THE 3rd EUROPEAN STUDENT COUNCIL SYMPOSIUM

Strasbourg, FRANCE

ESCS'14
3rd European Student Council Symposium
06 September 2014
Strasbourg, France
Sponsors

We thank our sponsors for sharing our vision and helping to make the 3rd European Student Council Symposium a success. Specifically, we are grateful to:
3rd European Student Council Symposium

Strasbourg, France
September 6, 2014
Contents

Agenda........................................................................................................................................... 2

Program ........................................................................................................................................... 4

Scientific Speed Dating ....................................................................................................................... 5

Keynote Speakers ................................................................................................................................. 6

Saprotrophics: a new natural habitat for bioinformaticians? ................................................................. 6

Scale matters! - The importance of scale in the analysis of chromatin landscapes, mutation profiles and protein network architecture ......................................................................................... 7

Network biology: large-scale data and text mining ................................................................................ 8

Oral Presentations ................................................................................................................................. 9

Session 1: Modelling ............................................................................................................................. 9

Session 2: Systems biology ................................................................................................................... 10

Session 3: Networks and Statistics ...................................................................................................... 12

Poster Presentations ............................................................................................................................. 14

Flash Poster Presentation .................................................................................................................... 24

Awards .................................................................................................................................................. 25

Best Presentation Award ...................................................................................................................... 25

Best Poster Award ............................................................................................................................... 25

Travel Fellowships ............................................................................................................................. 25

Acknowledgements ............................................................................................................................. 26

Regional Student Groups Initiative ....................................................................................................... 27

List of Participants ............................................................................................................................... 29

Organizing Committee ......................................................................................................................... 30

Social Event .......................................................................................................................................... 31
Welcome to the third European Student Council Symposium!

It is our great pleasure to welcome you to the 3rd European Student Council Symposium in Strasbourg, France. Our previous editions of the European Symposium, in Ghent (2010) and Basel (2012), were great meetings. We are therefore thrilled to continue our efforts in Strasbourg this year and as in previous years, we strived to create an opportunity for students to meet their peers from all over the world, promoting the exchange of ideas and networking.

We are honored to have Dr. Lennart Martens (University of Ghent), Dr. Jeroen de Ridder (Delft University of Technology), and Dr. Lars Juhl Jensen (The Novo Nordisk Foundation Center for Protein Research) as keynote speakers at this year’s Symposium. Their keynotes promise to be inspiring presentations of exceptional work relevant to everyone in the field.

We will start the Symposium with Scientific Speed Dating: a chance to meet colleagues in an informal and friendly way. Throughout the day we will hear oral presentations from a selection of 8 outstanding student abstracts spanning a wide-range of research areas. The poster session will offer exciting science in various domains, and give everybody a chance to discuss their research topics in more depth. This year the best posters will be selected for a flash poster presentation to give you all a chance to see some of the highest quality posters that were submitted. In the evening, we will end our Symposium with the traditional Social Event. A great place for networking and some fun!

Everyone involved in the organization of this Symposium contributed significantly to make this event happen. Our volunteers have spent many months preparing all aspects of this Symposium ranging from the invitation of keynote speakers, fundraising, advertising and organizing the peer-review process to such mundane things as maintaining a website. This year our team worked hard to secure funds to organize ESCS, give poster and oral presentation awards and, most importantly, offered four travel fellowships to ESCS delegates.

Make the most of this opportunity! Talk to other delegates, ask questions and show enthusiasm about your research if you are presenting! You can make this Symposium a starting point for fruitful future collaborations and another step towards a successful career in computational biology.

Do not forget to visit the Student Council Booth (Booth 6) to learn about the Student Council and its activities and how you can be part of this great effort.

Enjoy your time in Strasbourg!

Pieter Meysman & Margherita Francescatto
Chairs, 3rd European Student Council Symposium 2014

Note: This booklet went into print in August, please check http://escs.iscbsc.org for the latest updates and announcements.
<table>
<thead>
<tr>
<th>Time</th>
<th>Event/Activity</th>
<th>Details on Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>08:00-09:00</td>
<td>Registration &amp; Ice Breaking</td>
<td>5</td>
</tr>
<tr>
<td>09:00-09:05</td>
<td>Opening</td>
<td></td>
</tr>
<tr>
<td>09:05-09:50</td>
<td><strong>Keynote: Saprotrophics: a new natural habitat for bioinformaticians? Dr. Lennart Martens</strong>, Ghent University</td>
<td>6</td>
</tr>
</tbody>
</table>
| 09:50-10:10| Viral DNA replication: new insights and discoveries from large scale computational analysis  
*Darius Kazlauskas*, Institute of Biotechnology, Vilnius University, Lithuania | 9              |
| 10:10-10:30| Using the PDB to explore the conformational space of query proteins with at least one known conformation  
*Ayo Narunsky*, Tel Aviv University, Israel | 9              |
| 10:30-11:00| Coffee Break                                                                |                |
| 11:00-11:20| **Oral Presentations Session 1: Modelling**                                  | 10             |
| 11:00-11:20| Applications of Proteochemometrics – From Species Extrapolation to Cell Line Sensitivity Modelling  
*Isidro Cortes*, Institut Pasteur, Unité de Bioinformatique Structurale; CNRS, France |                |
| 11:20-11:40| Genotyping microsatellites in next- An exploration of the 3D chemical space has highlighted a specific shape profile for the compounds intended to inhibit protein-protein interactions  
*Mélaine A. Kuenemann*, University of Paris Diderot, France | 10             |
| 11:40-12:00| **Oral Presentations Session 2: Systems biology**                            | 11             |
| 11:40-12:00| Mining the human proteome for conserved mechanisms  
*Stefan Naulaerts*, University of Antwerp, Belgium |                |
| 12:00-14:15| Lunch & Poster Presentations                                                | 14             |
| 14:15-15:00| **Oral Presentations Session 3: Networks and Statistics**                   | 7              |
| 14:15-15:00| Keynote: Scale matters! - The importance of scale in the analysis of chromatin landscapes, mutation profiles and protein network architecture, Dr. Jeroen de Ridder, Delft University of Technology |                |
| 15:00-15:20| Tetranucleotide usage in Identification and Analysis of Methylation Call Differences between Bisulfite Microarray and Bisulfite Sequencing Data with Statistical Learning Techniques  
*Matthias Döring*, Max Planck Institute for Informatics, Germany | 12             |
| 15:20-15:40| Hybrid approaches for the detection of networks of critical residues involved in functional motions in protein families  
*Dagoberto Armenta Medina*, IBT/UNAM, Mexico | 12             |
<table>
<thead>
<tr>
<th>Time</th>
<th>Event</th>
</tr>
</thead>
</table>
| 15:40-16:00| **Improving duplicated nodes position in vertebrate gene trees**  
*Amélie Peres*, Ecole Normale Supérieure, Institut de Biologie de l’ENS, IBENS, France |
| 16:00-16:30| Coffee Break                                                |
| 16:30-17:00| **Oral Presentations Session 4**                            |
| 17:00-17:45| **Keynote: Network biology: large-scale data and text mining,**  
*Dr. Lars Juhl Jensen*, The Novo Nordisk Foundation Center for Protein Research |
| 17:45      | Symposium Final Words & Award Ceremony                      |
| 19:00      | Student Social Event                                       |
Program
Scientific Speed Dating

We are continuing the speed-dating event at this year’s Symposium! No, we are not going to help you find your life partner (even though it may be a side effect), we are talking about scientific speed dating to chat with your colleagues and break the ice in a convivial atmosphere.

Who are the other people who will spend the day with you? Where are they working? What are their research interests? Are they Ph.D. students? Is this the first time they attended the Symposium? Are they here to present a poster? Will they attend ECCB? Getting to know people during this event will help you make the most of your Student Council Symposium experience. And there is always lunch and coffee breaks to follow up on interesting beginnings.

Don’t be shy! Make the most of scientific speed dating!
Keynote Speakers

Saprotrophics: a new natural habitat for bioinformaticians?

