21st Annual International Conference
on Intelligent Systems for Molecular Biology

12th European Conference
on Computational Biology

An Official Conference of
the International Society
for Computational Biology

2013 ISMB ECCB

STUDENT COUNCIL SYMPOSIUM 9

JULY 19, 2013

www.iscb.org/ismbeccb2013

CONFERENCE
JULY 21–23

SIGS & TUTORIALS
JULY 19–20

9th ISCB Student Council Symposium

JULY 19, 2013

An Official Conference of
the International Society
for Computational Biology

www.iscb.org/ismbeccb2013

Contents

Welcome to the 9th ISCB Student Council Symposium!

The ISCB Student Council is pleased to welcome you to the 9th International Society for Computational Biologists Student Council Symposium in Berlin, Germany. Our previous Symposia in Long Beach (2012), Vienna (2011), Boston (2010), Stockholm (2009), Toronto (2008), Vienna (2007), Fortaleza (2006) and Madrid (2005) were successful meetings. We are therefore thrilled to continue our efforts in Berlin 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. This year we secured funds to organize the Symposium, give poster and oral presentation awards and, most importantly, offered nine travel fellowships for Symposium delegates.

We are honored to have **Dr. Gonçalo Abecasis (University of Michigan)**, **Dr**. **Alex Bateman** (**EMBL-European Bioinformatics Institute**), and **Dr. Satoru Miyano, (Institute of Medical Science of the University of Tokyo**) 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 your peers in an informal and friendly way. Throughout the day we will hear **oral presentations** from a selection of 10 outstanding student abstracts spanning a wide-range of research areas. In the evening, the **poster session** will offer exciting science in various domains, and give everybody a chance to discuss their research topics in more depth.

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 has worked hard to raise funds with the intention of making the 9th Student Council Symposium even better than before. We wish to continue this unique and well-received event as judged by registration and submission numbers.

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 check out other Student Council events during ISMB2013, such as the **Career Central/Student Council Lounge (Booth 1)**, our **Open Business Meeting**, the **Art and Science Exhibition** and, of course, our **Social Events**.

Enjoy your time in Berlin!

Tomás Di Domenico, Student Council Symposium Chair

Tomasz Stokowy, Student Council Symposium Co-Chair

Anupama Jigisha, Student Council Chair

Note: This booklet went into print in late June. Please check http://symposium.iscbsc.org for the latest updates and announcements.

Agenda

Time	Event/Activity	Details on Page
08:30	Registration	
08:45	Icebreaker: Scientific Speed Dating	*10*
09:15	Welcome	
09:30	**Keynote:** *Not just stamp collections: Molecular Biology Databases in the 21st century* **Dr. Alex Bateman**, EMBL-EBI	*11*
10:20	Coffee Break	
10:40	*Oral Presentations* **Session I** ***An integrated approach to understanding apicomplexan metabolism from their genomes*** *Achchuthan Shanmugasundram,* University of Liverpool, United Kingdom ***Inferring the yeast salt stress response subnetwork from diverse, complementary data sources using integer linear programming*** *Deborah Chasman,* University of Wisconsin-Madison, USA ***Nonsense-Mediated Decay is a Major Transcriptome Regulator*** *André Kahles,* Memorial Sloan-Kettering Cancer Center, USA ***Towards breaking the curse of dimensionality in computational methods for the conformational analysis of molecules*** *Han Cheng Lie,* Freie Universitaet Berlin, Germany	*16*
12:00	*Workshop* *Presenting your science visually* **Dr. Thomas Abeel**, Broad Institute of MIT and Harvard	
12:30	Lunch break + Posters Session I	
13:30	***Keynote:*** *Cancer Gene Network Analysis with Supercomputer* **Dr. Satoru Miyano**, University of Tokyo	*12*
14:20	*Oral Presentations* **Session II**	*19*

On the expansion of dangerous gene families in vertebrates
Severine Affeldt, Institut Curie, France

