considering that the latter dataset is enriched of antibodies with very long H3 loops. Another important aspect of the method is that it is significantly faster than Rosetta Antibody with an average CPU-time of 5 minutes (CPU speed: 2.5 GHZ and RAM: 5 GB) per antibody to be compared with hours and sometimes days. In conclusion, we developed a new method for predicting the structure of H3 loops of immunoglobulin, a rather elusive and complex problem that is however essential for obtaining an accurate view of the antibody-antigen binding sites. The method compared favorably with the most accurate available tool with an improvement of almost 1 Å in terms of average backbone RMSD, providing evidences that high-resolution modeling of H3 loops can be possible using comparative modeling techniques. We also showed that the H3 structure environment constitutes a valuable source of information that can be used to tackle, with an appropriate degree of success, the difficult problem of predicting proper conformations of long loops.
**Contact email:** anna.tramontano@uniroma1.it

Poster: Structure and Function – SF19

**Investigating the T cell receptor (TCR) activation mechanism: a combination of docking simulations and experimental results**

Vangone A, Bonvin A.M.J.J.

*Bijoveth Center for Biomolecular Research, Faculty of Science, Utrecht University, Utrecht, The Netherlands*

**Motivation:** Human T cells mediate potent antimicrobial immune response. To carry out their function, T cells express a T cell antigen receptor (TCR) after recognition of phosphorylated prenyl metabolites as antigens in presence of antigen-presenting cells. However, the molecular mechanism behind the activation of TCR remains unclear. Recent evidence point to the role of butyrophilin BNT3A1 in this response as antigen-presenting, with two possible (and controversial) mechanisms: i) direct presentation of phosphoantigen by BTN3A1 to TCR (De Libero et al., Nature Immunology 2013) or ii) binding of the phosphoantigen to the intracellular domain of BTN3A1 with consequent conformational changes of BTN3A1 and recognition of TCR (Adams et al., J. Biol. Chem. 2012). Considering the new large potential and interest in using nanobodies (camelids antibodies characterized by only the heavy chain still fully capable of antigen binding, that are very stable and so very promising in diagnosis and therapeutic application), the group of Prof. Van der Vliet from the Cancer Center Amsterdam tested nanobodies on this system to help in clarifying the possible activating mechanism of the TCR. They found that the nanobody named VHH19 actually binds and neutralize the TCR, blocking subsequent activation of the T cells. We then performed docking simulations on the whole TCR, VHH19 and BTNA31 system, to model it and direct further lab experiments.
**Methods:** The amino acids sequence of the nanobody VHH19 was provided from the group of Prof. Van der Vliet; the VHH19 structure was then modeled with Modeller (Blundell et al., J. Mol. Biol. 1993) based on the sequence alignment between the nanobody and the template obtained from a BLAST search. For the TCR and BTN3A1 proteins, the experimental x-ray structures (PDB entry 1HXM and 4F9L, respectively) were used for the docking simulations. They were performed by HADDOCK (Dominguez et al., JACS 2003) webserver, in which the interaction in the TCR-BTN3A1 and TCR-VHH19 complexes was tested and then compared, taking advantage of CONS-COCOMAPS web tool (Vangone et al., BMC Bioinformatics 2012), that analyses an ensemble of models complexes to point out the most promising key residues for the interaction.
**Results:** From the docking simulations of the TCR-VHH19 compelx and their consecutive analysis, we found that the TCR mostly involved its CDRs loops in the interaction with VHH19, identifying as key residues in particular amino acids belonging to γ2 and γ3 loops. Interestingly, this TCR binding region totally overlap with the one we found in the TCR-BTN3A1 simulations, showing that a possible of competition exists between the VHH19 and BTN3A1 since they both bind the same TCR region. Taking into account our results and the experimental data from our collaborators, since the presence of VHH19 neutralize the TCR, and both VHH19 and BTN3A1 seems to compete for the same TCR region, these result suggest the idea that BTN3A1 is actually binding the TCR to activate it.
**Contact email:** a.vangone@uu.nl

Poster: Structure and Function - SF20

**Loop insertions in helices: a survey of the Protein Data Bank**

Balasco N (1,2), Riccio R (1,3), De Simone A (4), Vitagliano L (1)

*(1) Institute of Biostructures and Bioimaging, CNR, via Mezzocannone 16, I- 80134 Napoli, Italy (2) Second University of Napoli, Via Vivaldi 43 - 81100 Caserta, Italy (3) University of Napoli "Federico II", via Mezzocannone 16, I- 80134 Napoli, Italy (4) Division of Molecular Biosciences, Imperial College South Kensington Campus London SW7 2AZ, UK*

**Motivation:** Insertions and deletions (indels) in secondary structure elements are rather unusual events. Indeed, insertions/deletions within alpha-helices and beta-strands are less tolerated than indels occurring in unstructured regions. Although literature data are somewhat contrasting, it is generally assumed that these mutations have remarkable destabilizing effects on the overall protein structure. Unlike amino acid substitutions, whose mechanisms have been studied intensively, indels remain less understood and pose several unanswered questions. Previous literature studies have reported that for helices, insertions occurred within four residues of the helix termini and typically resulted in a helical extension of a few residues. By comparing the crystal structures of proteins isolated from different species we have detected intriguing examples of loop insertions in helices. Particularly relevant is the case of the protein system of the elongation factors EF-1A/EF-Tu. Interestingly, a straight helix is present in the E. coli EF-Tu (Song et al. JMB 1999) whereas the same helix presents a large insertion in the homologue EF-1A isolated from the archaeon S. solfataricus (Vitagliano et al. EMBO J. 2001). Even larger is the insertion detected in the structure of the eukaryal EF-1A isolated from yeast(Andersen et al. Mol Cell 2000). In order to gain insights into the role and the frequency of these large insertions we searched the entire Protein Data Bank looking for similar structural motifs.
**Methods:** The statistical analyses were carried out on a non-redundant ensemble of X-ray PDB protein structures. In particular, we selected 6690 single polypeptide chains which satisfy the following criteria : resolution better than 2.5 Å, R-factor < 0.3, and sequence identities lower than 25%.
**Results:** The survey of the database highlighted the presence of nearly fifty large insertions (number of residues > 9) within helices. These insertions were located either at the middle or at the termini of the helices. Interestingly, the inserted residues could adopt different type of structures. In particular, they could (a) present an irregular loop structure, (b) form beta-hairpins or (c) be completely disordered in the crystal structure. The analysis of the conservation of these typically exposed motif in homologous protein classes indicates that they are generally well preserved although their length and amino-acid composition are characterized by a large variability. The implications of these findings for the biological role of these insertions and for the design of innovative protein motifs will be presented.

Poster: Proteomics – PRO1

**The High Clustering Coefficient Confers on Natural Protein Interaction Networks Self Organization Properties**

Galeota E(1), Ferrante A(1), Gravila C(1), Castiglione F(2), Bernaschi M(2), Cesareni G(1)

*(1) Department of Biology, University of Rome 'Tor Vergata', Rome, Italy (2)Istituto per le Applicazioni del Calcolo "M. Picone", CNR, V.le del Policlinico 137, Rome, Italy*

**Motivation:** Cell organization is governed and maintained via specific interactions between its constituent macromolecules. Here we investigate the role of the topological characteristics of the protein interaction network in promoting cell organization. Comparison of the experimentally determined protein interactions networks in different model organisms have revealed little conservation of the specific edges linking ortholog proteins. Nevertheless, some topological characteristics of the graphs representing the networks – namely non random degree distribution and high clustering coefficient – are shared by networks of distantly related organisms.
**Methods:** We have used ProtNet, a stochastic model representing a computer stylized cell to ask questions about the dynamic consequences of the topological properties of the static graphs representing protein interaction networks.
**Results:** By using a novel metrics of cell organization, we show that natural networks, differently from random networks, can promote cell self organization. Furthermore the ensemble of protein complexes that form in pseduocells, that self organize under the interaction rules of natural networks, are more robust to perturbations. The analysis of the dynamic properties of networks with a variety of topological characteristics lead us to conclude that self organization is a consequence of the high clustering coefficient, while the scale free degree distribution has little influence.
**Contact email:** eugenia.galeota@gmail.com

Poster: Proteomics – PRO2

**Microvesicles in cancer: a meta-analysis of proteins in different types of tumours**

Merelli I(1) Galluccio N(1) Mezzelani A(1) Milanesi L(1)

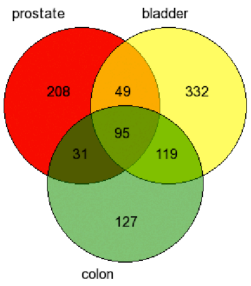
*(1) Istituto di Tecnologie Biomediche - Consiglio Nazionale delle Ricerche*

**Motivation:** Cancer cells release small membrane vesicles, called microvesicles, with an increased rate in malignant tumours. Although recent progresses in this area have revealed that tumour-derived microvesicles play multiple roles in tumour angiogenesis and metastasis, their biological significance is not completely understood. Microvesicles contain a multitude of biologically active molecules, including mRNAs, proteins and microRNAs that are transmitted to the target cells modifying their cellular physiology. For instance, tumour cell-derived mRNAs, which encode pro-angiogenic factors can be expressed in target cells upon their microvesicular transfer. In addition, several proteins, including pro-angiogenic factors can be directly transferred from tumour cells to target cells [1].
**Methods:** Many data about microvesicles are available in literature and in particular there are two public available databases, Exocarta [2] and Vesiclipedia [3], that collect data about experiments concerning microvesicles. We performed a meta-analysis of three types of cancer derived microvesicles (all analysed by using mass spectrometry): prostate cancer (from the experiment of Hosseini-Beheshti et al. [4]), bladder cancer (from the experiment of Welton et al. [5]) and colon cancer (pooling three experiments [6,7,8]). By annotating proteins found in the microvesicles released by these tumour using Gene Ontology, Pathway descriptions, and Protein Domains, we were able to find same interesting common characteristic shared by this kind of tumour. All the results provided are FDR corrected and computed using David Bioinformatics Resource [9].
**Results:** The first important result to highlight is that there is a considerable overlap between proteins that have been identified in the different microvesicles. Between the 45% and the 66% of the proteins are shared by at least two type of cancer and between the 16% and the 25% by all three types of cancer (Fig. 1). Concerning the Molecular Function GO annotation, GTPase ($p<6.1e-4$) is present in all the tumour, which provides hints of enhanced protein biosynthesis, control and differentiation during cell division, and translocation of proteins through membranes. Nucleotide binding ($p<1.6e-4$) has been also identified in all the cancers, which correlate with enhanced gene expression profiles. In relation to the Cellular Component GO annotation, the terms that appear in all the tumours are correlated with microvesicles release, communication and transfer: membrane-bounded vesicles ($q<4.3e-9$), vesicles-mediated transport ($q<4.0e-8$), and non-membrane-bound organelle ($q<1.4e-5$). Looking at the Biological Process GO annotation, cytoskeleton organization ($q<1.4e-4$) and cellular macromolecular complex assembly ($q<6.0e-4$) are among the most overrepresented terms. Also the annotation against Kegg Pathway provided interesting results, because there is an enrichment of proteins correlated with focal adhesion ($q<9.8e-2$), gap junction ($q<2.1e-2$), glycolysis/gluconeogenesis ($q<4.2e-3$), and regulation of actin cytoskeleton ($q<4.2e-3$). There is also a notable enrichment of proteins correlated with the Pathogenic Escherichia Coli infection ($q<6.2e-9$), which should be investigated in details, but may be generally correlated to immunological response (Fig. 1). Analysing the Reactome pathway annotations, there is general overrepresentation of terms correlated with the metabolism of proteins ($q<4.7e-6$), although looking at each single tumour other terms like cell cycle, telomerase maintenance, apoptosis, integrin/cell surface, and Wnt signalling have enriched patterns. Concerning Biogrid pathway, there are evidences concerning Glycolysis ($q<3.7e-1$), and Hypoxia ($q<3.5e-1$), which clearly relate to the lack of

oxygen in this type of cancers. The last annotation has been achieved looking at the distribution of Interpro domains, which suggests an enrichment of Tubulin (q<2.5e-6) and Hsp (q<5.6e-4) domains.