A large amount of public data is available, covering various aspects of biology and biomedicine. As a result, it is becoming increasingly interesting to perform meta-analyses on these data, foregoing the traditional need to first obtain novel data from a dedicated analysis or experiment. Indeed, the information encoded across many different experiments often provides access to knowledge that could not have been extracted from a single or a few experiments. At the same time, the signal in these experiments can easily be lost in the noise, as heterogeneity at all levels poses a substantial threat to the re-use or re-interpretation of large, unrelated collections of data. It is therefore paramount to employ purpose-built filters and/or quality control prior to repurposing existing data from the public domain. Such efforts are however hampered by the often fragmentary annotation of these various data sets, and it is therefore often necessary to second-guess a lot of the context. Based on the specific example of proteomics data, I will show that such challenges can be met, starting from the creation of a system for data exchange, all the way to the application of public data to extract novel knowledge and insight, both at the technical as well as the biological level. In the end, such analyses can create a wholly new branch of ‘saprotrophic' bioinformatics, in which data that have already served their intended purpose for the original authors, is essentially endlessly recycled in novel useful roles from public domain databases.

Lennart Martens

Ghent University

Dr. Lennart Martens is Professor of Systems Biology in the Department of Biochemistry in the Faculty of Medicine and Health Sciences at Ghent University, and Group Leader of the Computational Omics and Systems Biology (CompOmics) Group at VIB, both in Ghent, Belgium. He holds, or has held positions in various HUPO Initiatives, including the HUPO Proteomics Standards Initiative (PSI), and has served as the 2011 Chair of the ABRF Proteomics Informatics Research Group (iPRG). He studied his PhD in the Department of Biochemistry at Ghent University. During this time, he created the well-known PRIDE repository at EMBL-EBI as a Marie Curie fellow of the European Commission. After receiving his Doctorate in Sciences: Biotechnology from Ghent University in Belgium on the subject of proteomics informatics, he moved to EMBL-EBI in Cambridge, UK where he served as the PRIDE Group Coordinator until his return to Ghent University and VIB in 2009. He pioneered the meta-analysis of large-scale and often heterogeneous proteomics data, a topic that continues to be an active field of research for his current CompOmics group at VIB and Ghent University.
Scale matters! - The importance of scale in the analysis of chromatin landscapes, mutation profiles and protein network architecture

The notion of scale plays an important role in how we observe our environment. For instance, considering a tree from a distance of a light-year or a nanometer is meaningless, but would make sense at a distance of a few meters. Clearly, objects only exist as meaningful entities across a certain range of scales. The concept of scale is also omnipresent in genomics and molecular biology. Some cellular functions may involve a single direct protein interaction (small scale), whereas others require more indirect interactions, such as protein complexes (medium scale) and interactions between large modules of proteins (large scale). Other examples include the various levels at which the genome is organized within the cellular nucleus and the various levels at which the genome is modified by epigenetic marks. In this talk I will demonstrate the importance of incorporating scale in the analysis of large-scale molecular data. I will do this based on three examples. First of all, I will present the most comprehensive characterization of the genetic and epigenetic determinants of target site selection of genome-integrating elements to date. Our scale-aware analyses reveal that, at a genome-wide scale, biases are largely conserved across elements, whereas, at a small scale, the integration biases of different elements are driven by element-specific genomic features. I will also address the scale-aware analysis of cancer-causing retroviral integrations. These analyses demonstrate that, on top of the known clustering of insertions along the linear genome, clustering is also apparent at a different scale, which can only be revealed if the 3D conformation of the genome is taken into account. This knowledge has important implications for target gene determination and, more generally, provides new hypotheses on how (viral) enhancers act on their targets. Finally, I will discuss how to characterize topological structure in protein interaction networks. We developed scale-aware versions of known graph topological measures based on diffusion kernels to characterize the topology of networks across all scales simultaneously, generating a so-called graph topological scale-space. We demonstrate that graph topological scale-spaces capture biologically meaningful features that provide new insights into the link between function and protein network architecture.

Jeroen de Ridder

Bioinformatics for Health and Disease, Delft University of Technology

Dr. de Ridder did his PhD on pathway discovery in insertional mutagenesis data at the Delft Bioinformatics lab (Reinders group, TUD) and at the Netherlands Cancer Institute (Wessels group). His PhD work resulted in a statistical framework for the analysis of retroviral insertional mutagenesis data to identify cancer genes in mice and was used in the analysis of several mutagenesis datasets in collaboration with researchers at the Netherlands Cancer Institute and the Wellcome Trust Sanger Institute (Hinxton, UK). During his PhD, Dr. de Ridder visited the Shmulevich lab at the Institute for Systems Biology (Seattle, USA), where he worked on multi-scale methods for genomic data analysis. In addition, Dr. de Ridder has been an active member of the national and international bioinformatics communities, first as member of the board of the Dutch Regional Student Group, co-initiated by NBIC, and later as the elected chair of the Student Council, an initiative of the International Society for Computational Biology (ISCB). Dr. de Ridder currently holds a position as Assistant Professor in the Delft Bioinformatics Group.
Network biology: large-scale data and text mining

Methodological advances have in recent years given us unprecedented information on the molecular details of living cells. However, it remains a challenge bring together all the available data on individual genes to facilitate systems-level analyses of, for example, diseases. Networks have proven to be a very useful abstraction for bridging this gap between the single-gene and the systems level. In my presentation I will describe the STRING database (http://string-db.org), which scores and integrates evidence from a diverse range of curated databases, raw data repositories, automatic text mining, and computational prediction methods to provide the most comprehensive protein association network possible. I will also introduce a suite of three new web-based resources that use similar techniques to associate the proteins with cellular compartments (http://compartments.jensenlab.org), tissues (http://tissues.jensenlab.org), and diseases (http://diseases.jensenlab.org) to enable systems biology studies of diseases, taking into account both the interactions and the spatial localization of the proteins.

Lars Juhl Jensen

The Novo Nordisk Foundation Center for Protein Research

Lars Juhl Jensen started his research career in Søren Brunak’s group at the Technical University of Denmark (DTU), from which he in 2002 received the Ph.D. degree in bioinformatics for his work on non-homology based protein function prediction. During this time, he also developed methods for visualization of microbial genomes, pattern recognition in promoter regions, and microarray analysis. From 2003 to 2008, he was at the European Molecular Biology Laboratory (EMBL) where he worked on literature mining, integration of large-scale experimental datasets, and analysis of biological interaction networks. Since the beginning of 2009, he has continued this line of research as a professor at the Novo Nordisk Foundation Center for Protein Research at the Panum Institute in Copenhagen and as a co-founder and scientific advisor of Intomics A/S. He is a co-author of more than 100 scientific publications that have in total received more than 10,000 citations. He was awarded the Lundbeck Foundation Talent Prize in 2003, his work on cell-cycle research was named “Break-through of the Year” in 2006 by the magazine Ingeniøren, his work on text mining won the first prize in the “Elsevier Grand Challenge: Knowledge Enhancement in the Life Sciences” in 2009, and he was awarded the Lundbeck Foundation Prize for Young Scientists in 2010.
Oral Presentations
Session 1: Modelling

1. Viral DNA replication: new insights and discoveries from large scale computational analysis

Darius Kazlauskas and Ėčsolovas Venclovas
Institute of Biotechnology, Vilnius University, Lithuania

Ability to replicate is essential for all living entities. Duplication of genetic information is carried out by replication proteins. DNA replication is well studied in T7, T4 phages and herpesviruses; however, the information about replication mechanisms from other groups of viruses is either scarce or missing altogether. Double-stranded (ds) DNA viruses infect cells from all domains of life, evolve fast and are very diverse. Their genome size varies from 5 to 2500 kbp. To better understand viral DNA replication we identified replication proteins in dsDNA viruses using current state-of-the-art homology detection methods. Over 150,000 proteins from 1574 genomes were analyzed. We found out that the composition of replication machinery depends on virus genome size. Small viruses (<40 kbp) use protein-primed DNA replication or rely on replication proteins from the host. Large viruses (>140 kbp) have their own RNA-primed replication apparatus often supplemented with processivity factors and DNA topoisomerases to increase replication speed and efficiency. Latter insight led us to a search for „missing” replication components in large genomes. This has resulted in a discovery of single-stranded DNA binding (SSB) proteins in largest eukaryotic viruses. Surprisingly these proteins turned out to be homologs of SSB proteins previously thought to be specific for T7-like phages. Another surprise came from the analysis of the herpesviral helicase-primase complex. We found that its component (UL8) is a highly diverged inactivated B-family DNA polymerase.

2. Using the PDB to explore the conformational space of query proteins with at least one known conformation

Aya Narunsky, Haim Ashkenazy, Rachel Kolodny and Nir Ben-Tal
Tel Aviv University, Israel

Proteins function often involves conformation alteration, e.g., shift between active and inactive states of receptors, etc. The Protein Data Bank (PDB) can be exploited to study conformational changes, as our analysis showed that: (a) most of the proteins have more than one available structure, and (b) often proteins that share a similar conformation have also other conformations in common.