Systems Level Analysis of Breast Cancer Reveals the Differences between Lung and Brain Metastasis through Protein-Protein Interactions
Billur Engin, Koc University, Turkey

ConFind: Exploiting the protein databank to propose additional conformations of a query protein of known structure
Aya Narunsky, Tel Aviv University, Israel

15:20 *Research and Industry Partners presentation* *14*
Dr. Cheng Soon Ong , NICTA

15:35 *Coffee break + Posters Session II*

16:00 ***Keynote:*** *Computational Biology and Human Genetics* *13*
Dr. Gonçalo Abecasis, University of Michigan

16:45 *Oral Presentations* *21*
Session III

alleHap: An efficient algorithm to construct zero-recombinant haplotypes from parentoffspring pedigrees
Nathan Medina Rodriguez, Universidad de Las Palmas de Gran Canaria, Spain

ancGWAS: Graph-based Approach for Analyzing Sub-networks Underlying Ethnic Differences in Complex Disease Risk in Admixed Populations
Emile R. Chimusa, University of Cape Town, South Africa

Oqtans: A Multifunctional Workbench for RNA-seq Data Analysis
Vipin T. Sreedharan, Memorial Sloan-Kettering Cancer Center, USA

17:50 Closing remarks + Posters Session III

19:00 Symposium Ends

19:30 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 ISMB? 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

Alex Bateman

EMBL-European Bioinformatics Institute

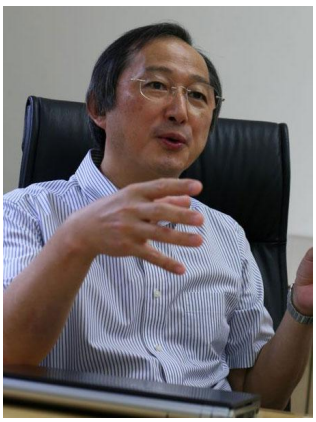
Dr. Alex Bateman is head of Protein Sequence Resources at European Bioinformatics Institute. Pfam/Rfam generalissimo and Wikipedia evangelist". Apart from being the leader of some of the most important databases and data curation efforts, Dr. Bateman is actively involved in bridging the gap between scientists and the public, particularly through the use of tools such as Wikipedia. His address will surely be of great interest and value to students and young researchers.

Not just stamp collections: Molecular Biology Databases in the 21st century.

There are now thousands of biological databases available covering almost every aspect of molecular and computational biology. Some of these have become so important and all pervasive that we scarcely even notice them any more, except when they go wrong. I have watched this growth for the past twenty years and have been involved in producing some popular databases such as Pfam and miRBase and now I am also leading the UniProt database. None of these would have been possible without the hard work and dedication of Biocurators. To me Biocurators are the unsung heroes of biology who order and make sense of the immense primary literature for us. The International Society of Biocuration has given this disparate group of researchers a forum to exchange the latest developments and a stronger voice on the issues they care about.

Satoru Miyano

Institute of Medical Science of the University of Tokyo



Dr. Satoru Miyano is a Professor of the Human Genome Center at the Institute of Medical Science of the University of Tokyo. His mission is to create computational strategies for systems biology and medicine towards translational bioinformatics based on the recent advances in biomedical research that have been producing large-scale, ultra-high dimensional, ultra-heterogeneous data.