**Contact email:** ivan.merelli@itb.cnr.it

**Supplementary information:**Figure Fig 1. a) Venn diagram of the proteins identified in microvesicles-derived from three types of cancer. b) Kegg Pathways enrichment analysis of proteins common to the three types of cancer. References [1] Ratajczak J et al. Embryonic stem cell-derived microvesicles reprogram hematopoietic progenitors: evidence for horizontal transfer of mRNA and protein delivery. Leukemia, 20, 847-856, 2006. [2] Mathivanan S et al. ExoCarta 2012: database of exosomal proteins, RNA and lipids. Nucleic Acids Research, D1241-4, 2012. [3] Kalra H et al. Vesiclepedia: A compendium for extracellular vesicles with continuous community annotation. PLoS Biology, 12, e1001450, 2012. [4] Hosseini-Beheshti E et al. Exosomes as biomarker enriched microvesicles: characterization of exosomal proteins derived from a panel of prostate cell lines with distinct AR phenotypes. Mol Cell Proteomics, 11(10):863-85, 2012. [5] Welton JL, Proteomics analysis of bladder cancer exosomes. Mol Cell Proteomics, 9(6):1324-38, 2010. [6] Choi DS et al. Proteomic analysis of microvesicles derived from human colorectal cancer cells. J Proteome Res., 6(12):4646-55, 2007. [7] Mathivanan S et al. Proteomics analysis of A33 immunoaffinity-purified exosomes released from the human colon tumor cell line LIM1215 reveals a tissue-specific protein signature. Mol Cell Proteomics, 9(2):197-208, 2010 . [8] Huber V et al. Human colorectal cancer cells induce T-cell death through release of proapoptotic microvesicles: role in immune escape. Gastroenterology, 128(7):1796-804, 2005. [9] Systematic and integrative analysis of large gene lists using DAVID Bioinformatics Resources. Nat Protoc., 4(1):44 -57, 2009.




| Term | P-Value | Benjamini |
|---|---|---|
| Pathogenic Escherichia coli infection | 1,0E-10 | 6,2E-9 |
| Antigen processing and presentation | 1,4E-4 | 4,4E-3 |
| Regulation of actin cytoskeleton | 2,0E-4 | 4,2E-3 |
| Glycolysis / Gluconeogenesis | 2,7E-4 | 4,2E-3 |
| Gap junction | 1,7E-3 | 2,1E-2 |
| Leukocyte transendothelial migration | 5,7E-3 | 5,7E-2 |
| Endocytosis | 8,7E-3 | 7,5E-2 |
| Focal adhesion | 1,3E-2 | 9,8E-2 |

Poster: Proteomics – PRO3

**Rigidity and flexibility properties of mutational Protein-protein interaction networks involved in muscular dystrophies**

Sharma A(1),Ferraro MB(1)(2), Maiorano F(1), Varavallo A(1), Guarracino MR (1)

*(1) Laboratory for Genomics, Transcriptomics and Proteomics, High Performance Computing and Networking Institute, National Research Council, Via P. Castellino, 111, Naples, Italy (2) Department of Statistical Sciences, La Sapienza University of Rome, P.le A. Moro 5, Rome, Italy*

**Motivation:** Protein-protein interactions are indispensable to the functioning of cells and they take part in many biological processes within a cell. The analysis of protein interaction networks represents a valuable methodology to understand complex biological systems. Biological systems often show rigidity, flexibility or intermediate state in their protein interaction network, depending on environmental cues, stress conditions or genetic diseases. Muscular dystrophies are known to have a genetic cause. This disorder gradually weakens muscles, ultimately leading to death of muscle cells and tissues. Only 33% of patients affected by DMD can be explained by defects in dystrophin gene and other muscle proteins. Our aim is to study the topological and functional effect of mutations on protein interaction maps related to muscular dystrophies.
**Methods:** We accumulated the first order network of corresponding proteins from a list of causative genes in listed by Kaplan et al, 2011[1] across various databases including high throughput methods and experimentally known interactions. Modular organization of network was assessed by ModuLand framework. Emitting model of ITM probe was used to determine the interference in the networks [2]. Rigidity and flexibility issues of the sub-networks with maximum mean interference were analyzed by KINARI-Lib [3]
**Results:** Muscular dystrophic diseases share mutated proteins in protein-disease network indicating the comorbidity in the muscular diseases. The modules in the first order protein interaction network of 207 proteins showing mutations in MD are involved in different molecular functions such as intracellular transport, with significant p-value. We have detected core nodes in modules receiving maximum visits from duplets of key seed nodes showing mutation with larger mean interference values. The CACNA1S and CALM1 sub-networks demonstrate maximum flexibility with largest percentage of degrees of freedom, while 14-3-3 proteins family, which binds functionally diverse signaling proteins, are observed to be the most rigid sub-networks in muscular dystrophic protein networks with least degree of freedom [2].
**Contact email:** ankush.sharma@na.icar.cnr.it
**Supplementary information:**Reference: 1. Kaplan J-C (2011) The 2012 version of the gene table of monogenic neuromuscular disorders. Neuromuscul. Disord. 21, 833–6 2. A.Stojmirovic and Y.-K. Yu. ITM Probe: analyzing information flow in protein networks. Bioinformatics, 25 (18):2447-2449, 2009. 3. Naomi Fox, Filip Jagodzinski, and Ileana Streinu, "KINARI-Lib: a C++ library for pebble game rigidity analysis of mechanical models", Minisymposium on Publicly Available Geometric/Topological Software, Chapel Hill, NC, USA, June 17 and 19, 2012.

**Hybrid simulation of chemically reacting systems with delays**

Caravagna G [1] Bortolussi L [2]

*[1] Dipartimento di Informatica, Sistemistica e Comunicazione, Università degli Studi Milano-Bicocca, Italy. [2] Department of Mathematics and Geosciences, Università degli Studi di Trieste, Italy.*

**Motivation:** In the last decades, we have assisted to an increased interest in computational modelling approaches of complex biological systems, also thanks to the easy access to high performance computing resources available in the so-called high-throughput era. However, despite a considerable methodological progress (e.g., deterministic, stochastic or hybrid approaches in a broad sense), much can still be gained from combining theory and algorithms from different research areas (i.e., biophysics, logic, statistics). This will increase our ability of modeling a biological system with the mathematical representation which best suits with the underlying "physics" of the system, and offers the best efficiency-accuracy trade-off to carry out model analysis based, for instance, on simulation.
**Methods:** In this respect, we present a novel technique to combine a hybrid representation of chemical reactions with delayed events, two features which have been studied, so far, separately. On the one hand, the hybrid representation is a convenient way to efficiently represent multiscale systems where either the molecular counts span over different orders of magnitude (e.g, from few proteins to abundant chemicals) , or the reactions involved happen at different time-scales (e.g., from translation events to cell-level kinetics). In this kind of systems, the analysis of which becomes quickly intractable with the standard purely stochastic approaches, hybrid approaches exploit a mean-field approximation for, e.g., aboundant chemicals, and a discrete-stochastic representation for molecules present in few copies. Thus, this joint deterministic-stochastic representation still captures the intrinsic fluctuations of some components, while resulting in a more efficient simulation. On the other hand, delays are often used to approximate missing dynamical components, at any level of abstraction. Generally, they are used to abstract complex and often only partially known sequential dynamics into a single step, once an estimate of the duration of the dynamics is available.
**Results:** In this work, we present a sound mathematical characterization of this wide class of models, coherently integrating the previous approaches to model hybrid systems and delayed systems. In particular, we are able to give a full mathematical specification of the process underlying the system, which is essentially a non-Markov process (due to the non-Markovian delayed transitions), combined with the vector field modeling the flow of the continuous (i.e., abundant) mean-field variables, which is now given by a set of delayed differential equations. We characterize this time-inhomogenous process by a master equation that gives the probability, in time, of the modeled system to be in any of its possible reachable states, for any initial condition. Despite this equation being unsolvable, as usual, we derive an exact algorithm to sample from that distribution, so to gain insights on the temporal dynamics of the system by repeatedly simulating such a process. We show the applicability of the approach by considering some simple multiscale systems that, however, could not be easily analyzed with standard Gillespie-like approaches. In addition, we test the approach to analyze the effects of a delayed immune-response on the growth of an immunogenic neoplasm. In particular, we investigate the emergence of non-linear behaviors, such as oscillations, to the presence of delays.

Poster: Stochastic modeling – SM2

**PyTSA: Analysis of stochastic biochemical time-series made easy**

Caravagna G, De Sano L. Antoniotti M. (1)

*(1) Dipartimento di Informatica Sistemistica e Comunicazione, Università degli Studi Milano-Bicocca.*

**Motivation:** The study of biological systems witnessed a prominent cross-fertilization between experimental investigation and computational methods, thanks to many different modelling approaches developed within different research areas (e.g, biophysica, logics) to describe natural systems. In most cases, the considered models are purely deterministic but, recently, an increasing interest towards stochastic approaches has emerged. In general, regardless of the mathematical framework modeling the stochasticity within the system components (e.g., the Gillespie exact and approximated approaches, non-Markovian simulation or hybrid systems to name but a few), most analyses rely on evaluating ensemble of simulations under different model conditions and parameters. This data prediction phase eventually requires an efficient data analysis pipeline to interpret the model predictions via, e.g., averages and probability distributions provided by the model intrinsic/extrinsic fluctuations. Despite most of the efforts are often devoted to the definition of simulation techniques, less has been done to automatize data analysis, a task which is inherently time-consuming and definitely error prone. Thus, researchers would benefit from a tool specifically tailored to automatize, in easy fashion and with stochastic-specific language, data analysis.
**Methods:** For this reason, we developed a tool which allows to analyze results coming from stochastic biochemical simulation, in a very simple way. More precisely, PyTSA (Python Time-Series analyzer) implements some complex data analysis over a dataset of time-realizations of a stochastic model.
**Results:** The current pyTSA version supports the following plots, in many graphical formats: single traces (single panel or multi panel), average and standard deviation of a dataset (with barplot or traces, single panel or multi panel), 2D/3D probability density function of a quantity at some specific time point (with normalization and gaussian fit), 2D/3D time-varying probability density function of a quantity in a time-interval (heatmap or surface); 2D/3D phase-space and (synthetic) western blots. pyTSA scripts can be processed in a pipeline with any simulation tool outputting time-series, and intuitive commands allow to perform the above analysis. All in all, pyTSA can be also embedded in any non-biological stochastic simulation framework, and should eventually allow the researcher to focus on system's simulation rather than on data analyses, thus speeding up the usual pipeline of model-definition, testing and refinement.
**Contact email:** caravagn@di.unipi.it