We used this to develop a method and a webserver, ConTemplate, aiming at predicting putative conformations for a query protein with at least one known conformation, based on its similarity to other proteins in the PDB, and alternative conformations of these proteins.

Briefly, ConTemplate creates an ensemble of conformations for the query using a three steps process. First, the entire PDB is scanned, and proteins whose structural-similarity to the query is above a preset threshold are collected. Second, for each of the collected proteins, additional known conformations are indicated, and clustered. In the third step, the server calculates models of the query in various conformations using the structure-based sequence alignments found in the first step, and the cen-


ters of the clusters found in the second step as templates.

We demonstrate the method with the kinase domain of the Epidermal Growth Factor Receptor (EGFR). Using the inactive conformation as our query, we reproduce the active conformation with root mean square deviation of 1.76Å, based on the query's structural similarity to the inactive conformation of Abl tyrosine-kinase, and the known active conformation of that kinase. The sequence identity between the two kinase domains is only 40%, and the fact that they share similar active and inactive conformations might not be obvious.

**Session 2: Systems biology**

3. *Applications of Proteochemometrics – From Species Extrapolation to Cell Line Sensitivity Modelling*

*Isidro Cortes, Gerard J.P. van Westen, Daniel S. Murrell, Eelke B. Lenselink, Andreas Bender and Thérèse Malliavin*

Institut Pasteur, Unité de Bioinformatique Structurale; CNRS, France

Proteochemometrics (PCM) is a computational technique to model the bioactivity of multiple ligands against multiple targets, e.g. proteins or cell lines, simultaneously. Therefore, PCM has enabled the exploration of the selectivity and promiscuity of ligands on different protein classes [1,2]. Indeed, the simultaneous inclusion of both chemical and target information permits the extra- and interpolation to predict the bioactivity of compounds on yet untested targets [3]. In this contribution, we will firstly show a methodological advance in the field [4], namely how Bayesian inference (Gaussian Processes) can be successfully applied in the context of PCM for (i) the determination of the applicability domain of a PCM model; (ii) the prediction of compounds bioactivity as well as the error estimation of the prediction; and (iii) the inclusion of the experimental uncertainty of bioactivity measurements during model training. Additionally, we will describe how PCM can be useful in medicinal chemistry to concomitantly optimize compounds selectivity and potency, in the context of two application scenarios, which are: (a) modelling isoform-selective cyclooxygenase inhibition; and (b) large-scale cancer cell line drug sensitivity prediction.


4. *An exploration of the 3D chemical space has highlighted a specific shape profile for the compounds intended to inhibit protein-protein interactions*

*Mélaine A. Kuenemann, Laura M.L. Bourbon, Céline M. Labbé, Bruno O. Villoutreix and Olivier Sperandio*

University of Paris Diderot - Inserm UMR-S 973, France

The vital role of Protein-Protein Interactions (PPI) for Life makes them the subject of a growing number of drug discovery projects. Yet, the specific properties of PPI (often described as flat, large and hydrophobic) require a dramatic paradigm shift in our way to design the small
compounds meant to modulate them with therapeutic perspectives. To this end, successful inhibitors of PPI targets (iPPI) may be used to discover what singular properties make this type of inhibitors capable of binding to such intricate surfaces. Among the properties from which lessons could be learnt, the 3D characteristics of iPPI have been pinpointed as essential. Understanding the putative shape profile of iPPI could therefore help the design of a new generation of inhibitors with improved ligand efficiencies.

In an attempt to identify such putative 3D characteristics, we have collected the bioactive conformations of 58 orthosteric iPPI and compared them to those of 1623 inhibitors of conventional targets (e.g. enzymes) collectively from different databases (2P2I, PDBind, PDB). Because the known heavier and more hydrophobic character of iPPI could conceal other characteristics, we have imposed that none of the identified descriptors correlate with the hydrophobicity or the size of the compound. The newly identified properties were further confirmed as characteristic to iPPI using the data of much larger datasets including our iPPI-DB, eDrugs3D and a representative subset of the bindingDB. Most noticeably, the essential property revealed by this study illustrates how iPPI manage to bind to the hydrophobic patches of PPI. Interestingly, the absence of correlation of such property with the hydrophobicity and the size of the compounds, that can be a liability for drug developments, opens new ways to design potent iPPI with a better balance for some of the pharmacokinetic features.

5. Mining the human proteome for conserved mechanisms

Stefan Naulaerts, Pieter Meysman, Wim Vanden Berghe, Bart Goethals and Kris Laukens

University of Antwerp, Belgium

All cells find themselves in continually changing environments to which they have to adapt, using their sensory system to provide input for the regulatory systems that integrate the information and trigger the eventual effectors. Recently, several advancements in the understanding of these signaling cascades have been propelled by “omics” approaches and vast public compendia of protein interactions have been created for several model organisms. Although these information-rich resources exist, the adaptive nature of protein complexes and signalling cascades remain poorly understood, as current predominant approaches are not suited to investigate the dynamic relationships between proteins. This complicates the analysis of any protein interaction data. It thus remains a challenge to find out how biological entities cooperate to regulate cellular response to stimuli.

Here, we present an integrative method, reliant on advanced pattern mining approaches, to gain a deeper understanding of protein network dynamics. To this end, we created a large Homo sapiens compendium of proteomics papers that report differentially expressed proteins in cell lines. We analysed this collection with frequent itemset mining to identify proteins that are often co-occurring and used these patterns as the backbone structure of our further analysis. These patterns were then enriched with additional attributes, such as gene expression correlation, protein localization and Gene Ontology, and used as a filter on top of an integrated protein interaction network, obtained by fusing several of the most popular resources.

As a case study, we compared patterns derived from gene and protein expression data for several cancer types and states (pre- and post-metastasis), which resulted in distinctly differ-
ent patterns. We found that pattern-based analysis of the cascade of up- and down-regulation on multiple "omics" levels can help to identify the cellular logic circuits.

Session 3: Networks and Statistics

6. Identification and Analysis of Methylation Call Differences between Bisulfite Microarray and Bisulfite Sequencing Data with Statistical Learning Techniques

Matthias Döring, Gilles Gasparoni, Jasmin Gries, Karl Nordström, Pavlo Lutsik, Jörn Walter and Nico Pfeifer
Max Planck Institute for Informatics, Germany

Methylation of DNA is an epigenetic modification known to play a prime role in gene silencing and is an important topic in epigenetic research. Although accurate methods for measuring DNA methylation exist, technology-dependent errors give rise to inconsistencies between method results. We studied DNA methylation measurements from the Infinium HumanMethylation450 microarray (β450K) and whole genome bisulfite sequencing (βWGBS) to evaluate whether there are locus-specific measurement differences, Δβ = β450K − βWGBS, and whether this effect is predictable using statistical learning techniques. The ability to determine inconsistent methylation measurements based on bisulfite-converted sequencing data would be valuable for epigenome-wide association studies, in which such positions could be treated accordingly. We built support vector regression models based on Illumina bisulfite-sequencing data of HepaRGd7R2 to predict Δβ. A measure for read similarity was obtained via numerical and string kernels as well as set kernels. We introduced the notion of hybrid string kernels to afford a similarity measure for both, numeric and string input simultaneously.

Using a read-based set kernel, we found that the predicted values of Δβ correlated significantly with the observed outcomes (r = 0.73, p-value < 2.2 X 10^{-16}). This model utilized CpG positions reflecting methylation. To obtain a model independent of βWGBS, we excluded the CpG positions and still found a significant correlation (r = 0.37, p-value < 2.2 X 10^{-16}). Features beside the reads played only a minuscule role in the emergence of inconsistent methylation measurements. To our knowledge, this is the first time someone was able to show that differences between β450K and βWGBS are predictable from the sequence. Furthermore, our new hybrid string kernels gave the best performance among the non-set-kernel approaches and could be a valuable approach also in other settings.