Cancer Gene Network Analysis with Supercomputer

Cancer is a very complex disease that occurs from accumulation of multiple genetic and epigenetic changes in individuals who carry different genetic backgrounds and have suffered from distinct carcinogen exposures. These changes affect various pathways which are necessary for normal biological activities and gene networks are driving these pathways in disorder in the center. We present our computational methods and their analyses in Cancer Systems Biology that use the supercomputer system at Human Genome Center of The University of Tokyo (225 TFLOPS, 4.4PB storage). A big challenge is to development of a systematic methodology for unraveling gene networks and their diversity lying over genetic variations, mutations, environments and diseases. We present our challenge for uncovering systems in cancer by supercomputer from gene expression profiles. NetworkProfiler is a method that will exhibit how gene networks vary from patient to patient according to a modulator, which is any score representing characteristics of cells, e.g. survival. First we defined an EMT (epithelial-mesenchymal transition) modulator and analyzed gene expression profiles of 762 cancer cell lines. Network analysis unraveled global changes of networks with 13,508 genes of different EMT levels. By focusing on E-cadherin, 24 genes were predicted as its regulator, of which 12 have been reported in the literature. A novel EMT regulator KLF5 was also discovered in this study. We also analyzed Erlotinib resistant networks using 160 NSCLCs with GI50 as a modulator. Hubness analysis exhibited that NKX2-1/TTF-1 is the key gene for Erlotinib resistance in NSCLCs. Our microRNA/mRNA gene network analysis with Bayesian network method called SiGN-BN also revealed subnetworks with hub genes (including NKX2-1/TTF-1) that may switch cancer survival. For dynamic system modeling, we devised a state space model (SSM) with dimension reduction method for reverse-engineering gene networks from time-course data, with which we can view their dynamic changes over time by simulation. We succeeded in computing a gene network with prediction ability focused on 1500 genes from data of about 20 time-points. We applied this SSM model to human normal lung cell treated with (case)/without (control) Gefitinib, and we identified genes under differential regulations between case and control. This signature of genes was used to predict prognosis for lung cancer patients and showed a good performance for survival prediction of stage 1 patients that has been considered very difficult. On-going cancer research using K-computer (10PFLOPS supercomputer in Kobe, Japan) is also introduced.

Gonçalo Abecasis

University of Michigan

Dr. Goncalo Abecasis is a Professor of Biostatistics. He received his D.Phil. in Human Genetics from the University of Oxford in 2001 and joined the faculty at the University of Michigan in the same year. Dr. Abecasis' research focuses on the development of statistical tools for the identification and study of genetic variants important in human disease. Software developed by Dr. Abecasis at the University of Michigan is used in several hundred gene-mapping projects around the world.

Computational Biology and Human Genetics

During the past 10 years, human genetics has progressed at a rapid pace - aided not just by improvements in laboratory technology and the available of large population samples for study, but also by the creative and key contributions by computational biologists. I will review challenges and opportunities in human genetic research, with an emphasis on the important contributions that young computational biologists can make.

Research and Industry Partners Session

Dr Cheng Soon Ong

Victoria Research Laboratory, National Information and Communications Technology Australia

Dr Cheng Soon Ong is Senior researcher at the Bioinformatics Group, National ICT Australia, located at theUniversity of Melbourne.He is interested in enabling scientific discovery by extending statistical machine learning methods. In recent years, he has developed new optimization methods for solving problems such as feature selection and structured output prediction, with the aim of solving scientific questions in collaboration with experts in other fields. This has included diverse problems in genomics, systems biology, fMRI analysis and medical imaging. He is also interested in open source software and reproducible research in the context of machine learning. His PhD work on kernel methods was completed at the Australian National University in 2005. He had a short postdoc stint at the Statistical Machine Learning Group, in NICTA, Canberra, followed by a longer one at at both the Max Planck Institute of Biological Cybernetics and the Fredrich Miescher Laboratory in Tübingen, Germany.

About the Talk

NICTA Bioinformatics Research Group develops novel statistical methods that lead to efficient computational tools for systematically analyzing large scale biomedical data. In close collaboration with medical specialists and pathologists, they develop software tools that extract timely information from the sea of data. The talk will provide an example of one of their up and coming projects.