Poster: Stochastic modeling – SM3

**Stochastic pattern formation for reaction-diffusion systems on networks**

Di Patti F(1), Fanelli D(1)

*(1) Dipartimento di Fisica e Astronomia, Università degli Studi di Firenze and INFN, Firenze*

**Motivation:** The process of pattern formation in reaction-diffusion systems is a fascinating field of investigation with applications ranging from physics to chemistry and biology. Despite patterns are classically studied in the mean-field deterministic limit, recently a growing interest has focused on the individual-based stochastic models which result in finite-size corrections to the mean-field dynamics. It has been shown that under specific conditions, microscopic fluctuations are enhanced by a resonance mechanism and yield organized spatiotemporal patterns. More specifically, the measured concentration that reflects the distribution of the interacting entities (e.g., chemical species and biomolecules) can display spatially patched profiles, collective phenomena which testify on a surprising degree of macroscopic order, as mediated by the stochastic component of the dynamics. Moreover, in the last years, the study of pattern formation has been extended to the networks, a relevant direction of investigation due to the similarity of network topologies to some cellular environments. The aim of this work is to describe the dynamics of particles free to diffuse on a network and to characterize analytically the emergence of stochastic patterns.
**Methods:** First of all we will introduce a stochastic formulation of a general reaction-diffusion system on a network. Then, we will proceed with an analytical technique which will allow us to write down the analytical power spectrum: a localized peak in the shape of the power spectrum signals the presence of patterns.
**Results:** Through the study of the power spectrum of the fluctuations we will point out the existence of stochastic patterns, seeded by inherent stochasticity, outside the region predicted by the deterministic analysis. For a specific model, the theoretical predictions will be successfully tested though numerical simulations obtained implementing the Gillespie algorithm.

Poster: Stochastic modeling – SM4

**A tool for the stochastic modelling of non-Markovian biochemical processes**

Chiarugi D(1), Falaschi M(2), Hermith D(2), Olarte C(3), Torella L(2)

*(1) Istituto di Scienza e Tecnologie dell'Informazione "A . Faedo" - CNR, Pisa (2) Dipartimento di Ingegneria dell'Informazione e Scienze matematiche - Università di Siena (3) Dep. de Electronica y Ciencias de la Computacion, Univ. Javeriana. Cali, Colombia*

**Motivation:** The improvements of wet-lab techniques is making available novel experimental data about biochemical reactions occurring in living cells. This information provides evidences that random events arising at the molecular level play important roles in determining the overall behaviour of biological organisms. As a consequence, new interest is rising towards stochastic computational models of biochemical systems as useful tools for bridging the gap between the experimental observations and the complex dynamics typical of living entities. In the 1970s Gillespie provided a seminal contribution in this field, formulating the Chemical Master Equation (CME) and providing a very popular Stochastic Simulation Algorithm (SSA) for generating trajectories of the Markov process described analytically by the CME. This SSA embeds a Monte Carlo technique, the Inverse Sampling (IS), for generating random numbers. In fact, the CME is a set of differential equations describing a Continuous time Markov Chain (CTMC). In CTMC models of biochemical systems, states are defined by the number of molecules of each reacting species present at a given instant. Each of the allowed state transitions correspond to the occurrence of a chemical reaction. The system evolves from a state to another according to given probability density functions (PDFs) which rule both transition probabilities and the timing of the process. Importantly, CTMC are characterised by the so-called Markov property: informally, the probability of transition towards the possible future states of the process depend only on the current state. It can be proven that if the Markov property holds, the random variable describing the time between two subsequent events (waiting time) must be distributed according to a negative exponential PDF. In other words, the Markov property do not hold when an observed process shows up as a series of random events occurring with non exponentially distributed waiting times. These kind of processes is sometimes referred to as non-Markovian and, often, cannot be described properly through a CTMC . Many observed biological processes turns out to occur as stochastic processes with non-exponential waiting times. Some examples are the single-enzyme catalysis or the mRNA degradation process. Computational simulations of these non-Markovian processes are often performed using the Gillespie's SSA and, hence, approximating them as CTMCs. The effects of this approximation on the simulation results are difficult to evaluate and, thus, alternative SSA turns out to be desiderable.

**Methods:** We present a software tool which grounds on a SSA which allows to simulate the evolution of biochemical systems as discrete-state continuous-time stochastic processes with arbitrary waiting times. In its essence our method can be seen as an enhancement of the original SSA, which overcomes some shortcomings of the IS. The IS in the Gillespie's SSA is used for generating random waiting times. The crucial step of the IS is the solution of the equation $F(\tau) = r$, where $F(\tau)$ is the cumulative function of the PDF $f(\tau)$ for waiting times and r is a random number generated from the uniform distribution. In order to compute a solution for $F(\tau) = r$ we use the XRI library, an implementation of a real interval constraint system. The XRI library implements efficient techniques such as Hull, Box and kB-Consistency for solving sets of numerical constraints. The advantage of this technique with respect to classical numerical methods is that: 1) no initial parameters for iteration are required; 2)

interval arithmetic allows us to bound numeric errors since real numbers are represented by means of intervals instead of single floating-point numbers; and 3) constraints represent relations on the variables rather than assignments to values. Hence, by using the XRI on top of the Mozart system, we can impose the constraint $F(\tau) = r$ and we are able to both determining r given a $\tau$ and computing $\tau$ given some random number r. This is a very general method that can be applied to all continuous PDFs. Indeed, we can easily obtain correct random samples from both distributions describing experimental data and general PDFs such as Erlang, Gamma and Hyperexponential.

**Results:** In this work we extend an earlier proposal, presenting a working software for the computational simulation of biochemical systems as stochastic processes with waiting times which are not necessary exponentially distributed. In this way we provide a tool which allows us to simulate biochemical phenomena without the need of approximating them as CTMCs as in the Gillespie's SSA. Moreover, we exploit our tool addressing some case-studies. We then compare the outcomes of these simulations with analogous results obtained through the Gillespie's SSA to show the impact of the CTMC approximation on the obtained dynamics.

Poster: Stochastic modeling – SM5

**Novel approximation methods for stochastic biochemical kinetics**

Grima R (1), Thomas P (2)

*Department of Biological Sciences, University of Edinburgh, United Kingdom*

**Motivation:** It is well known that for any chemical system composed of at most first-order reactions, one can obtain the full stochastic properties of the system via an exact solution of the moment equations for the chemical master equation. In contrast for systems composed of bimolecular reactions such an exact solution is not generally possible. Given that most biochemical systems involve such interactions, there is the need of systematic approximation methods to probe the intrinsic noise statistics of realistic biochemical networks.
**Methods:** In this talk I will present the Effective Mesoscopic Rate Equation (EMRE) formalism which provides accurate approximations to the mean concentrations predicted by the chemical master equation for systems with molecule numbers as small as the order of ten molecules.
**Results:** These equations correct the conventional rate equations and predict a wide variety of phenomena including amplification of substrate concentrations in metabolic networks and concentration inversions in trimerization and genetic networks. I will also discuss an extension of these ideas to predict noise-induced oscillations in circadian oscillators and a comparison of the EMRE based approaches with conventional moment-closure methods. Finally I'll showcase iNA (intrinsic noise analyzer), our new open-source software with a user-friendly graphical interface, which uses the aforementioned approximation methods to calculate the intrinsic noise statistics of a biochemical network specified by an SBML file.
**Contact email:** ramon.grima@ed.ac.uk

**Models of anomalous diffusion for explaining the "super-concentration" effect**

Allegrini P(1), Chiarugi D(2), Paradisi P(2).

*(1) Istituto di Fisiologia Clinica (2) Istituto di Scienza e Tecnologie dell'Informazione "A. Faedo" - CNR, PISA (Italy)*

**Motivation:** A series of recent experiments aimed at constructing synthetic minimal living cells, revealed possible deviations from the number of macromolecules expected to be entrapped inside liposomes. In one kind of these experiments, e.g., cryo-TEM tmicroscopy was used to study the entrapment of the protein ferritin into liposomes. In particular, liposomes were left free to spontaneously form into a ferritin-containing solution, for obtaining lipid vesicles which were expected to contain a certain number of ferritin molecules. Surprisingly, the observed distribution of ferritin molecules entrapped into the vesicles, strongly deviated from the expected Poisson distribution. Instead, a power-law distribution appeared, characterized by many “empty” vesicles (no or very little entrapped solute), and a long decreasing tail with extremely crowded vesicles. Indeed, the cryo-TEM analysis showed some extremely crowded liposomes adjacent to empty ones. These observations has been further confirmed in a series of similar experiments and the observed “anomalous” behaviour has been called “super-concentration”. Due to the possible important implications on the study of the origin of life, the “super-concentration effect” is currently intensively studied. Even though some hypothesis has been formulated for describing the possible causes underlying this phenomenon, up to the authors knowledge no exhaustive explanation has been provided yet. In this work we aim at addressing this issue proposing a model in which the crowding of molecules into water solution turns out to arise spontaneously if anomalous diffusion of molecules is assumed.
**Methods:** Anomalous diffusion has been observed in living cells and is thought to depend on the heterogeneity of the environment. According to our hypothesis, similar conditions occur also in the protein and lipid solutions and drive the observed super-concentration. We investigate this issue through a modelling approach based on Lévy stable densities, such as Lévy flights, Lévy Walks or Continuous Time Random Walks with fractal waiting time distributions and/or self-similar cluster in the jump probability. In any case, we assume that: (i) liposomes and protein molecules moves into the solution independently from each other (ii) liposomes are driven by a standard Brownian motion, while (iii) the protein molecules are supposed to move according to a random walk model giving rise to anomalous diffusion.
**Results:** Through our approach we provide possible explanations of the “super-concentration” effect. In particular we show that, if some models of anomalous diffusion are assumed, the crowding of the diffusing particles may arise spontaneously and the distribution of the clusters dimensions follow a power law. Our findings suggest possible directions for further experimental efforts aiming at verifying in which conditions anomalous diffusion shows up.