7. Hybrid approaches for the detection of networks of critical residues involved in functional motions in protein families

Dagoberto Armenta Medina and Ernesto Perez Rueda
IBT/UNAM, Mexico

Background: Currently there is great interest to identify critical residues to have a better understanding and engineering of protein families. Diverse approaches combine sequence information, structural data, dynamics analysis and functional description to determine the importance amino acids associated to protein function. In this work, we proposed a hybrid approach for the identification of critical residues in proteins combining the use of evolutionary information (coevolution) and cross-correlation of atomic fluctuations derived from
Anisotropic Normal Mode Analysis simulations (ANMA), which in turn was also contrasted with previous approaches.

Results: Combining the information of covariance matrix derived from Statistical Coupling Analysis (SCA) and the cross-correlation matrix of atomic fluctuations derived from ANMA, it was possible to identify a network of evolutionarily coupled residues involved in relevant motions in the Adenylate kinase protein family. The outstanding sites revealed by our hybrid approach (ANMA.SCA) show a high correspondence with experimental data confirming the critical role of these sites in the functional mobility of proteins. Additionally we find that our approach ANMA.SCA is complementary and maintains a good correspondence with previous approaches derived from extensive molecular dynamics but being faster and less expensive in computation resources.

Conclusions: The hybrid approach ANMA.SCA by means of the detection of networks of critical sites and its topological study opens a wide range of possibilities in the study of functional motion of protein families and reveals hidden aspects of its dynamic personality.

8. Improving duplicated nodes position in vertebrate gene trees

Amélie Peres and Hugues Roest Crollius
Ecole Normale Supérieure, Institut de Biologie de l’ENS, IBENS, France

Gene phylogenies are essential for many biological evolutionary studies. However, phylogenetic reconstructions are difficult to model, especially when they include gene duplications. In this study, we have developed a method to improve the positions of duplications in gene trees produced by TreeBest.

We first investigated methods to automatically identify incorrectly positioned duplications with different independent criteria. One criterion relies on a “confidence score”, a measure comprised between 0 and 1 and assigned to each duplicated node. The score reflects the ratio between the number of species with a duplicated genes and the total number of species under this node. A well-supported duplication will thus have a score close to 1. Once a node is considered to be poorly supported, we edit the node by replacing it by a speciation node, and testing the nodes just below using the same criterion. If the nodes pass the test, the duplication is created at this new position in the tree.

In order to assess the quality of the new edited trees, we tested criteria on all phylogenetic trees available in the Ensembl compara database (20194 gene trees in Ensembl 71) and then used the new gene tree databases and the initial Ensembl gene tree database to reconstruct ancestral genomes using AGORA. AGORA is an algorithm developed in our laboratory to reconstruct ancestral gene orders, and its performances are very sensitive to the quality of the input gene trees. In particular, the length of the reconstructed ancestral chromosomal regions varies substantially depending on the quality of the input gene trees. We find that using the confidence score criterion significantly improves the positions of duplications within gene trees compared to the initial Ensembl gene tree database. The optimal value is a threshold score of 0.3, at which 39% of the 197 894 duplication nodes are edited, resulting in a 200% increase in the N50 length of ancestral Boreoeutheria reconstruction.
Poster Presentations

9. In silico identification and in vitro validation of potential cholestatic BSEP inhibitors utilizing a 3D ligand-based pharmacophore model

Susanne Hermans, Rick Greupink, Frans Russel and Tina Ritschel
Computational Discovery and Design, Centre for Molecular and Biomolecular Informatics, Radboud University Medical Centre, Nijmegen, The Netherlands

Drug-induced cholestasis is a frequently observed side effect of drugs and is often caused by an unexpected interaction with the bile salt export pump (BSEP/ABCB11). BSEP is the key membrane transporter responsible for the transport of bile acids from hepatocytes into bile. Here, we developed a pharmacophore model that describes the molecular features of compounds associated with BSEP inhibitory activity. To generate input and validation data sets, in vitro experiments with membrane vesicles over-expressing human BSEP were used to assess the effect of compounds (50 µM) on BSEP-mediated 3H-taurocholic acid transport. The model contains two hydrogen bond acceptor/anionic features, two hydrogen bond acceptor vector features, four hydrophobic/aromatic features and exclusion volumes. The pharmacophore was validated against a set of 59 compounds, including registered drugs. The model recognized 9 out of 12 inhibitors (75%), which could not be identified based on general parameters, such as molecular weight or SlogP, alone. Finally, the model was used to screen a virtual compound database. A number of compounds found via virtual screening was tested and displayed statistically significant BSEP inhibition, ranging from 13±1% to 67±7% of control (P<0.05). In conclusion, we developed and validated a pharmacophore model that describes molecular features found in BSEP inhibitors. The model may be used as an in silico screening tool to identify potentially harmful drug candidates at an early stage in drug development.

10. Druggability prediction performances related to different pocket estimations

Alexandre Borrel, Leslie Regad, Henri Xhaard, Michel Petitjean and Anne-Claude Camproux
Molécules Thérapeutiques in silico (MTi), Université Paris Diderot, France and University of Helsinki, Finland

Therapeutical molecules bind to preferred sites of action, which are located within proteins or at their surface. Estimation and characterization of pockets is a major issue in drug target discovery. Among the molecules, “drug-like molecules” are small molecules with particular properties as able to cross the digestive tract. Pocket “druggability”, the ability of a pocket to bind “drug-like” molecules, is essential for discovering new targets.

Identifying druggable pockets is possible by different statistical models of prediction which differ in methods used to estimated pockets, in descriptors used to characterized pockets and in statistical methods used. Moreover, the quality of these approaches is limited by the few data available, and by the pocket estimation, pockets changes with the estimations. However of new target discovery, it is important to be able to predict the druggability of a pocket in its apo form that means when it is not yet bound to a ligand and deformed by the interaction with one ligand. We propose a model to predict pocket druggability from holo or apo form. From one protein set, we used different ap-
proaches to estimate pockets. With ligand, defines pockets as protein atoms less than 4 Å away from the ligand or without ligand, based on two geometric estimations DoGSite and fpocket. Pockets estimated using three approaches, were characterized, using 57 descriptors. Three pocket sets are generated. From each pocket sets, we built a pocket “drug-gability”, based on a Linear Discriminant Analysis (LDA). The construction of these models consisted in the selection of LDA models with the best accuracy and containing as few descriptors as possible. The model, giving the best performance is conserved. Finally we used a consensus of 4 LDA models, built from pockets estimated by fpocket and which present a good accuracy (close to 80%), better than literature, on other pocket sets and on apo pockets.

11. LBIBCell: A Cell-Based Simulation Environment for Morphogenetic Problems

Simon Tanaka and Dagmar Iber
ETH Zurich, Switzerland

We developed a cell-based simulation framework to study morphogenetic problems. The tissue model is based on the visco-elastic cell model approach [1-2], which models interactions between the viscous interstitial fluid and cytoplasm, and the elastic structures such as cell membrane, cell-cell junctions and cytoskeleton using the immersed boundary method. The tissue model is coupled to a system of advection-diffusion-reaction partial differential equations to model signaling, and thus is predestined to study cell-based mechanisms based on the interaction of tissue dynamics and signaling, as appearing in many morphogenetic problems.

We studied the dynamics of a Turing-type reaction-diffusion signaling system, which locally controls the proliferation rate, and demonstrate that, in comparison to the continuous formulation counterpart, surprising effects (such as dividing cell clusters) emerge. The software will be published under the GPL license.

12. Phosphate and ribose structural isosteric replacement in the Protein Data Bank

Yuezhou Zhang, Alexandre Borrel, Leslie Regad, Anne-Claude Camproux, Gustav Boije Af Gennäs, Jari Yli-Kauhaluoma and Henri Xhaard
Molécules Thérapeutiques in silico (MTi), Université Paris Diderot, France and University of Helsinki, Finland

In this work, we have studied the structural isosteric replacements of phosphate and ribose found in protein-ligand complexes available in the Protein Data Bank database (June of 2014 release). We developed a computational protocol that was used to construct 157 datasets. Each of these datasets is composed of several superimposed ligands, including POP, AMP, ADP or ATP to use as references, derived from superimposed molecular complexes. Structural replacements of ribose and phosphate groups were then extracted and studied: we identified a set of 15 common structural isosteres of phosphate and 43 structural isosteres of ribose. In addition to classical isosteres of phosphate, we found unexpected types of replacements that do not conserve charge or polarity, for example phosphate and ribose replaced by aliphatic groups, phenyl, or carbamoyl groups. The structural mechanism involved in structural isosteres appears varied: New interactions may be created, water molecules are important, in some case ion plays a role, and of course large and small conformational changes do occur at the binding sites. This study has implications
both in the field of medicinal chemistry, i.e. it expands our knowledge of structural isosteres, and in the field of chemoinformatics, since our results have implications with respect to the definitions of chemical similarity.