Genome-wide association studies (GWAS) genotype a large number of single nucleotide polymorphisms (SNPs) and attempt to determine their association with a given phenotype. Typically, this analysis examines the association of each SNP individually. Yet many common diseases have complex aetiologies that involve combinations of SNPs from different genes and different combinations within the population of affected individuals. Recent progress on detecting epistatic interaction using log linear models (Wan et. al., AJHG, 2010) and ROC-based strategies (Goudey et. al, BMC Genomics, 2013) have enabled exhaustive search for such combinations. The availability of such tools allows researchers to efficiently filter for putative pairs of SNPs that are strongly associated with disease, leading to new challenges. These include: tools for interactive visualization of the discovered interactions to enable human curation and investigation of the putative pairs; algorithms for quantifying replicability and stability of the ranked SNP pairs; and techniques for linking putative SNP pairs to gene interaction networks and other biological pathways

Oral Presentations

Session I

1. An integrated approach to understanding apicomplexan metabolism from their genomes

Achchuthan Shanmugasundram, Faviel Gonzalez-Galarza, Jonathan Wastling, Olga Vasieva, Andrew Jones
University of Liverpool,United Kingdom

The Apicomplexa is a large phylum of intracellular parasites that show great diversity and adaptability in the various ecological niches they occupy. They are causative agents of human and animal infections including malaria, toxoplasmosis and theileriosis and have a huge economic and social impact. In order to develop new drugs and vaccines it is essential to have a complete understanding of the metabolism and host-parasite relationships of these pathogens. A number of apicomplexan genomes have been sequenced and are publicly available. However, the prediction of gene models and annotation of gene function remains challenging. We have utilised an approach called 'metabolic reconstruction', in which genes are systematically assigned to functions within pathways/networks. Functional annotation and metabolic reconstruction was carried out using a semi-automatic approach, integrating genomic information with biochemical evidence from the literature. The functions were automatically assigned using a sequence similarity-based approach and protein motif information. Experimental evidence was also accommodated in the confirmation of functions and the grouping of genes into metabolic pathways. The functions required to complete metabolic pathways, and that are missing in gene models, were also identified. A web database named Library of Apicomplexan Metabolic Pathways (LAMP, http://www.llamp.net) was developed and contains the near complete mapping of genes to metabolic functions for Toxoplasma gondii, Neospora caninum, Cryptosporidium and Theileria species and Babesia bovis. Each metabolic pathway page contains an interactive metabolic pathway map, gene annotations hyperlinked to external resources and detailed information about the metabolic capabilities. We also carried out a comparative analysis of the overall metabolic capabilities of apicomplexan species in terms of their ability to synthesise or dependence on the host for a metabolite. The comparative analysis of the metabolism of apicomplexans also provides clues about their life cycles. This database provides complete metabolism of apicomplexan parasites and leads to identification of putative drug targets. The metabolic pathway annotations of LAMP have been linked out from the respective gene pages of T. gondii primary database, ToxoDB (release 8). We expect the LAMP database will become a valuable resource for the Apicomplexa community both for fundamental and applied research.

2. Inferring the yeast salt stress response subnetwork from diverse, complementary data sources using integer linear programming

Deborah Chasman, David Berry, James Hose, Anna Merrill, M. Violet Lee, Elisha Yi-Hsuan Ho, Josh Coon, Audrey Gasch, Mark Craven
University of Wisconsin-Madison, USA

To adapt to environmental stress, Saccharomyces cerevisiae undergoes widespread changes in gene regulation. While some key signaling proteins have been identified and characterized, the complex signaling network that coordinates these changes is incompletely understood. We present an integer linear programming-based approach to distill large-scale experimental data into predicted signaling pathways that control the yeast salt stress response. Our approach takes as input experimental data that examines the salt response from multiple, complementary perspectives: single gene mutants that confer a defect in stress resistance, proteins with an altered phosphorylation state during stress, and genes whose salt-responsive expression is dysregulated in fifteen single gene mutants corresponding to known regulators. Using these data, a background network of publicly

available protein-protein, protein-DNA, and protein-RNA interactions, and an integer linear program, we infer a subnetwork that provides directed pathways through which the fifteen signaling regulator mutants perturb the regulation of their affected downstream targets. While previous methods for signaling network inference have addressed the task of inferring a directed subnetwork from expression profiles, our method also exploits additional sources of condition-specific experimental data. The objective function maximizes the inclusion of genes and proteins identified by fitness and phosphoproteomic assays while minimizing the total number of proteins in the inferred subnetwork. The resulting inferred subnetwork includes known or likely regulators of the salt response. Additionally, we predict the involvement of additional proteins that have not been previously annotated as being instrumental in the salt response; these are promising candidates for future experimental work.