Poster: Synthetic and systems biology – SSB2

**Synthetic Biology: A detailed analysis of size and solute distributions of liposomes reveals intriguing features in synthetic cell models.**

Fanti A. (1)*, Gammuto L. (1)*, Mavelli F. (2), Luisi PL. (3), Stano P. (3)^, Marangoni R. (1, 4)^

*(1) Department of Biology, University of Pisa (2) Department of Chemistry, University of Bari (3) Department of Sciences, University of Roma Tre, Roma (4) CNR - Biophysics Institute, Pisa. * joint first-Author ^ joint last-Author*

**Motivation:** Anomalous solutes entrapment phenomena have been detected when liposomes are generated at nanoscale range (volume ca. 10exp(-18) L), revealing that a power-law distribution adequately describes the amount of encapsulated solutes, rather than the expected Poisson law. This causes the "supercrowding effect": the existence of experimentally detectable liposomes with a very high inner solute concentration. This phenomenon plays a crucial role in origin of life scenarios, being probably at the roots of the origin of early cells in primitive diluted solutions. Here we extend the investigation on solute concentration inside small vesicles when these are generated from larger (giant) vesicles (GVs, size > 1 ?m). Solute-filled GVs can be mechanically divided in order to generate solute-filled smaller vesicles, mimicking, in a very simple and rough way, early cell division. By combining theoretical and experimental approaches, we asked whether anomalous size/solute distributions occur, by testing the redistribution of solutes when combined with vesicle division.
**Methods:** From a theoretical point of view, we simulated both the GVs population generation and their division into submicron vesicles. The size distribution is generated accordingly with the standard theory. The solutes have been assigned to vesicles following different stochastic partition models. From the experimental point of view, we first enclose different solutes (low MW: pyranine, and calceine; high MW: dextranes and bovine serum albumine, bound to fluorescent markers), inside GVs, describing the solutes entrapment distribution inside these vesicles. Then, we produce a derived vesicles population via extrusion, thus obtaining a population of submicron vesicles (SUVs, radius ca. 0.4 ?m) from the original GVs population. Confocal microscopy, dynamic scattering and cytofluorimetric measurements are employed to assess the vesicle size distribution and the solute concentration in both populations.
**Results:** While the size distribution of GVs populations is found to be in accord with the theoretical forecasted ones, the size distribution of SUV populations show a deviation from the expected, due to the presence of larger vesicles, coexisting with the vesicles of expected size. This suggests that during extrusion specific membrane reorganization processes occur. By comparing the solutes distribution in GVs and in extruded vesicles, significant differences are observed. In particular, a correlation between GVs size and solute content is evident. Extruded vesicles, on the other hand, show a size-independent mean solute content, but stochastic fluctuations increase when vesicle volumes decrease. Intriguingly, tailed solute distributions are often observed. This study has been repeated for each tested solutes, and experimental results have been compared with stochastic predictions. Results are tentatively discussed in terms of different chemical-physical properties of the solutes tested and their interaction with the membrane during vesicle extrusion. The relevance of this investigation in the context of origin of life and synthetic biology is shortly commented.

**Studying the effects of cell-to-cell variability on the quantitative behaviour of synthetic biological networks through stochastic simulations**

Politi N(1,2), Pasotti L(1,2), Zucca S(1,2), Magni P(1,2)

*(1) Laboratorio di Bioinformatica e Biologia Sintetica, Dipartimento di Ingegneria Industriale e dell'Informazione, Università degli Studi di Pavia (2) Centro di Ingegneria Tissutale, Pavia*

**Motivation:** Noise is considered as a key feature in biological systems and is currently under investigation by different research groups. The contribution of noise was studied both theoretically and via experimental data in gene expression systems, showing that it can explain apparently unpredictable circuit behaviours and can also propagate through interconnected networks. Variability could result from two sources of noise: intrinsic noise, which is the inherent variability in the expression of a gene of interest, and extrinsic noise, which, conversely, is linked to the fluctuations affecting cellular resources such as amount of polymerase, ribosomes, etc. Considering these features, a study on the effect of such variability sources affecting genetic circuits has been carried out: stochastic simulations can elucidate the predicted effects as a function of circuit architecture, as well as provide a sensitivity analysis tool to find the conditions in which noise least affects the network outcome.
**Methods:** Steady-state mathematical models of inducible systems were considered, assuming that protein production follows a Hill equation which is function of the specific input concentration. The systems considered in this work receive exogenous signals or regulator proteins as inputs; it was assumed that proteins are the only species affected by Gaussian additive or Lognormal multiplicative noise. Samples were extracted from distributions with a constant standard deviation or coefficient of variation. Intrinsic and extrinsic noise were tuned by changing the correlation level between two or more extracted samples. Two different topologies for interconnection were studied: feed-forward and feed-back.
**Results:** The genetic circuits considered in this work are widely used to characterize biological noise. However, here we focus on different aspects compared to already published works. In particular, we test if modularity of parts is affected by the stochastic nature of biological components by simulating the output of feed-forward circuits driven by several input devices which exhibit different cell-to-cell variability. Moreover, close-loop networks are considered: their feed-back regulation could be driven by an endogenous signal (e.g., an intracellular transcription factor) or an exogenous one (e.g., a signalling molecule). We simulate the output of such genetic networks and compare the two described options for the feed-back signal. The described work can be a complement of experimental studies on model systems to understand if unexpected behaviours could be explained by noise, for example the apparent non-modularity resulting by the interconnection of two quantitatively characterized devices which yield an unpredictable output. Finally, the work can also support the design of feed-forward or feed-back circuits to choose robust control signals and systems topology.

**Modeling the effects of promoter and regulator copy number variation on the output of inducible systems in engineered genetic circuits.**

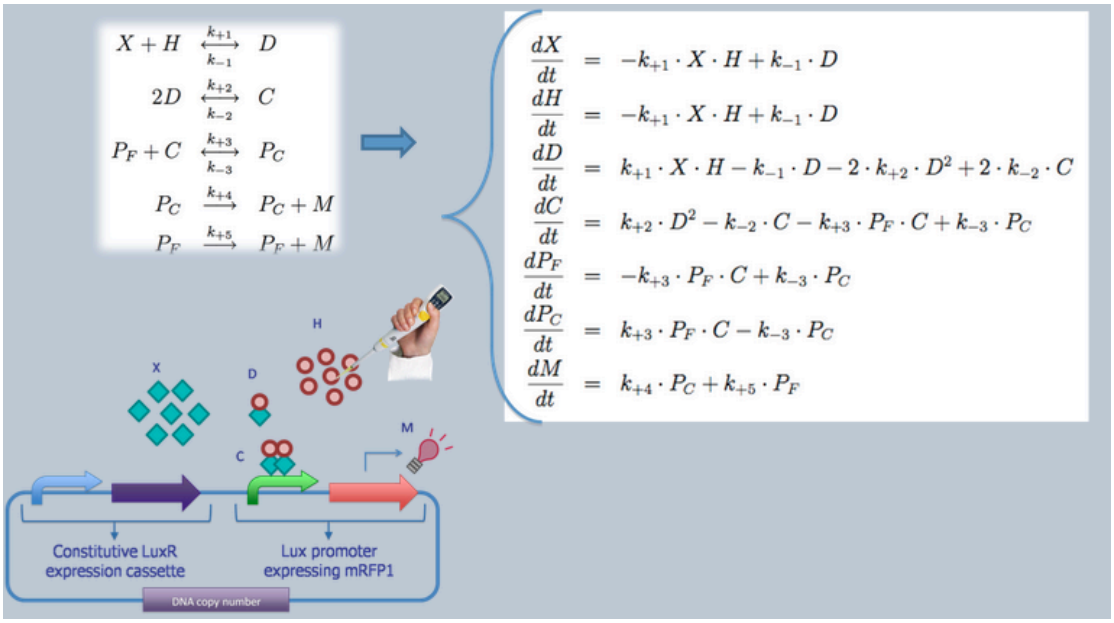
Zucca S (1,2) Pasotti L (1,2) Politi N (1,2) Casanova M (1,2) Cusella De Angelis MG (2) Magni P (1,2)

*(1) Laboratorio di Bioinformatica e Biologia Sintetica, Dipartimento di Ingnegneria Industriale e dell'Informazione, Università degli Studi di Pavia (2) Centro di Ingegneria Tissutale, Università degli Studi di Pavia*

**Motivation:** DNA copy number (CN) variation is a key element which strongly affects the outcome of a gene network in synthetic biology. A deep analysis on CN related phenomena is essential to exploit this parameter as a genetic knob for the fine regulation of the network behavior, with the aid of both experimental data and mathematical modelling. Most frequently used mathematical models rely on the use of empirical equations that describe biological phenomena (e.g., Hill-like equations to describe the activation of an inducible transcriptional promoter), with the limitation of the lack of biological significance of the parameters. Mechanistic models overcome this issue and permit to deepen the knowledge about every single reaction. Moreover, they enable to explicitly introduce CN as a structural parameter. Here, a simple gene network composed by the Lux promoter, which can be activated in presence of the LuxR transcription factor (constitutively expressed) and the exogenous quorum sensing molecule 3-oxoexanohyl-homoserine lactone (HSL) has been studied with the aid of an ad-hoc defined mechanistic model (Fig. 1), based on the law of mass-action kinetics. Given the complexity of the network, the effects of CN variation of the Lux promoter and the LuxR transcription factor, individually and in concert have been numerically studied. Sensitivity analyses have also been performed on model parameters and a sub-portion of the model has been analytically studied.
**Methods:** All the simulations were performed via Matlab R2009b (MathWorks), using the 'ode23s' routine for the numerical system solving. Parameter values were retrieved from the literature. The steady-state input-output characteristic has been computed by simulating the model behaviour at the steady state in response to different HSL concentrations. Three indicators have been defined to describe the shape of this characteristic: Pmax is the maximum output value, H50 is the induction required to obtain an output value equal to 50% of Pmax and H10 and H90 are the HSL values for which the output is 10% and 90% of Pmax, respectively. Finally, the relative dispersion measurement $H^*=(H90-H10)/H50$ was computed. Sensitivity analysis on model parameters was performed by varying them individually 1 to 3 orders of magnitude from the nominal value. For each parameter variation, the corresponding induction curve was computed.
**Results:** First, an analytical study of a small portion of the model has been performed, in order to determine the analytical relations among the model parameters, species and the system outputs. An analytical dissertation of the full model is not feasible, thus numerical methods need to be applied. The system response to CN variation has been computed and numerical results were compared with experimental data. The model previsions were qualitatively in accordance with experimental data, obtained in our lab in engineered Escherichia coli containing the described network, despite the absolute predicted quantities were deviated from the experimental values of about one order of magnitude. Since the used parameter values have been estimated via in vitro experiments, they may not accurately describe in vivo occurring processes: to overcome this issue, sensitivity analysis on model parameters has been implemented, to validate the robustness of the system

output. The parameters of major impact on system quantitative behavior have been individuated. Further computational and experimental efforts are required to fill the gap between the biological knowledge of the system and the measurement of kinetic rates regulating the process.

**Contact email:** susanna.zucca@unipv.it



Poster: Synthetic and systems biology – SSB5

**Differential connectivity in neoplastic co-expression networks**

Creanza TM(1,2), Anglani R(1), Liuzzi VC(3), Piepoli A(4), Panza A(4), Andriulli A(4), Ancona N(1).