13. MultiGeMS: Detection of SNPs from Multiple Samples Using Model Selection on High-Throughput Sequencing Data

Gabriel Murillo, Na You, Xiaoquan Su, Kang Ning and Xinping Cui
Department of Statistics, University of California, Riverside, USA

Recent advances in high-throughput sequencing (HTS) promise revolutionary impacts in science and technology, including the areas of disease diagnosis and pharmacogenomics. An important way to analyze the increasingly abundant HTS data, is through the use of single nucleotide polymorphism (SNP) callers. Building on previously published work with the single sample SNP caller Genotype Model Selection (GeMS), a suite of GeMS SNP detection procedures, including a multiple sample version of GeMS, will be introduced. As with single sample GeMS before it, simulation studies demonstrate that multiple sample GeMS has the best balance of sensitivity and positive predictive value (PPV) among a selection of popular SNP callers. Real data analyses also support this conclusion. Finally, future work regarding GeMS SNP calling with metagenomics data will be given.


Thomas Coudrat, John Simms, Denise Wootten, Arthur Christopoulos and Patrick Sexton
Monash University, Australia

G Protein-Coupled Receptors (GPCRs) are a superfamily of transmembrane proteins that mediate cellular responses to their environment upon binding of an effector to their extracellular-facing binding pocket. With over 800 human GPCRs playing key roles in modulating tissue/cell physiology and homoeostasis, they represent a major target for pharmaceutical intervention.

Recently, an increasing number of GPCR X-ray crystal structures were solved; Computer-Aided Drug Discovery (CADD) methods such as Virtual Screening (VS) leverage that structural information to rationalise drug discovery efforts. VS attempts to identify new drug leads by ranking libraries of small molecules based on the predicted complementarity between small molecules and the target GPCR binding pocket. The success of VS is highly dependent on the conformation of the target binding pocket; therefore, a key step of CADD is to refine the binding pocket within a protein structure.

Here we present a new computationally efficient Ligand Directed Modelling (LDM) method for GPCR binding pocket refinement. This method aims to establish the global energy minimum of a GPCR binding pocket in complex with a known active ligand. Our LDM method samples the GPCR structure and docks a single known active ligand for that GPCR on each generated structure. All binding pocket/ligand complexes are scored and the best are selected to start a subsequent round of sampling, docking and selection. This results in a recursive refinement of the GPCR binding pocket in complex with its known active ligand.

To benchmark the method, we have used family A GPCR structures that have been crystallised with more than one known active ligand, and tested our LDM in a range of different refinement scenarios. In each LDM experiment, the resulting structures were scored and compared
to both the starting and final X-ray crystal structures. This benchmark provides a guideline for the application of this LDM method in future CADD projects.

15. The CAD-score webserver: contact area-based comparison of structures and interfaces of proteins, nucleic acids and their complexes

Kliment Olechnovic and Ceslovas Venclovas
Vilnius University, Lithuania

The Contact Area Difference score (CAD-score) web server provides a universal framework to compute and analyze discrepancies between different 3D structures of the same biological macromolecule or complex. The CAD-score method is based on calculating differences of contact areas derived from the Voronoi diagram of balls that correspond to heavy atoms of van der Waals radii. CAD-score was initially developed for proteins and recently has been extended to RNA/DNA 3D structures. The CAD-score web server accepts both single-subunit and multi-subunit structures and can handle all the major types of macromolecules (proteins, RNA, DNA and their complexes). It can perform numerical comparison of both structures and interfaces. In addition to entire structures and interfaces, the server can assess user-defined subsets. The CAD-score server performs both global and local numerical evaluations of structural differences between structures or interfaces. The results can be explored interactively using sortable tables of global scores, profiles of local errors, superimposed contact maps and 3D structure visualization. The web server could be used for tasks such as comparison of models with the native (reference) structure, comparison of X-ray structures of the same macromolecule obtained in different states (e.g. with and without a bound ligand), analysis of nuclear magnetic resonance (NMR) structural ensemble or structures obtained in the course of molecular dynamics simulation. The server is freely accessible at http://bioinformatics.ibt.lt/cad-score.

16. Computational Analysis of DNA Polymerases and their Homologs in Bacterial Genomes

Kęstutis Timinskas and Ėslovas Venclovas
Institute of Biotechnology, Vilnius University, Lithuania

DNA polymerases are essential to bacterial survival. All known bacterial DNA polymerases belong to five families: A, B, C, X and Y. Most experimental data is available for Escherichia coli, which has polymerases of four families: Pol III (C family) - the primary polymerase in DNA replication; Pol I (A family) - a supporting replicative polymerase, also involved in DNA repair; Pol IV and Pol V (Y family) - translesion synthesis repair polymerases; Pol II (B family) - repair and putative backup replicative polymerase. Although biochemical data from some other bacteria are available, it is insufficient to draw a broader picture. To better understand the distribution, combinations and functions of bacterial DNA polymerases of all families we have collected them from nearly 2000 fully sequenced bacterial genomes. We then analyzed each family in more detail: divided into smaller groups, identified known functional domains in all sequences and compared sequence and structure conservation between families and groups.

Polymerases of the C-family are consistently encoded in all bacterial genomes and appear to be universally responsible for chromosomal DNA replication in bacteria. Polymerases of A
and Y families are also widespread among bacteria. Most of the A family polymerases cluster together according to the sequence similarity, suggesting that they may function similarly to E. coli Pol I. On the other hand, Y family polymerases can be subdivided into several groups: a relatively narrow group containing E. coli Pol V, several groups similar to Pol IV and two groups of ImuB (an inactive polymerase, involved in error-prone DNA repair). Interestingly, the B-family, essential in eukaryotes, is not widespread among bacteria. Surprisingly, at least 2% of analyzed bacteria were found to contain only C-family polymerases in their genomes, suggesting that C-family alone may be sufficient to carry out all the essential DNA synthesis tasks in the bacterial cell.

17. **Loads-to-nodes: a framework to distribute biological database and execute sequential programs in parallel environments**

*Ahmad Salah and Li Kinli*
Hunan University, China

Background: Data-analytical programs that handle biological data face massive databases sizes that force researchers to provide a parallel version of their programs. These parallel versions are critical for contributors during the development phase, to evaluate the program performance on huge databases, and for final users, to harness the advances of the hardware for better performance and bigger databases. Herein, we present a software framework that removes the parallelism burden from the researchers' shoulders. While there are general purpose task scheduling tools like HTCondor the proposed framework is more specific for biological data. Moreover, it has an optimization module that selects the optimal number of resources, node and/or cores. Since for each specific problem size increasing the computational resources may lead to performance degradation. The framework divides and distributes the biological databases and then converts the sequential program into a parallel one or extends the targeted parallel resource from multi-core to grid levels. The framework is able to divide databases, of FASTA, PDB and ent (SCOP format) file formats, into balanced size portions. Herein, we have three test cases, but the framework is not limited to these cases. The test cases include sequence comparison, protein structure comparison and CpG island finder.

Results: The speedup of the proposed framework is linear, close to optimal. The source code and executables are available at [http://biocloud.hnu.edu.cn/loads-to-nodes/](http://biocloud.hnu.edu.cn/loads-to-nodes/).

18. **Spatial organization and distribution of linear motifs in the Ankyrin repeat protein family and its binding partners**

*Nina S. Verstraete, Ignacio E. Sánchez and Diego U. Ferreiro*
Universidad de Buenos Aires – CONICET, Argentina

Background: Interactions between proteins regulate cellular physiology. Many of these interactions involve the recognition of short peptides regions (i.e. short linear motifs, SLiMs) which can be characterized by simple sequence patterns, usually found in intrinsically disordered regions or in loops connecting globular or transmembrane domains. These peptide-domain interactions are typically transient and often involve folding upon binding, challenging the lock-and-key paradigm of protein recognition.

Ankyrin-repeat domains are one of the most frequently observed protein-protein interactors
in nature. These domains are composed of tandem arrays of recurrent amino acids that cooperatively fold into elongated structures that mediate molecular recognition with high specificity. Many ankyrin-binding sites are either predicted or demonstrated to correspond to extended peptides mimicking SLiMs.