3. Nonsense-Mediated Decay is a Major Transcriptome Regulator

André Kahles, Gabriele Drechsel, Philipp Drewe, Jonas Behr, Andreas Wachter, Gunnar Rätsch
Memorial Sloan-Kettering Cancer Center, USA

Nonsense-mediated mRNA decay (NMD) is a surveillance pathway in eukaryotes able to detect and target aberrant transcripts for degradation. NMD substrates typically arise from mutations, transcription errors, or differential transcript processing such as alternative splicing. Thus, NMD functions not only in the clearance of erroneous transcripts, but also plays an important role in regulating gene expression via degradation of mRNA variants. In order to identify NMD substrates on a transcriptome-wide level, we generated single and double mutants defective in the NMD factors UPF1 and UPF3 in *Arabidopsis thaliana*. The single mutants upf1 and upf3 show only minor developmental defects. The upf1upf3 double mutants are arrested in early development and show substantial accumulation of known NMD target transcripts. We sequenced the transcriptomes (Illumina) of the mutants as well as wildtypes treated with the translation inhibitor cycloheximide. We designed a computational pipeline to accurately identify novel alternative splicing events and developed a method for accurately detecting significant differential exon/intron usage combining and exploiting evidence from multiple samples taking biological variance into account. We found a surprisingly large number of differentially spliced events in the double mutants (3,361) and in cycloheximide-treated wildtypes (3,238) at a false discovery rate of 10%. A first conservative estimate suggests that at least 17.4% of protein-coding genes possess at least one NMD transcript variant, implying an important role for NMD in transcription regulation. We experimentally analyzed ten randomly selected cases with false discovery rate lower than 30% and could confirm differential isoform expression in nine cases. Intriguingly, 92.3% of the NMDresponsive mRNAs contain classical NMD-eliciting features, supporting their authenticity as direct targets. In summary, our study provides the first transcriptome-wide, splicing-sensitive analysis of NMD in plants and suggests an important and widespread role of this surveillance pathway in shaping the transcriptome. Further dissecting the features of this surveillance mechanism with respect to target characteristics and connecting them to different modes of operation is an important next step towards the systemic understanding of RNA processing-regulation. As NMD is a general to all eukaryotes, our pipeline can be readily applied to RNA-seq data of any organism. Further information is available under
http://www.raetschlab.org/suppl/nmdtools/nmdtools.

4. Towards breaking the curse of dimensionality in computational methods for the conformational analysis of molecules

Han Cheng Lie
Freie Universitaet Berlin, Germany

In computational molecular biology, the conformational analysis of a molecule refers to the identification of conformations - the clusters of molecular configurations sharing a

large-scale geometric structure - and the description of conformation dynamics. These data help describe the tructures and functions of molecules which are relevant in biology. In applications such as computational drug design, they help identify suitable compounds for disrupting processes involved in disease mechanisms. Today, many computational methods for conformational analysis are mesh-based: they involve discretizing configuration space using regular lattices, and simulating trajectories on energy landscapes. These methods suffer from the 'curse of dimensionality' in that the cost increases exponentially with the number of atoms in the molecule. Given that many important molecules in biology - such as proteins - have thousands of atoms, the curse of dimensionality severely limits the applicability of these methods. Finding an alternative approach to conformational analysis that does not suffer from the curse of dimensionality would enlarge the set of molecules and biological processes which may be studied using computational methods. We construct a mesh-free method for conformational analysis. We provide proof of concept by testing the method on a toy model of a molecule from the literature and find that the results of our method are comparable to those obtained by an exact, mesh-based method. We establish that the computational cost of our method increases only polynomially with the size of the molecule. We randomly sample points on the energy landscape and assign these to discretization regions. We compute the Gibbs-Boltzmann probability of each region by taking averages, and adjacency relations between discretization regions using linear programs. Applying a novel square root approximation to the probabilities and adjacencies yields a transition rate matrix, which gives a Markov Chain approximation of the continuous dynamics. Perron Cluster Analysis and coarsegraining yield the desired conformational data. Results from linear programming theory establish that our method does not suffer from the curse of dimensionality. Our findings demonstrate that one does not need meshes for conformational analysis and emphasize the utility of using different mathematical ideas in devising new computational methods.