*(1) Institute of Intelligent Systems for Automation, National Research Council (CNR-ISSIA), Bari, Italy (2) Center for Complex Systems in Molecular Biology and Medicine, University of Torino, Torino, Italy (3) Department of Bioscience, Biotechnology and Biopharmaceutical, University of Bari, Italy (4) Department of Medical Sciences, Division and Laboratory of Gastroenterology, IRCCS Casa Sollievo della Sofferenza, San Giovanni Rotondo, Italy*

**Motivation:** Cancer-causing alterations in coding regions and post-translational modifications can change the protein activity without affecting the coding gene expression level, although altering its interaction scheme with other genes. Hence, an analysis merely based on differential expression measures could be ineffective for revealing key drivers in neoplastic lesions. Moreover, a differential expression approach does not take into account the interactions among genes, while it is widely accepted that the understanding of the molecular mechanisms underlying a complex disease like cancer requires the investigation on the interplay between genes. These considerations motivate a study of the rewiring of gene interaction networks as a consequence of the aforementioned alterations in order to complement the informative content of the differential expression.
**Methods:** We analyzed five different cancer types in order to study the structural properties of cancer networks spotting common crucial traits among the different tumor tissues. Moreover, we presented a differential connectivity approach that searches for changes in gene interactions that characterize the cancer tissues with respect to the healthy controls by comparing the condition-specific inferred co-expression networks. To this aim, we investigated on the single genes with a connection structure strongly modified between the two biological phenotypes. Genes with degree variations significantly associated to pathology were detected by adopting non-parametric permutation tests. Finally, we analyzed the relative merits of an enrichment pathway analysis that considers both differential expression and differential connectivity measures with respect to the classic enrichment studies in the identification of known cancer programs.
**Results:** The loss of connectivity in cancer tissues turns out to be a general trait of the different kinds of tumors. Next, our approach based on differential connectivity fingers toward known cancer-driver genes not revealed by differential expression measures. Moreover, the over-representation of known cancer genes in our findings allow to make reasonable hypotheses that top-ranked genes represent putative tumor biomarkers not yet recognized. Finally, our integrated pathway analysis enlightens novel cancer-related drivers shifting the focus on mechanisms of carcinogenesis and tumor progression not still properly investigated.
**Contact email:** ancona@ba.issia.cnr.it

**Modelling the gene regulatory network underlying early T-cell development**

Cacace E (1,2), Collombet S (2), Thomas-Chollier M (2), Thieffry D (2)

*(1) Scuola Superiore Sant'Anna, Pisa, Italy (2) Ecole Normale Supérieure - Département de Biologie IBENS - UMR ENS - CNRS 8197 - INSERM 1024 46 rue d'Ulm, 75230 Paris Cedex 05, FRANCE*

**Motivation:** The hallmark of T-cell development is the sequential exclusion of alternative differentiative potentials: it is only after suppressing B cell, myeloid, dendritic cell and NK lineage pathways that a T-cell attains its definite commitment. Since many of the factors needed for T-cell specification are required in other hematopoietic differentiation pathways, the ultimate specification of the T-cell fate is most likely the result of fine-tuned combinations of these factors that are activated in a time and environment-dependent fashion. To investigate the interplay among the diverse differentiation programmes, we dynamically modelled a comprehensive and integrative network of the factors involved.
**Methods:** We mapped the gene regulatory network that underpins T-cells differentiation until the DN4 stage, integrating data from literature and ChIP-seq data from publicly available datasets. We then used the software GINsim (Gene Interaction Network simulation: http://ginsim.org/) to evaluate the influence of different combinations of transcription factors at crucial stages of T cell development, applying logical formalism to model the interactions among the network components.
**Results:** We focused on two major checkpoints of T-cell specification: (i) the DN2a-DN2b transition, when a first commitment occurs due to the Bcl11b-mediated downregulation of stem cell-associated genes, marking the beginning of TCR gene rearrangement, and (ii) the beta-selection step, where a functional and signalling pre-TCR permits the transition to the DN4 stage. In addition, we refined and enriched a previous model of the lymphomyeloid switch that governs the choice between these two alternative fates and is centered on the interactions between the targets of the transcription factor Pu.1 and the Notch signalling pathway. By applying logical formalism to these key steps of T-cell development, we were able to test the influence of different combinations of transcription factors at the branching points among different lineage programmes, thereby shedding light on T-cell early plasticity and resiliency.
**Contact email:** e.cacace@sssup.it

**Integrating protein-protein and signaling information to model and simulate signed oriented biological networks.**

Calderone A (1) Cesareni G (1,2)

*(1) Department of Biology, University of Rome Tor Vergata, Rome, Italy. (2) Fondazione Santa Lucia Istituto di Ricovero e Cura a Carattere Scientifico (IRCCS), Rome, Italy.*

**Motivation:** Discovering pathways is a key step to understand biological processes. Despite the large quantity of information available, scientists still need to search various databases in order to reconstruct and understand the steps that lead to known processes and to understand possible new interactions. The amount of interaction data has now grown to such an extend that it has become hard to filter out relevant information. Bioinformatics approaches can offer tools and strategies to manipulate such information automatically.
**Methods:** In the past years, the establishment of a common curation policy for protein-protein interaction (PPI) information allowed the possibility of integrating protein interaction in a more accurate way, simplifying the process of merging data derived from different sources. In one of our previous works, mentha[1], we have implemented a workbench to build custom PPI networks. Its procedure creates a large interactome containing PPI information extracted from five different databases and it assigned to each interaction a score that takes into account the number of papers and the kind of experiments conducted on each interaction, thus proportional to reliability of the interactions. Using mentha and algorithms to extract relevant interconnections among proteins of interest, it is possible to extract sub-networks from the global interactome. An orientation algorithm makes it possible to predict the signal direction[2], for instance, from receptors toward transcription factors. Activation and inhibition can be predicted using database containing phenotypes from RNA interference (RNAi) screens[3]. Furthermore, in order to complement the information extracted from PPI and phenotypes databases, we are developing SIGNOR, a database that contains curated signaling information. Having signed oriented biological networks will allow us to write transition functions for each node and, using Boolean formalisms, we can create Discrete Boolean Networks to simulate the behavior of the reconstructed network. Through simulation, it is possible to discover steady states and loops and eventually to predict the effect of perturbations of specific node.
**Results:** Sub-networks extraction has produced promising results with known pathways, the sub-network extracted contains a high number of known interactions and a part of "unknown" steps that can possibly hide relevant interconnections. The scoring function can approximate relevant interactions in known pathways and thus reveal new relevant interactions. The automatic procedure that computes Boolean transition functions makes the predicted network ready for simulations. Automatically generated Boolean expressions seem to approximate the natural behavior of the predicted network and possibly reveal new dynamics.
**Supplementary information:**References 1. Calderone A, Castagnoli L, Cesareni G (2013) mentha: a resource for browsing integrated protein-interaction networks. Nat Methods 10: 690–691. doi:10.1038/nmeth.2561. 2. Gitter A, Klein-Seetharaman J, Gupta A, Bar-Joseph Z (2011) Discovering pathways by orienting edges in protein interaction networks. Nucleic Acids Res 39: e22–e22. doi:10.1093/nar/gkq1207. 3. Vinayagam A, Zirin J, Roesel C, Hu Y, Yilmazel B, et al. (2014) Integrating protein-protein interaction networks with phenotypes reveals signs of interactions. Nat Methods 11: 94–99. doi:10.1038/nmeth.2733.
**contact email:** sinnefa@gmail.com

**Functional and network analysis of selenoproteins modulated in hepatocellular carcinoma**

Rusolo F(1), Guariniello S(2), Colonna G(2), Della Sala M(1), Arena A(1), Castello G(1), Costantini S(2)

*(1) Istituto Nazionale Per Lo Studio E La Cura Dei Tumori "Fondazione Giovanni Pascale", IRCCS, Italy (2) Biochemistry, Biophysics and General Pathology Department, Second University of Naples, Naples, Italy*

**Motivation:** Selenium (Se), an essential trace element for mammals, can assist cells to resist oxidative damage. In vivo, Se is primarily present as selenoproteins to maintain the balance of the cellular redox state. In particular, 25 selenoproteins have been found in humans and most of them play important roles in redox regulation, detoxification, immune-system protection and viral suppression. However, the biological functions of some newly identified selenoproteins remain unknown. Recently in our laboratory we have carried out the analysis of the global gene expression of hepatoma cell line compared to normal human hepatocytes by microarray studies focusing the attention on the selenotranscriptome family to identify what selenoprotein(s) might be used as marker(s) for HCC prognosis (or diagnosis) and/or treatment [1-2]. Moreover, we tested in vitro the effect of sodium selenite on two human hepatoma cell lines to assess its effect on the expression of selenoproteins and showed that in treated cells four selenoproteins, i.e. GPX1, SELS, SELK and SELM, increased their expression whereas DIO1 remained unchanged. Therefore, functional analysis and interactomics studies were performed on these selenoproteins to evaluate if they are correlated between them and in what metabolic pathways are involved.
**Methods:** Network studies were performed by Ingenuity Pathway Analysis (IPA) program whereas functional and pathway analysis by Panther and David tools. Moreover, Interactome3D web service was used for the structural annotation of protein-protein interaction networks and for visualizing and downloading structural information for interactions involving a set of proteins or interactomes.
**Results:** Interactomic studies performed by IPA program have evidenced that these selenoproteins are involved in two networks of which one presents DIO1 as hub node correlated with other 15 proteins and the other composed by GPX1, SELS, SELK and SECIS binding protein 2 that is SECISBP2 already known to be markers of hepatocellular carcinoma. The pathway analysis of selenoproteins has evidenced that GPX1 and DIO1 can be categorized in different metabolisms whereas the other three selenoproteins are less annotated and, hence, further functional studies are necessary. In particular, Panther analysis evidenced that: i) GPX1 is involved in lipid metabolic process, oxidoreductase, immune system process , antioxidant activity, catalytic activity and hydrolase, and ii) DIO1 in catalytic activity and hydrolase activity. Moreover, to visualize and download structural information for interactions involving DIO1, GPX1, SELS, SELK and SELM we used Interactome3D web service. It showed that SELS correlates with SRPK (serine/arginine protein kinase), SELK with CLN8 (Ceroid-lipofuscinosis, neuronal 8) and GPX1 with MAPK6 (mitogen-activated protein kinase 6). Moreover, this analysis suggests also information about possible templates that we can use to perform structural analysis. Therefore, in future we will perform studies related to structure and function relationships for these proteins. Some results were already obtained for SELM.
**Contact email:** susan.costantini@unina2.it
**Supplementary information:**References [1] R. Raucci, G. Colonna, E. Guerriero, F. Capone, M. Accardo, G. Castello and S. Costantini. Structural and functional studies of the human

selenium binding protein-1 and its involvement in hepatocellular carcinoma. Biochim Biophys Acta: Proteins and Proteomics (2011) 814(4): 513-22 [2] S. Costantini, G. Di Bernardo, M. Cammarota, G. Castello and G. Colonna. Gene Expression segnature of human hepatoma cells". Gene (2013) 518:335-345. [3] F. Rusolo, B. Pucci, G. Colonna, F. Capone, E. Guerriero, MR Milone, M. Nazzaro, MG Volpe, G. Bernardo, G. Castello, S. Costantini. Evaluation of selenite effects on selenoproteins and cytokinome in human hepatoma cell lines. Molecules (2013) 18:2549-62.