Description: We present here an exhaustive analysis of linear motif identification in Ankyrin proteins and their binding partners. We searched for enriched or depleted SLiMs with respect to a random exploration of the sequence-space in the Ankyrin protein family and their partners. We also analyzed the spatial distribution of SLiMs along the protein sequences and describe how particular SLiMs are structurally distributed in the Ankyrin-containing proteins.

Conclusions: This computational work presents sequence and structure-based approaches to analyze linear motif-mediated protein interactions in the Ankyrin repeat protein family. We discuss that the presence of functional constraints can conflict with the Ankyrin-repeats domains folding dynamics which in turn modulate the evolution of biological interactions.

19. Population genomics of the phytopathogenic fungi Microbotryum violaceum

Hélène Badouin, Jerome Gouzy, Alodie Snirc, Sophie Siguenza, Antoine Branca and Tatiana Giraud

ESE, CNRS, Université Paris Sud, France

Understanding how species adapt to their environment is a key question in evolutionary biology. To achieve this goal, scientists want to identify what genes are responsible for local adaptation, how many there are and how they are distributed within genomes. Studying patterns of genetic variation within and between populations can reveal signatures of positive selection.

Pathogens are excellent models to study adaptation, because they tend to evolve rapidly in response to the coevolving host environment. The basidiomycete Microbotryum violaceum is a species complex that parasites more than a hundred species of Caryophyllaceae. The fungus hijacks the reproductive system of its host and causes its sterility.

The analysis of EST sequences of Microbotryum violaceum has revealed genes that have potentially experienced positive selection between species adapted to different hosts. Nevertheless, genes causing host adaptation remain largely unknown. The genome of Microbotryum violaceum has been recently sequenced. In order to identify candidates for local adaptation, we sequenced 50 individuals of two sister-species, M. violaceum ssp silenes-dioicae and M. violaceum ssp lychnidis-dioicae and identified 200,000 single nucleotide polymorphisms.

We analyzed the genetic structure of both species, and found no recent hybrids in our dataset. We confirmed that European M. lychnidis-dioicae populations are structured in 3 distinct genetic clusters, more strongly differentiated than previously thought based on microsatellites. Since strong genetic structure can bias tests for positive selection, we performed selection tests on each cluster separately. We ran genomic scans for positive selection using Tajima's D and Fay and Wu's H standardized by Zeng (Z). We found potential positively selected region, including two regions containing genes coding for small secreted proteins such as and Major Facilitator proteins.

20. A modular pipeline for de novo transcriptome assembly
De novo assembly is nowadays routinely used to study the whole transcriptome from non-model organisms. This technique is indeed widely popular because it provides from massive amounts of high-throughput sequences, an initial set of transcripts without ever referring to pre-obtained genomic data. De novo assembly is however still challenging because it is computationally intensive, and recovering rare transcripts from a datasets with heterogeneous sequence coverage, require multiple k-mer strategy, with additional potential for mis-assembly To date, a multitude of de novo transcriptome assemblers are available, implementing the most efficient method for short read de novo assembly, the so-called de Bruijn graphs method. Although both cloud computing and multiple bioinformatics tools are available, they have not been employed as broadly as reference-based pipelines because programming knowledge is required.

We therefore propose the establishment of a pipeline implemented in Python for processing RNAseq data from non-model organisms. Our pipeline aims to construct a user-friendly de novo transcriptome assembly pipeline that compares the output of multiple programs and automatically analyses this data for downstream applications. Our pipeline uses Velvet-Oases and Trinity for the initial assembly, then performs transcripts metrics comparisons, and finally fuses the transcripts using GAM-NGS if significant metrics improvement is detected. Quality transcripts obtained are then used for various analyses with downstream programs, including BLAST and Blast2GO for transcripts identification, DESeq2 for differential expression analysis, etc.. Our pipeline has been built to run on a low-cost cloud-computing network, and will be usable from a standard desktop computer. The implementation of our pipeline, mostly modular, allows also adaptation to further analyses to answer any potential biological question of the future users.

21. Incorporating biological a priori in Gene Regulatory Networks Inference using Graph Cuts

Pirayre Aurélie, Couprie Camille, Bidard-Michelot Frédérique, Duval Laurent and Pesquet Jean-Christophe
IFP Energies nouvelles, France

Background: Inferring gene networks from high-throughput data is useful to visualize a sketch of regulatory relationships. Many inference methods ignore biological a priori.

Method: Weighting all possible pairwise gene relationships by the probability of edge presence, we formulate the regulatory network inference as an optimization problem. The optimal edge labeling minimizes an objective function that takes into account three terms. The first term favors the presence of strong weights, while the two other terms enforce biologically plausible a priori between groups of genes, e.g. promoting a higher connectivity of the transcription factors. The optimization is performed by the popular Graph Cuts method: the construction of an intermediate flow network allows us to compute the optimal binary labeling corresponding to a minimum cut in the graph.

Results: Our Graph Cuts approach uses weights defined by the mutual-information-based Context Likelihood Relatedness (CLR) method to decipher the Escherichia coli network. We compare the inferred networks to the state-of-the-
art Genie3, winner of the DREAM4 challenge. Simulations are performed on real data from a compendium of Escherichia coli. We used the RegulonDB database to evaluate the inference performance computing the Area under Precision-Recall curve (APR). Results show an APR of 5.0 for CLR, 5.1 for Genie3 and reach 6.4 with our approach. Up to 20 additional verified interactions are obtained over Genie3 in a 178-genes network. A reduction of the number of false positives additionally suggests a higher predictive power for larger graphs. For 4345 genes, our approach only requires less than 1/100th of the time spent in weight computation with CLR or Genie3.

Conclusion: Using network flow theory for network inference, we significantly improve the state-of-the-art. Our generic approach and fast Graph Cuts algorithm make it possible to improve results of other methods by using their weights as input.

22. 3D Modeling of Group I Intron structures by comparative modeling with ModeRNA and de novo RNA folding with SimRNA

Deepak Kumar, Michael Boniecki and Janusz Bujnicki

Laboratory of Structural Bioinformatics, Institute of Molecular Biology and Biotechnology, Collegium Biologicum, Adam Mickiewicz University, Poland

Group I introns is a family of widespread non-coding RNA molecules well known for self-splicing from the host precursor RNA. Thus far only Azoarcus, Tetrahymena and Twort are known and well studied structures of this family. Commonly, group I introns are classified into 14 subfamilies based on conserved core sequences and peripheral structures. However, introns from particular groups have high length diversity and weak sequence similarity, which makes structure prediction for these RNAs very difficult.

To provide 3D structural models of representatives of group I introns from all families, we used a combination of comparative and de novo RNA structure modeling. We manually prepared alignment of 11 representatives (from subfamilies with unknown structures) with sequences and structures of representatives with known structures. ModeRNA software was used to generate initial models of group I intron core structures by a comparative modeling approach. We defined the structural core based on the available secondary and tertiary structures, P4-P6 domain containing P4, P5 and P6 and P3-P9 domain containing P3, P7, P8 and P9. Azoarcus, Tetrahymena and Twort sharing significant similarity and common secondary structure with the representatives were chosen as templates for modeling. Fragments of models without counterparts in templates were then added and folded with a de novo modeling approach, as implemented in the SimRNA method (Boniecki, Bujnicki, and coworkers, manuscript in preparation).

The generated models of group I intron structures accurately depict the global topology, secondary and tertiary interactions. Expectedly, the accuracy is highest in the core, with RMSD between 3-4 Å, whereas deviations are larger for peripheral regions that differ substantially between different introns. The results of this analysis provide a 3D perspective for studying group I introns and for interpretation of their sequence evolution in a structural context.

23. In-Silico molecular docking study of Pyrazolone derivatives: Discovery of new inhibitors of DNA Topoisomerase-I
Syed Asif Husain Naqvi, Esraa Bassam Taha, Anu Gupta, Priya Tomar, Deepshikha Agarwal, Vijay Mahida, Prashant Pandey, Kavitha Raj. V and Zain Taha

BioDiscovery Group, India

DNA Topoisomerase I releases the supercoiling and torsional tension of DNA introduced during the DNA replication and transcription by transiently cleaving and rejoining one strand of the DNA duplex. It introduces a single-strand break via transesterification at a target site in duplex DNA. Due to this reason it is the molecular target of a diverse set of anticancer compounds. These compounds bind to a transient top1-DNA covalent complex and inhibit the resealing of a single-strand nick that the enzyme creates to relieve superhelical tension in duplex DNA.