Session II

5. On the expansion of dangerous gene families in vertebrates

Severine Affeldt, Herve Isambert, Param Priya Singh, Giulia Malaguti
Institut Curie, France

We report that the expansion of "dangerous" gene families, defined as prone to dominant deleterious mutations, can be traced to two rounds of whole genome duplication dating back from the onset of jawed vertebrates some 500MY ago. We argue that this striking expansion of "dangerous" gene families implicated in severe genetic diseases such as cancer is a consequence of their susceptibility to deleterious mutations and the purifying selection in post-whole-genome-duplication species.
Our data mining analyses, based on the 20, 506 human protein coding genes, first revealed a strong correlation between the retention of duplicates from whole genome duplication (so-called "ohnologs") and their susceptibility to dominant deleterious mutations in human. It appears that the human genes associated with the occurence of cancer and other genetic diseases (8, 095) have retained significantly more duplicates than expected by chance (48% versus 35%; 48% : 3, 844/8, 095; $P = 1.3 \times 10{-}128$, $\chi 2$). We also investigated an alternative hypothesis frequently invoked to account for the biased retention of ohnologs, namely the "dosage-balance" hypothesis. While this hypothesis posits that the ohnologs are retained because their interactions with protein partners require to maintain balanced expression levels throughout evolution, we found that most of the ohnologs have been eliminated from permanent complexes in human (7.5% versus 35%; 7.5% : 18/239; $P = 1.2 \times 10{-}18$, $\chi 2$). Our results also show that the gene susceptibility to deleterious mutations is more relevant than dosage balance for the retention of ohnologs in more transient complexes. To go beyond mere correlations, we performed mediation analyses, following the approach of Pearl, and quantified the direct and indirect effects of many genomic properties, such as essentiality, expression levels or divergence rates, on the retention of ohnologs. Our results demonstrate that the retention of human ohnologs is primarily caused by their susceptibility to deleterious mutations. All in all, this supports a non-adaptive evolutionary mechanism to account for the retention of ohnologs that hinges on the purifying selection against dominant deleterious mutations in post-whole-genome-duplication species. This is because all ohnologs have been initially acquired by speciation without the need to provide evolutionary benefit to be fixed in these populations.

6. Systems Level Analysis of Breast Cancer Reveals the Differences between Lung and Brain Metastasis through Protein-Protein Interactions

Billur Engin, Atilla Gursoy, Ozlem Keskin, Baldo Oliva, Emre Guney
Koc University, Turkey

According to American Cancer Society, breast cancer is the second most common cause of cancer death among women. Generally, the reason of fatality is the metastasis in another organ, not the primer tumor in the breast. A better understanding of the molecular mechanism of the metastatic process may help to improve the clinical methods for approaching the disease. For this purpose, we have used protein structure and protein networks together at the system level to explain genotype-phenotype relationship, and applied it to breast cancer metastasis. We have built a comprehensive human PPI network, by combining the available protein-protein interactions data from various databases. Then, we have ranked all the interactions of human PPI network according to their relevance to genes known to be mediating breast cancer to brain and lung metastasis. We have formed two distinct metastasis protein-protein interaction (PPI) networks from high ranked interactions. We have preformed functional analyses on brain/lung metastasis PPI networks and observed that the proteins of the lung metastasis network are also enriched in "Cancer", "Infectious Diseases" and "Immune System" KEGG pathways. This finding may be pointing to a cause and effect relationship between immune system-

infectious diseases and lung metastasis progression. We have enriched the metastasis PPI networks with structural information both with available data in Protein Databank and with our protein interface predictions. In the interface prediction step, the most common protein-protein interface templates in lung metastasis are observed to be coming from bacterial proteins. This finding reinforced our claim about the relationship between lung metastasis and infectious diseases. We have build two different breast cancer metastasis PPI networks. Besides we have also constructed structural PPI networks of the phenotypes. These network models may provide a foundation for future studies and may also be helpful for finding escape pathways of breast cancer metastasis.