**Ensemble evolutionary flux balance analysis for metabolic network modeling**

Chiara Damiani(1,2), Dario Pescini(1,3), Riccardo Colombo(1,2), Sara Molinari(1,4), Marco Vanoni(1,4), Lilia Alberghina(1,4), Giancarlo Mauri(1,2)

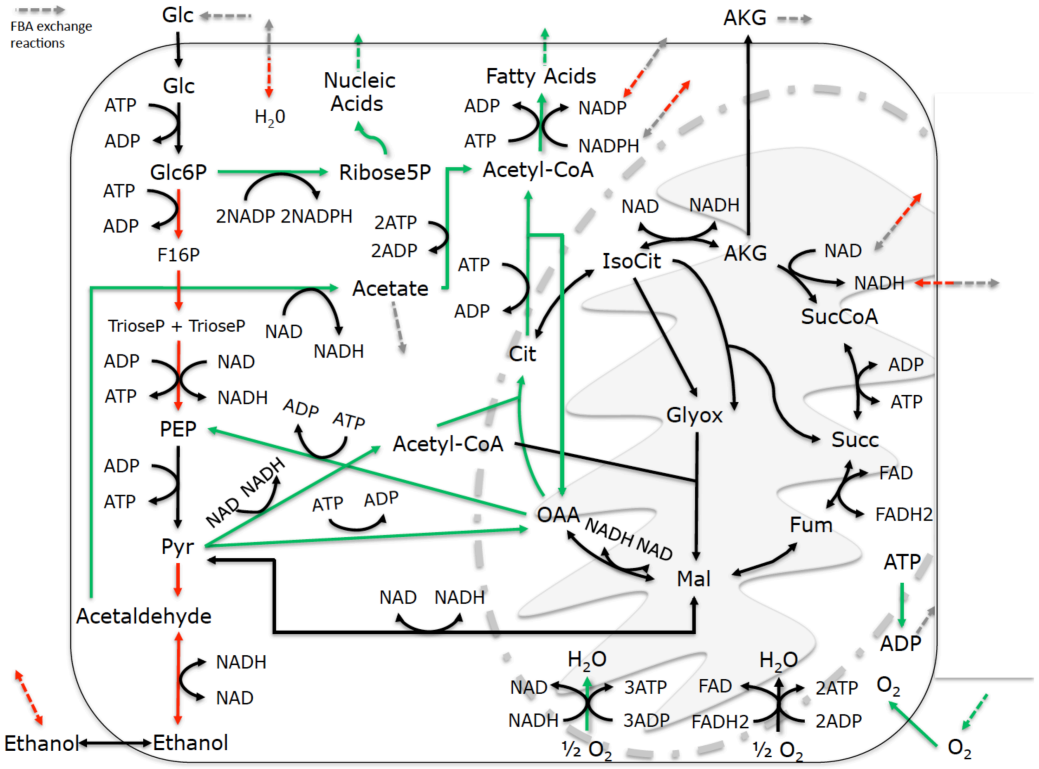
*(1) SYSBIO - Centre for Systems Biology, Piazza della Scienza 2,20126 Milano, Italy (2) Università degli Studi di Milano-Bicocca, Dipartimento di Informatica, Sistemistica e Comunicazione, Viale Sarca 336, 20126Milano, Italy (3) Università degli Studi di Milano-Bicocca, Dipartimento di Statistica e Metodi Quantitativi, Via Bicocca degli Arcimboldi 8, 20126 Milano, Italy (4) Università degli Studi di Milano-Bicocca, Dipartimento di Biotecnologie e Bioscienze, Piazza della Scienza 2-4, 20126 Milano,Italy*

**Motivation:** Metabolism plays a major role in cell functioning. The study of its control mechanisms and dysregulation is gaining increasing attention and it is nowadays an area of intense application of modeling efforts. As the complexity and dimension of metabolic reconstructions grow, mechanistic modeling of their dynamics becomes impracticable. For this reason, the computational investigation of genome-wide models makes typically use of a constraint-based approach, which exploits the knowledge about the structure of cell metabolism, while disregarding dynamic intracellular behavior, on the basis of a pseudo-steady state assumption. The imposition of additional constraints, such as irreversibility or capacity constraints, leads to a space of feasible flux distributions, each one representing a functional state. Assuming that cell behavior is optimal with respect to an objective, flux balance analysis (FBA) can be used to calculate a single optimal flux distribution. This approach has proven effective in implementing bioengineering design goals, such as the maximization of the cell production of a biochemical compound of industrial use. However, when one is interested in understanding which flux distributions are compatible with a given metabolic behavior, the defined objective must be as closer as possible to the physiological one. In this regard, the assumption of maximization of biomass yield as the objective function has revealed successful in predicting some phenotypical characteristics of microorganisms. Nevertheless, when dealing with multicellular organisms, the definition of a plausible objective function is not straightforward. For this reason, in this work we extend the classic constraint-based modeling approach to overcome the problem of defining an objective function.

**Methods:** The proposed approach neither focuses on a given solution nor analyzes the properties of the entire solution space; on the contrary it aims at analyzing the generic properties of a specific set of solutions - hence the name ensemble approach - which correspond to those solutions that match the physiological cell dynamics. The expected phenotype must be formally defined according to biological evidence. To this end, we propose to define it in terms of a metabolic response to environmental condition variations. For instance, we may want to observe a redistribution of fluxes as a consequence of a variation in nutrient availability (that can be simulated by increasing the boundaries that constrain the corresponding flux) that is compliant with experimental observations. The approach is still based on FBA, because several distinct objective functions are used in order to sample the solution space. As opposed to the dominant sampling algorithm of choice, the Markov chain Monte Carlo method, our sampling approach allows to maintain the network condition fixed (the objective), while varying the environmental conditions. The sampling method is then combined with a search algorithm in order to find solutions. As metabolic dynamics are described more accurately, the approach may benefit from the adoption of an evolutionary algorithm - hence the name evolutionary approach.

**Results:** We show how two distinct ensembles of solutions, one in agreement with the definition of Crabtree-positive yeasts and the other in agreement with the definition of Crabtree-negative yeasts, can be obtained. Then we show how the ensembles can be further characterized and refined by means of a cluster analysis. Finally we investigate which are the differentially expressed pathways in the two kinds of yeast.The fluxes that significantly differ between the two ensembles (according to a Kolmogorov-Smirnov test) are indicated in figure: red arrows indicate fluxes that are higher in the Cratbtree-positive ensemble as compared to the Cratbtree-negative one, green arrows indicate the opposite.

**Contact email:** chiara.damiani@unimib.it



**A mathematical perspective on cancer stem cell dynamics**

Fornari C(1,2), Beccuti1 M(1), Lanzardo S(2), Conti S(2), Balbo G(1,3), Cavallo F(2), Calogero R(2), Cordero F(1)

*(1) Department of Computer Science, University of Torino, Torino, Italy (2) Molecular Biotechnology Center, University of Torino, Torino, Italy (3) Faculty of Computing and Information Technology, Rabigh, King Abdulaziz University, Rabigh Branch, Kingdom of Saudi Arabia*

**Motivation:** The Cancer Stem Cell (CSC) involvement into tumor progression, tumor recurrence, and therapy resistance is one of the most studied subject in the current cancer research [1]. The CSC theory focuses on those tumors whose structure and behaviour are influenced by CSCs. Specifically, these cells drive both the tumor growth and evolution. Moreover, CSC-tumors are hierarchically structured and characterized by different subpopulations of cells - CSCs, Progenitor Cells (PCs), and Terminal Cells (TCs). This heterogeneity is also considered the cause of the failure of many conventional treatments. Therefore, the CSC theory is widely studied, being fundamental to fully understand the mechanisms underlying these type of tumors, and to better characterize their progression and treatments responses. However, due to the complex dynamics involved, a comprehensive theory has not been established yet. To this end, some advises can be obtained combining mathematical models and experimental data [2]. Modelling, indeed, is a powerful instrument which may drive the comprehension of cancer providing a clear description of its essential dynamics.
**Methods:** Mathematical models about cell populations dynamics [3, 4, 5] provide a useful tool to achieve such goal. In this work we describe cell population dynamics through a system of ordinary differential equations (ODEs), on which we perform both qualitative and quantitative analyses. Moreover, the model sensitivity - with respect to parameter values - is analyzed to identify key kinetics in tumor progression. Biological data are integrated in our model, making it suitable to verify hypotheses and to speculate about new theories. Specifically, data describing the tumor growth in mice and the population proportions over time are used to verify the correctness of the model fitting, constraining the ODE numerical integration.
**Results:** We propose a compartment model which describes the CSC-tumor progression at tissue level, exploring the CSC kinetics both in vitro and in vivo. Our model represents a new interesting way of investigate the tumorigenic power of CSCs, being the experimental results mathematically analyzed in order to extract the hidden knowledge underlying the available biological data. Specifically, we focus on the inter-dependencies among the different sub-populations of cells, trying to identify the essential mechanisms of cancer progression - at population level. Through a quantitative and qualitative analysis of our model is hence possible to define rules controlling the ErB2 positive breast cancer progression, and to characterize the three different scenarios of cancer evolution with respect to CSC dynamics. Moreover, the integration of experimental results in our model allows to better characterize the regulatory feedback involving the tumor cell growth. More generally, we combine mathematical techniques with biological results - arising from in vitro and in vivo experiments - to improve our understanding of the ErbB2+ breast cancer and to address new biological hypotheses. ## Bibliography ## [1] R. Pardal, M.F. Clarke, S.J. Morrison, Applying the principles of stem-cell biology to cancer. Nature Review Cancer, 3, 2003. [2] A.R.A. Anderson, V. Quaranta, Integrative mathematical oncology. Nature Reviews Cancer,

8, 2008. [3] C. Fornari, M. Beccuti, S. Lanzardo, L. Conti, G. Balbo, F. Cavallo, R.A. Calogero and F. Cordero, A mathematical-biological joint effort to investigate the tumor-initiating ability of cancer stem cells, PloS Computational Biology, under review. [4] F. Cordero, M. Beccuti, C. Fornari, S. Lanzardo, L. Conti, F. Cavallo, G. Balbo and R.A. Calogero, Multi-level model for the investigation of oncoantigen-driven vaccination effect, BMC Bioinformatics, 14, 2013. [5] X. Liu, S. Johnson, S. Liu, D. Kanojia, W. Yue, U.P. Singh, Q. Wang, Q. Nie, H. Chen, Nonlinar growth kinetics of breast cancer stem cells: implications for cancer stem cell targeted therapy, Scientific Reports, 2013.
**Contact email:** chiafornari@gmail.com

**Highly-Specific Transcribed-Ultra Conserved Regions in pluripotent stem cells**

Galasso M(1), Dama P(1), Previati M(1),Corrà F(1), Palatini J(1), Zerbinati C(1), Minotti L(1),Croce CM(2) & Volinia S(1,2)

*1 LTTA, Dept. of Morphology, Surgery and Experimental Medicine, Università degli Studi, Ferrara, Italy, 2 Comprehensive Cancer Center, Wexner Medical Center, and Biomedical Informatics, Ohio State University, Columbus, OHIO*