We have performed a Molecular Modeling study that comprised of Virtual screening and Molecular Docking of around 5000 molecules all of structure similarity of pyrazolone. Molecular docking approach using Lamarckian Genetic Algorithm was carried out to find out the potent inhibitors for DNA TOPOISOMERASE I on the basis of calculated ligand-protein pairwise interaction energies. The grid maps representing the protein were calculated using auto grid and grid size was set to 60*60*60 points with grid spacing of 0.375 Å. Docking was carried out with standard docking protocol on the basis of a population size of 150 randomly placed individuals; a maximum number of 2.5 *107 energy evaluations, a mutation rate of 0.02, a crossover rate of 0.80 and an elitism value of 1. Ten independent docking runs were carried out for each ligand and results were clustered according to the 1.0 Å rmsd criteria. The docking result of the study of 5000 molecules demonstrated that the binding energies were in the range of -1.47 kcal/mol to -6.68 kcal/mol with the minimum binding energy of -6.68 kcal/mol. Good number of molecules showed H-Bond interaction with residues of binding pocket around Asp-533 on, Arg-364, Gln-421, Lys-374 and they also showed promising ADMET properties. Further in-vitro and in-vivo study is required on these molecules as the binding mode provided hints for the future design of new inhibitors for DNA TOPOISOMERASE I.

24. Adding Structural Resolution to Edges in Protein-Protein Interaction Networks to Distinguish Driver and Passenger Mutations in Cancers

Jakob Berg Jespersen
Broad Institute, USA

During the last decade, we have had revolution in cancer research, thousands of tumors have been collected at Hospitals, and their genomes have been sequenced. By comparing the genome of the cancer to the genome of the patient, it is possible to discover differences, which are unique for the cancer cells. There is a possibility that these differences are mutations that have occurred, and are responsible for driving the disease in the tumor.

Cancer is a disease where biological networks have been perturbed, due to many mutations. In order to try to understand how biological networks are affected by mutations, we have decided to focus on deleterious single nucleotide variants (dSNVs) causing missense mutations, occurring in protein-protein interacting sites, for protein pairs with existing three dimensional structures.

It was possible with statistical significance to find existing cancer driver genes, that was classified in the COSMIC census database.
25. Model-based Generation of CRC Screening Guidelines

Brian Lang and Niko Beerewinkel
D-BSSE ETH Zurich, Switzerland

Colorectal cancer, CRC, can be avoided entirely by many at-risk patients with adherence to screening guidelines. These guidelines dictate time points at which an individual should undergo a screening colonoscopy. If completed at the correct time for the patient, these colonoscopies can detect and remove pre-cancerous polyps, thus reducing CRC incidence in the population. While CRC incidence reduction is a clear goal, health economics encourages us to also consider the concept of life years gained through early screening as well as the cost of a given set of screening guidelines. To this end we derive analytically an expression for the expected time needed for a clonal population to reach a given size. We derive numerically the expected waiting time to cancer from a clonal initiation event. We fit a four-stage clonal expansion model to the Surveillance Epidemiology and End Results, SEER, data (1973-2010) to gain age-specific rates of CRC incidence. We then use these three derivations to examine cost-optimal screening strategies and make our recommendations.

26. Extension of Aho-Corasick algorithm to match biological motif of variable gaps

Bilal Lounnas and Moussaoui Abdelouahab
Department of Computer Science, M'sila University, Algeria

An enormous number of motif discovery algorithms have been proposed in the literature in the past decade due to the importance of this operation in computer science in general and bioinformatics in particular. Located motifs within biological sequence are very crucial, from this perspective point of view we proposed an extension of Aho-Corasick (AC) algorithm, one of the most powerful and well know pattern matching algorithm that shows a significance progress in the field of pattern matching. The proposed extension is to adapt AC algorithm to detect biological motif by overriding two of its based function, the first one is responsible of building a keyword tree from a set of patterns and applied a modified failure function in order to build relations between these keywords based on biological patterns roles. While the second function is responsible for matching process, in this function we used substitution matrix to obtain a probabilistic pattern matching score. As a result of this study, we perform a comparisons based on robustness, accuracy, and run-time between our proposed algorithm, AC algorithm, and KMP algorithm. We found that our extension can outperform the AC algorithm especially in the context of biological motif discovery, it also outperform in terms of accuracy and run-time against KMP algorithm, and AC algorithm.
Flash Poster Presentation

This year, for the first time, the European Student Council Symposium provides the unique opportunity to present outstanding posters to fellow researchers. This session will consist of a series of selected flash poster presentations.

The outstanding posters of the Symposium will be chosen through voting by all delegates at the Symposium. The top five posters will have the opportunity to be presented in the 5 minutes flash presentation. The presentation includes a brief exposition of the poster topic that highlights the most important facts (3 minutes) and a few short questions (2 minutes).
Awards

The outstanding posters and oral presentations of the 3rd European Student Council Symposium will be recognized and awarded with the support of our sponsors.

Best Presentation Award

This award will acknowledge the Best Oral Presentation by a student at the Symposium. All Symposium delegates will be asked to score their favorite presentations. The winner will be determined by the vote.

In addition to the Best Presentation Award, a Best Presentation Runner-Up Award will also be given out.

Best Poster Award

The best poster of the Symposium will be chosen through voting by the delegates at the Symposium. The top five posters will be presented in the Flash poster presentation session. The jury of Student Council Leaders will select the two best posters among them based on the presentation.

Travel Fellowships

The ISCB Student Council has teamed up with this year’s sponsors to give several students the opportunity to attend the 3rd European Student Council Symposium in Strasbourg, France. Thanks to generous support from our sponsors, we were able to award 4 Travel Fellowships this year.

Congratulations to all the Travel Fellowship awardees!
Acknowledgements

The success of an event the size of the European Student Council Symposium depends on the commitment of many. We would like to thank everyone involved in the organization this year for their contribution, be it a 15-minute job or months of work. For some efforts we are extraordinarily grateful and they deserve to be mentioned explicitly:

We are greatly indebted to ECCB 2014 conference chairs Marie-Dominique Devignes and Yves Moreau for giving us the opportunity to have the 3rd European Student Council Symposium in Strasbourg. We are especially thankful for the logistical support and invaluable advice of the ECCB organizing committee; specifically the Workshops and Tutorials chairs Olivier Poch and Mario Albrecht, and our ECCB-intermediary Magali Michaut. We deeply appreciate their continued support of the ISCB Student Council and the Symposium.

Further, we would like to acknowledge the support of the ISCB Board of Directors and their trust in our vision. The Student Council would also like to thank our keynote speakers; Dr. Martens, Dr. de Ridder and Dr. Jensen. They are all very busy people, yet they are volunteering their time to contribute to the success of the Symposium and to promote the next generation of computational biologists.

Furthermore, we would like to thank everyone on the organizing committee, without them, there would be no Symposium! There’s a lot of work that goes unseen behind the scenes when organizing a symposium. We are extremely lucky to have such a talented and dedicated team this year and very much appreciate their contributions.

We are also extremely grateful for the financial support that we received from our sponsors. Without their help many of the exciting opportunities that we offer to the delegates at the 3rd European Student Council Symposium would not have been possible.

Thank you all!
Regional Student Groups Initiative

The ISCB Student Council (SC) has always strived to reach out to Students of Computational Biology and Bioinformatics around the world and promote communication between them to create a vibrant global network of peers. To accomplish this more effectively, in 2006 the SC conceptualized the setting up of Regional Student Groups (RSGs). Regional Student Groups work to fulfill the broad mission of the SC at their regional level by organizing events and initiatives tailored to the requirements of the local student community.

The RSGs initiative has turned out to be an extremely popular and successful initiative. In the past six years, the RSG network has grown to include twenty RSGs from all over the world. Our active RSG network has seen RSGs organize symposia, conduct workshops and contests, initiate discussion groups and even work with each other on trans-national collaborative student projects. As supra-institutional organizations, RSGs are perfectly placed to foster inter-institutional contacts and collaborations in their region and where possible, even serve as a link between students and the local industry. Most RSGs have also formed their own network of members using mailing lists, discussion forums or other means to ensure quick and efficient dissemination of useful information within the community.

The minimal leadership team required to run an RSG are a President and a Secretary working under the guidance of a Faculty Advisor. Since the RSGs are affiliated to the SC membership to an RSG is free. Only the President, Secretary and the Faculty Advisors are required to hold an ISCB membership. Individual RSGs are of course free to put in place a more elaborate administration team if needed. This uncomplicated administrative structure and low operating costs associated with the RSGs has made it feasible for students in many developing countries to begin and develop RSGs in their countries.