**Motivation:** There are 481 ultraconserved regions sequences (UCRs) longer than 200 bases in the genome of human, mouse and rat. These DNA sequences are absolutely conserved and show 100% identity with no insertions or deletions. Genome-wide profiling revealed that UCRs are frequently located on overlapping exons in genes involved in RNA processing and can be found in introns or at fragile sites and in cancer-associated genomic regions (CAGRs). Recently, transcribed-UCR were find differentially expressed in leukemia and solid tumors. The interest to understand the role of this class of non-coding RNA has been growing.
**Methods:** The Ohio State University Comprehensive Cancer Center custom microarray was used for T-UCR expression profiling following previously published protocols. GEO describes the OSU-CCC 4.0 platform under the accession number GPL14184. For each UCR two 40-mer probes were designed, one corresponding to the sense genomic sequence (named "+" or "plus") and the other to the complementary sequence (named "+A" or "minus"). Human Tissue samples were collected at ArrayExpress and respective accession numbers are: E-TABM-969 and E-TABM-970 for normal tissues. Data related to ESC and IPS cell lines were obtained by GSE16654, GSE12390 and GSE9440. Tissue samples were collected from mice. All the mice experiments for this study were approved by the Institutional Animal Care and Use Committee and University Laboratory Animal Resources of The OSU. Differentially expressed transcribed UltraConserved RNAs were identified by using the Class Comparison Analysis of BRB tools version 3.6.0 (http://linus.nci.nih.gov/BRB-ArrayTools. html). First, all samples were classified according to their organ-, tissue- and cell-type; then the normal samples were grouped in specific systems. To assess the specificity of UCR expression across groups, we needed to estimate what fraction of the total, for a given UCR belonged to each single group, the information content (IC) value. The specificity score, which varies between 0, when the expression level of the T-UCR m is the same across all tissues, and log2 (number of tissue types), when only one tissue expresses the UCR. Therefore, it is possible that we missed some specifically expressed T-UCRs that in our data had either very low expression or were specific to tissues/statement that we did not sample sufficiently. Banjo was used to infer the Bayesian network for the different classes of cells. For each class all the mature expressed and varying miRNAs were used as input to Banjo. The expression values were preprocessed with BRB array tools to only filter out nonvarying miRNAs. The static Bayesian network inference algorithm was run on the miRNA expression matrix by using standard parameters, with a discretization policy of q2. Consensus graphs, based on top 100 networks, were obtained from at least 8 x 109 searched networks. We applied the MCL graph-based clustering algorithm to extraction of clusters (i.e., groups of densely connected nodes) from gene networks. MCL (Neat) has been shown to enable good performances in extracting coregulated genes from transcriptome networks. yEd graph editor (yFiles software, Tubingen, Germany) was employed for graphs visualization.
**Results:** We tested the expression of UCRs in 374 normal samples from 46 different tissues, enabled us to perform a detailed analysis of UCRs expression. Tissue specific UCRs can

correctly differentiate cell types. For example, we identified brain, as well as epidermis specific T-UCRs. We then studied the expression of UCRs in human embryonic stem cells, induced pluripotent stem cells and a series of differentiated cell types (trophoblast, embryonic bodies, at 7 days and 14 days, definitive endoderm and spontaneous differentiated monolayer). One T-UCR in particular, uc.283 plus was highly specific for pluripotent stem cells and this finding was confirmed by Real Time PCR (RT-PCR). We hypothesized a possible network of the uc.283 plus in the gene pluripotent landscape.
**Contact email:** glsmrc@unife.it

**Design and analysis of a Petri Net model of the von Hippel-Lindau (VHL) tumor suppressor interaction network**

Minervini G(1), Panizzoni E(1), Giollo M(1,2), Masiero A(1), Ferrari C(2), Tosatto S(1).

*(1) Department of Biomedical Science, University of Padua, Padova (2) Information Engineering, University of Padua, Padova*

**Motivation:** Von Hippel-Lindau (VHL) syndrome is a hereditary condition predisposing to the development of numerous different cancer forms, related with the germline inactivation of the homonymous tumor suppressor pVHL. The best characterized function of pVHL is the ubiquitination dependent degradation of Hypoxia Inducible Factor (HIF-1α) via the proteasome. pVHL resulted involved also in several cellular pathways acting as molecular hub and interacting with more than 200 different proteins. Molecular details of its molecular plasticity remain in large part unknown. Here, we present a novel manually curated Petri Net (PN) model of the main pVHL functional pathways.
**Methods:** A Petri Net model was built using functional information derived from the literature with the final aim to reproduce a realistic description of the protein involved in VHL outcome. The model includes all major pVHL functions and is able to reproduce a credible description of VHL pathway at the molecular level.
**Results:** Interestingly, PN analysis suggests that variability of different VHL manifestations are correlated to concomitant inactivation of different metabolic pathways.
**Contact email:** giovanni.minervini@bio.unipd.it

**Simbio-System: a Statecharts based environment for biochemical pathways modeling and simulation**

Panagrosso M(1,2), Fioravanti F(3), Helmer-Citterich M(1), Nardelli E(2)

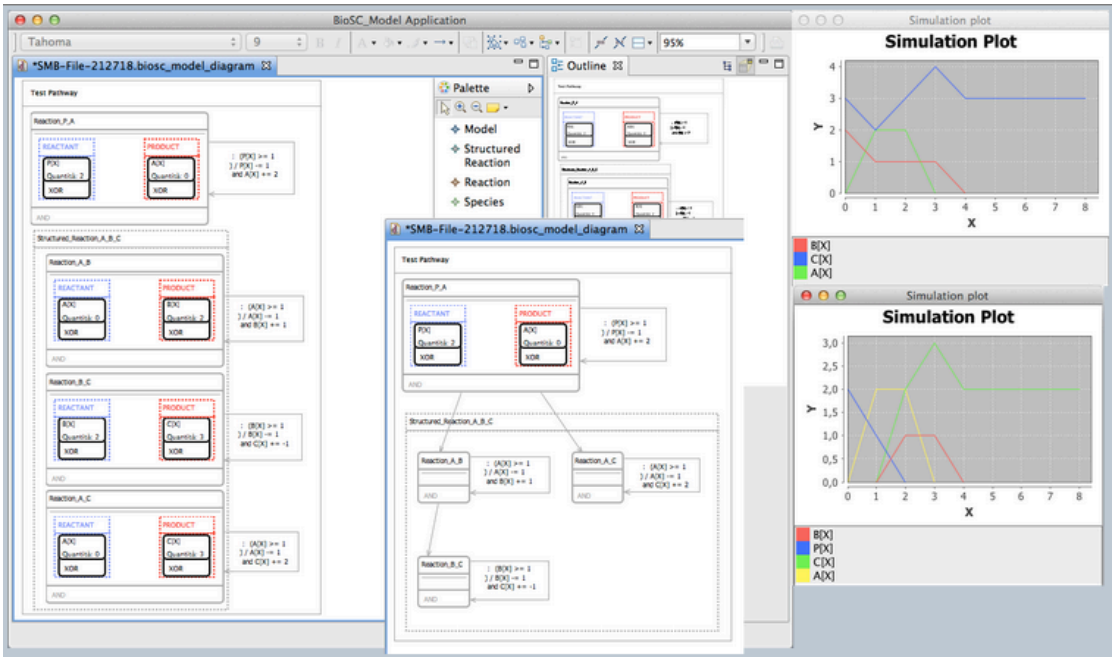
*(1) Department of Biology, University of Roma "Tor Vergata" (2) Department of Mathematics, University of Roma "Tor Vergata" (3) Department of Sciences, University of Chieti-Pescara "G. D'Annunzio"*

**Motivation:** Models used to represent biological systems have to be meaningful and natural to use for biologists. They also have to be based on formal semantics so as to make it possible to execute them and to test and analyze their behavior. Statecharts is widely used for modeling software systems and increasingly adopted for modeling biological system [1, 2]. In [3] we described how they could be used to automatically derive an executable representation of the SBML description [4] of a biological pathway. We now have implemented Simbio-System, a software environment for representing, editing, and simulating biochemical pathways, starting from their SBML descriptions.
**Methods:** The Reactome web site [5] is used as a source of SBML descriptions for Simbio-System. Pathways are visualized on the Reactome web site with their hierarchical structure, which unfortunately is not described in the downloaded SBML, even if it is highly useful to examine a pathway at the appropriate level of aggregation. Simbio-System therefore completes the SBML representation of a pathway with hierarchical data, deduced from the Reactome web site, and gives the user the possibility to visually edit the completed representation. Moreover, the user can interactively explore the representation at various aggregation levels. For example, she can collapse atomic (i.e., unstructured) reactions (hiding their constituent species), and iteratively collapse higher level (structured) reactions (hiding their component reactions), so as to consider the overall pathway with a coarser granularity. To support this interactive navigation at the various aggregation levels, Simbio-System automatically derives what we call precedence relations: these are arrows showing the precedence relations between reactions derived by the fact that species produced by a reaction are consumed by another reaction. When the user collapses some reactions, the system displays the precedence relations possibly existing between the collapsed reactions. The user can edit the visualized model in various ways. Once satisfied she can run a simulation of the execution of the pathway, after having possibly adjusted initial quantities for species, and visualize the results of the simulation. The simulation currently uses linear equations for all reaction kinetics. At each step of the simulation each reaction whose reactants and modifiers are present in sufficient quantities is executed, breaking by a random selection any possibly existing ties between two or more reactions needing the same scarce quantity of a reactant.
**Results:** In the figure (a composition of hardcopies of actual runs on a Mac) we show how Simbio-System represents a test pathway and simulates its behavior. It has four reactions in total. In the atomic reaction, 1 unit of species P produces 2 units of species A. The structured reaction is made up by three atomic reactions: in the first one 1 unit of A produces 1 unit of B, in the second one 1 unit of B consumes 1 unit of C, in the third one 1 unit of A produces 2 units of C. On the left we show the pathway at the level of its atomic reactions, while in the middle bottom we show a view of the pathway where the user has collapsed the atomic reactions in the structured reaction and the system has therefore visualized the precedence relations. The represented model shows the initial quantities (P=2, A=0, B=2, C=3) producing the temporal behavior appearing (only for selected species A, B, and C) on the top-right plot

in the figure. A second plot (bottom-right) shows the temporal behavior for all species when the simulation is run starting from a different set (P=2, A=B=C=0) of initial quantities. Both plots highlight the possibility of producing, even with a simple reaction kinetics based on linear equations, interesting temporal evolution of species. The top-right plot shows an oscillatory behavior of species C, while the bottom-right one shows how species C reaches a peak and then stabilizes.

**Contact email:** nardelli@mat.uniroma2.it

**Supplementary information:**[1] Hillel Kugler, Antti Larjo, David Harel, Biocharts: a visual formalism for complex biological systems, J.R.Soc.Interface 7(48):1015-1024, Jul.10, published online 18dec09. [2] Fabio Fioravanti, Manuela Helmer-Citterich, Enrico Nardelli: Modeling Gene Regulatory Network Motifs using Statecharts. BMC Bioinformatics, vol.13, suppl.4, S20, March 2012. [3] Fabio Fioravanti, Manuela Helmer-Citterich, Enrico Nardelli: A statechart based representation for SBML descriptions, Network Tools and Applications in Biology (NETTAB-10), Dec. 2010. [4] Hucka M., et al. "The systems biology markup language (SBML): A medium for representation and exchange of biochemical network models". Bioinformatics 19 (4): 524–531. 2003. http://www.sbml.org [5] Matthews L., et al. Reactome knowledgebase of human biological pathways and processes. Nucleic Acids Res., 37:D619-D622, 2009. http://www.reactome.org/



Poster: Synthetic and systems biology – SSB14

**Ranking of candidate microRNA-target pairs sensitive to disease and cell-line conditions**

Russo F(1,2), Baglioni M(3), Rizzo M(4), Berti G(4), Evangelista M(4), Geraci F(3), Rainaldi G(1), Pellegrini M(1)

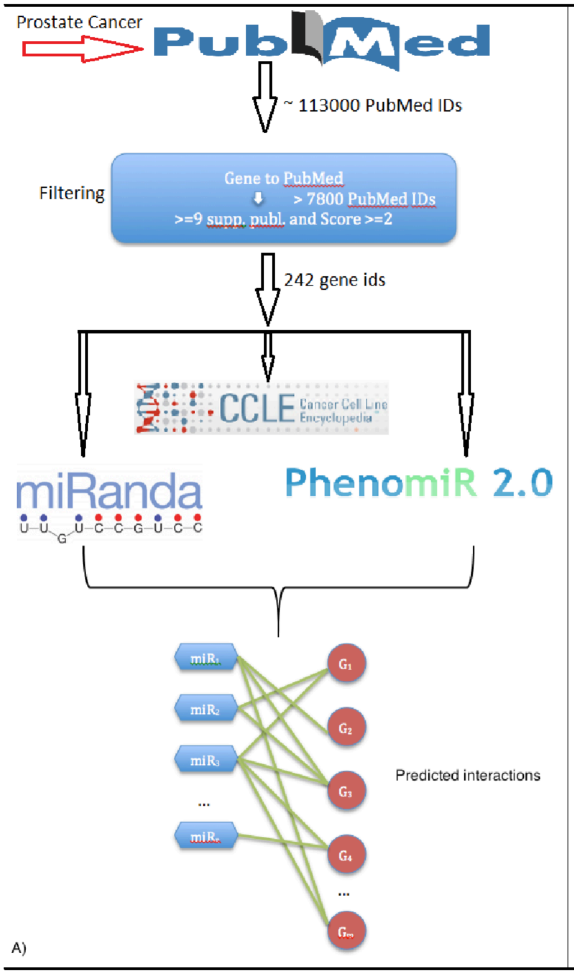
*(1)Laboratory of Integrative System Medicine (LISM), Institute of Informatics and Telematics (IIT) and Institute of Clinical Physiology (IFC), National Research Council (CNR), Pisa, Italy (2)Department of Informatics, University of Pisa, Pisa, Italy (3)Institute of Informatics and Telematics (IIT), National Research Council (CNR), Pisa, Italy (4)Laboratory of Molecular and Gene Therapy, Institute of Clinical Physiology (IFC), National Research Council (CNR), Pisa, Italy.*

**Motivation:** microRNAs (miRNAs) are post-transcriptional regulators that play important roles in cellular development, differentiation and diseases. They are often deregulated in many pathologies (e.g. cancer and cardiovascular diseases). One essential step to understand the regulatory effect of miRNAs is the reliable prediction of their target mRNAs. However, prediction tools by themselves are not enough, in that they are solely based on sequence information, and they do not take into account the specific cell line features and the gene expression. To better understand a pathology we need to collect the list of genes related to it, the miRNA-mRNA relationships, the specific cell line data and the gene expression. This work presents a general method applied to cancer, to accomplish this aim.
**Methods:** Our first step is retrieving the list of genes related to a disease by using the Entrez Programming Utilities, developed by the NCBI for searching publications from the MEDLINE. To do this we obtained the list of PubMed IDs containing gene information, then we filtered it by using the gene-to-publication link table provided by Entrez-Gene. We then applied the Hypergeometric Test (HT) on the number of publications, as done in Jourquin et al., to rank the retrieved genes and we applied the –log(HT) to obtain the gene score: the higher the score the more correlated the gene is to the pathology. Then, we compared the gene list with predicted miRNA targets and we selected those belonging to both the lists. Lastly, we incorporated the mRNA and miRNA expression from the Cancer Cell Line Encyclopedia (CCLE) and PhenomiR respectively. CCLE provides a detailed genetic characterization of a large panel of human cancer cell lines. For each gene in each cell line CCLE provides a single normalized expression value. To determine the regulation level of the gene we computed the median among the specific cancer cell line. The gene is declared to be up-regulated if its value is above the median value, and it is down-regulated otherwise. PhenomiR provides the miRNA expression on specific disease and cell line. Since miRNA generally repress their target mRNAs, a straightforward way to validate miRNA targeting mRNAs is detecting whether their expressions are inversely correlated. Based on this hypothesis, we retained only anti-correlated predicted miRNA-target interactions.
**Results:** We tested our approach on Prostate Cancer (see Figure 1.A). Among about 113000 papers related to this disease, about 7800 contained gene information and 3700 gene IDs (in Human). From this gene list we selected 242 genes with more than 9 supporting publications and a gene score greater than 2, as done by Jourquin et al. By using CCLE and PhenomiR data we selected a subset of miRNA-target interactions showing very high probability of binding based on target prediction tools. We have molecular and cellular data showing that miR-26a and miR-28 are both down-regulated in two prostate cancer tumor cell lines (DU-145 and PC-3) and that their over-expression inhibited cell proliferation. On the basis of these results we focused on these miRNAs and their targets. For DU-145 we found 11 genes while for PC-3 we found 8 genes, as shown in Figure 1.B. In particular we discovered some validated

miRNA-target interactions for specific cell line such as miR-26a-PTEN for DU-145 (Tian et al.), miR-26a-EZH2 for both cell lines (Koh et al.). Notably, no association between PTEN and miR-26a was found for PC-3 (it can be explained because generally PTEN is mutated in PC-3 cell line). Furthermore, the importance of considering the specific cell line is shown by the different subset of targets obtained for the two miRNAs. It is also to notice that while no relationship can be found in literature between miR-28 and Prostate Cancer genes, this approach was able to predict novel miRNA targets important for the disease. For this reason we are experimentally verifying the importance of miR-28 in Prostate Cancer.

**Contact email:** francesco.russo@iit.cnr.it

**Supplementary information:**Jourquin J et al. GLAD4U: deriving and prioritizing gene lists from PubMed literature. BMC Genomics. 2012;13 Suppl 8:S20. Tian L et al. Four microRNAs promote prostate cell proliferation with regulation of PTEN and its downstream signals in vitro. PLoS One. 2013 Sep 30;8(9):e75885. Koh CM et al. Myc enforces overexpression of EZH2 in early prostatic neoplasia via transcriptional and post-transcriptional mechanisms. Oncotarget. 2011 Sep;2(9):669-83.




| target | miRNA | target_exp | miRNA_exp |
|---|---|---|---|
| EZH2 | hsa-miR-26a | up | down |
| PAWR | hsa-miR-26a | up | down |
| EZH2 | hsa-miR-28-5p | up | down |
| JAZF1 | hsa-miR-26a | up | down |
| VEGFA | hsa-miR-26a | up | down |
| PIM1 | hsa-miR-26a | up | down |
| PTGS2 | hsa-miR-26a | up | down |
| NCOA4 | hsa-miR-26a | up | down |
| PTGS2 | hsa-miR-28-5p | up | down |
| PTEN | hsa-miR-26a | up | down |
| TRPM8 | hsa-miR-26a | up | down |
| SOD2 | hsa-miR-26a | up | down |
| CYP24A1 | hsa-miR-26a | up | down |

B) DU-145 cell line target prediction

| target | miRNA | target_exp | miRNA_exp |
|---|---|---|---|
| EZH2 | hsa-miR-26a | up | down |
| EZH2 | hsa-miR-28-5p | up | down |
| JAZF1 | hsa-miR-26a | up | down |
| PIM1 | hsa-miR-26a | up | down |
| PTGS2 | hsa-miR-26a | up | down |
| PTGS2 | hsa-miR-28-5p | up | down |
| ETV1 | hsa-miR-26a | up | down |
| ESR1 | hsa-miR-26a | up | down |
| IGFBP1 | hsa-miR-28-5p | up | down |
| CYP24A1 | hsa-miR-26a | up | down |

B) PC-3 cell line target prediction

**A computational network-based framework to study cancer progression using quantitative proteomics data**

Zanzoni A(1), Brun C(1)

*(1) TAGC, Inserm, Aix-Marseille Université, Marseille, F-13288, France.*

**Motivation:** Cancer is an ensemble of diverse diseases arising from genomic alterations occurring in the cell. The technological advances in cancer genomics and the development of successful computational methods have provided new clues on the molecular features of these diseases. However, the detailed mechanisms of cancer onset and progression have not been completely elucidated. The recent launch of high-resolution proteomics techniques will assist the cancer research in the forthcoming years by providing a more precise quantitative picture of cancer phenotypes. Therefore, there is a clear need for implementing computational approaches in order to exploit such data to improve our understanding of cancer progression.
**Methods:** We have developed a computational network-based framework that, by combining quantitative proteomics data with protein interaction analysis, aims to identify dys-regulated cellular functions that can be involved in cancer development. We have applied our framework on a high-resolution proteomic profile of a panel of cell lines that recapitulates breast cancer progression [1]. Our approach consists of the following steps: (i) generation of an high-confidence human interactome that is partitioned into network modules using the Overlapping Clusters Generator (OCG) algorithm [2]; (ii) annotation of network modules with functional information gathered from the Gene Ontology and several pathway databases such as KEGG, Reactome and NCI-PID; (iii) integration of quantitative proteomics data; (iv) statistical assessment of network module dysregulation using non-parametric analysis of variance methods (Kruskal-Wallis, Trend test).
**Results:** Among the 414 network modules analyzed, the ˜35% of them, which contains around 3'200 proteins, is significantly dysregulated. The majority of the network modules show an increased up-regulation during progression representing function and pathways related to cell cycle, transcription and splicing. We also observed a significant down-regulation of modules associated to cell-cell interactions and adhesion. Finally, we will present some preliminary results of a structure-based analysis to assess the wiring of dysregulated modules and identify mutually exclusive (XOR) and concurrent (AND) interactions using protein abundance data as recently suggested [3].
**Contact email:** andreas.zanzoni@inserm.fr
**Supplementary information:**References: 1 - Geiger T, Madden SF, Gallagher WM, Cox J, Mann M. Proteomic portrait of human breast cancer progression identifies novel prognostic markers. Cancer Res. 2012 May 1;72(9):2428-39. 2- Becker E, Robisson B, Chapple CE, Guénoche A, Brun C. Multifunctional proteins revealed by overlapping clustering in protein interaction network. Bioinformatics. 2012 Jan 1;28(1):84-90. 3 - Kiel C, Verschueren E, Yang JS, Serrano L. Integration of Protein Abundance and Structure Data Reveals Competition in the ErbB Signaling Network. Sci Signal. 2013 Dec 17;6(306):ra109.

## NOTES

# List of contributors