As recognition of the importance of the RSGs to the Student Council’s overall mission, the RSGs funding program was initiated in July 2010, thanks to funding support by the ISCB. As a part of this program, RSGs are invited to submit proposal for events and initiatives they plan to organize and after a peer review process some of those proposals are selected to be funded by the SC. So far, RSGs have utilized these funds to organize workshops, hackathons, discussion groups and more. Visit http://iscbsc.org/node/65/rsg-funding for more details about the funding program.
Snapshots from RSG events organized with funding support from the SC

The success of the RSGs initiative is due only to the enthusiasm and commitment shown by the RSG leaders and the support that they have received from faculty advisors and other interested professors. And with these motivated students leading our RSGs, we only expect to see this initiative grow from strength to strength in the coming days.

If you would like to find out more about the RSGs initiative or find out how you too can get involved in this, please visit http://iscbsc.org/content/regional-student-groups or send an email to rsg@iscbsc.org
# List of Participants

<table>
<thead>
<tr>
<th>Name</th>
<th>e-mail</th>
</tr>
</thead>
<tbody>
<tr>
<td>Armenta medina</td>
<td><a href="mailto:armentamedina@gmail.com">armentamedina@gmail.com</a></td>
</tr>
<tr>
<td>Babaei</td>
<td><a href="mailto:s.babaei@tudelft.nl">s.babaei@tudelft.nl</a></td>
</tr>
<tr>
<td>Badouin</td>
<td><a href="mailto:helene.badouin@u-psud.fr">helene.badouin@u-psud.fr</a></td>
</tr>
<tr>
<td>Borrel</td>
<td><a href="mailto:alexandre.borrel@helsinki.fi">alexandre.borrel@helsinki.fi</a></td>
</tr>
<tr>
<td>Cortes</td>
<td><a href="mailto:isidrolauscher@gmail.com">isidrolauscher@gmail.com</a></td>
</tr>
<tr>
<td>Coudrat</td>
<td><a href="mailto:thomas.coudrat@monash.edu">thomas.coudrat@monash.edu</a></td>
</tr>
<tr>
<td>Döring</td>
<td><a href="mailto:mdoering@mpi-inf.mpg.de">mdoering@mpi-inf.mpg.de</a></td>
</tr>
<tr>
<td>Francescatto</td>
<td><a href="mailto:margherita.francescatto@dzne.de">margherita.francescatto@dzne.de</a></td>
</tr>
<tr>
<td>Hermans</td>
<td><a href="mailto:s.hermans@radboudumc.nl">s.hermans@radboudumc.nl</a></td>
</tr>
<tr>
<td>Jespersen</td>
<td><a href="mailto:jbjesper@broadinstitute.org">jbjesper@broadinstitute.org</a></td>
</tr>
<tr>
<td>Kazlauskas</td>
<td><a href="mailto:d.kazlauskas@ibt.lt">d.kazlauskas@ibt.lt</a></td>
</tr>
<tr>
<td>Kuenemann</td>
<td><a href="mailto:melaine.kuenemann@univ-paris-diderot.fr">melaine.kuenemann@univ-paris-diderot.fr</a></td>
</tr>
<tr>
<td>Kumar</td>
<td><a href="mailto:deepak.k.choubey@gmail.com">deepak.k.choubey@gmail.com</a></td>
</tr>
<tr>
<td>Lang</td>
<td><a href="mailto:brianlang@bsse.ethz.ch">brianlang@bsse.ethz.ch</a></td>
</tr>
<tr>
<td>Mahé</td>
<td><a href="mailto:mahe@rhrk.uni-kl.de">mahe@rhrk.uni-kl.de</a></td>
</tr>
<tr>
<td>Mahfouz</td>
<td><a href="mailto:a.mahfouz@tudelft.nl">a.mahfouz@tudelft.nl</a></td>
</tr>
<tr>
<td>Meng</td>
<td><a href="mailto:arnaud.meng@gmail.com">arnaud.meng@gmail.com</a></td>
</tr>
<tr>
<td>Meysman</td>
<td><a href="mailto:pieter.meysman@uantwerpen.be">pieter.meysman@uantwerpen.be</a></td>
</tr>
<tr>
<td>Narunsky</td>
<td><a href="mailto:ayanarun@post.tau.ac.il">ayanarun@post.tau.ac.il</a></td>
</tr>
<tr>
<td>Naulaerts</td>
<td><a href="mailto:stefan.naulaerts@uantwerpen.be">stefan.naulaerts@uantwerpen.be</a></td>
</tr>
<tr>
<td>Olechnovic</td>
<td><a href="mailto:kliment@ibt.lt">kliment@ibt.lt</a></td>
</tr>
<tr>
<td>Peres</td>
<td><a href="mailto:amelie.peres78@gmail.com">amelie.peres78@gmail.com</a></td>
</tr>
<tr>
<td>Pirayre</td>
<td><a href="mailto:aurelie.pirayre@ifpen.fr">aurelie.pirayre@ifpen.fr</a></td>
</tr>
<tr>
<td>Rot</td>
<td><a href="mailto:gregor.rot@uzh.ch">gregor.rot@uzh.ch</a></td>
</tr>
<tr>
<td>Timinskas</td>
<td><a href="mailto:kestutis.timinskas@bti.vu.lt">kestutis.timinskas@bti.vu.lt</a></td>
</tr>
<tr>
<td>Verstraete</td>
<td><a href="mailto:verstraete.nina@gmail.com">verstraete.nina@gmail.com</a></td>
</tr>
</tbody>
</table>
Organizing Committee

**Pieter Meysman, Chair**  
Department of Mathematics and Computer Science  
University of Antwerp, Belgium

**Margherita Francescatto, Co-Chair**  
Genome Biology for Neurodegenerative Diseases  
German Center for Neurodegenerative Diseases (DZNE) - Tübingen, Germany

**Esmeralda Vicedo, Finance Chair**  
Department for Bioinformatics and Computational Biology  
Technical University Munich, Germany

**Susanne Hermans, Program Chair**  
Chemical Discovery and Design, Centre for Molecular and Biomolecular Informatics  
Radboud University Medical Centre, Nijmegen, The Netherlands

**Sepideh Babaei, Booklet Chair**  
Bioinformatics Lab, Pattern Recognition Group  
Delft University of Technology, The Netherlands

**Anni Jacobsen, Communications Chair**  
IBIVU Centre for Integrative Bioinformatics  
VU University Amsterdam, The Netherlands

**Nina Verstraete, Organization Team**  
Protein Physiology Laboratory, Department of Biological Chemistry  
University of Buenos Aires - CONICET, Argentina

**Alexandre Borrel, Organisation Team**  
INSERM UMR S973, Molécules Thérapeutiques In Silico - University of Paris Diderot, France  
Computational Drug Discovery group, Faculty of Pharmacy - University of Helsinki, Finland

**Mélaine Kuenemann, Organisation Team**  
INSERM UMR S973, Molécules Thérapeutiques In Silico  
University of Paris Diderot, France

**Dan DeBlasio, Website Committee**  
Department of Computer Science  
University of Arizona, USA

**Bilal Lounnas, Design Committee**  
Department of Computer Science  
M'sila University, Algeria

Disclaimer
The ISCB Student Council has made all efforts to provide accurate information but does not guarantee the correctness of any information provided in this booklet. The ISCB Student Council is a committee of the International Society for Computational Biology (ISCB), which is incorporated as a 501(c)(3) non-profit corporation in the United States.

Copyright © 2014 European Student Council and contributing authors. All rights reserved. This booklet may be reproduced without permission in its original form.
Social Event

We wish to invite all Symposium delegates to take part in our Social Event evening. It will be a focused event of drinks, new friends and a fun social activity. The evening program provides you the opportunity to network with other young researchers.

On the evening, following the Symposium's activities for the day, you can relax on a leisurely walk through the center of Strasbourg. Following, drinks and the social activity will be provided at Académie de la Bière.

Please join us for this Social Event.

We will leave for the walk:
When: 18:30
Where: Conference Centre (STRASBOURG CONVENTION CENTRE)

The Event details:
When: 19.00 - Onwards
Where: Académie de la Bière, 17 Rue Adolphe Seyboth, 67000 Strasbourg