7. ConFind: Exploiting the protein databank to propose additional conformations of a query protein of known structure

Aya Narunsky, Nir Ben-Tal
Tel Aviv University, Israel

Proteins often alternate between several conformations, e.g., active and inactive states of receptors, open and closed states of channels, etc. However, in many cases only one conformation is known. The detection of more (biologically-relevant) conformations of a protein would provide more insight on its function in health and disease. We present the ConFind computational tool for modeling putative conformation(s) of a query protein with one known conformation by assuming that pairs of structurally similar proteins may also share similar conformational changes. A three-step procedure is used: First, the protein databank is searched for structurally similar proteins to the query. Second, pairwise structural alignments are built between the query protein and each of the structurally similar proteins. Third, other known conformations of these proteins are indicated. By using the alignments found in the second step, and modeling on the structural templates found on the third step, the method suggests new conformations for the query protein. We demonstrate the method with tyrosine kinases. Using the Epidermal Growth Factor Receptor (EGFR) kinase domain active conformation as our query, we reproduced the inactive conformation with root mean square deviation (RMSD) under 1.35Å, based on the structural similarity to the active conformation of C-SRC tyrosine-kinase and the known inactive conformation of this protein. The sequence identity between the two kinase domains is only 37% and the fact that they share similar active and inactive conformations might not be obvious. The idea of inferring new conformations of a protein of interest based on known conformations in related proteins is not new. However, to the best of our knowledge ConFind is its first automated implementation.

Session III

8. alleHap: An efficient algorithm to construct zero-recombinant haplotypes from parentoffspring pedigrees

Nathan Medina Rodriguez, Angelo Santana del Pino, Ana M Wagner, Jose M Quinteiro Gonzalez
Universidad de Las Palmas de Gran Canaria, Spain

Haplotype inference is an essential stage in genetic linkage analysis and estimation methods are also very frequently used to reconstruct haplotypes in current genetic association studies. Most of the latter are focused on haplotype phasing from recombinant DNA areas of unrelated individuals and use likelihood-based methods to infer the presence of alleles in several loci with very time-consuming probabilistic algorithms. So far, literature does not analyze zero-recombinant haplotypes composed by large number of alleles because of the genetic distance among the loci and the possible presence of recombination hotspots. But some intergenic regions are very stable against parent-offspring DNA recombination (such as HLA loci). Nevertheless there are hardly any alternative methods for constructing haplotypes from non-recombinant DNA areas, even considering that computational inference by probabilistic models may cause a large number of incorrect inferences. We have developed an algorithm called alleHap, which is able to impute alleles from parent-offspring pedigree databases with missing family members, to later construct their corresponding, unambiguous haplotypes. alleHap comprises the following stages: data loading, pre-processing, computing and evaluation. The data loading stage will correspond to the load of a database, either real or simulated. In the preprocessing step the loaded dataset is filtered to detect genotyping errors or missing values. Subsequently, in the computing phase, allelic missing values are imputed and resulting haplotypes are then constructed. Finally, the evaluation stage reveals an optimum performance of the algorithm even with a large number of families (>1000) considering pedigrees of up to 9 individuals or 7 offspring. Simulations also demonstrated good computational performance with haplotypes consisting of 6 or more alleles. Our algorithm allows the reconstruction of haplotypes from parent-offspring datasets. The proposed algorithm has been tested by simulations and also with the Type 1 Diabetes Genetics Consortium database. The alleHap algorithm is very robust against inconsistencies within the genotypic data and consumes very little time, even when handling large amounts of data. The missing data imputation may improve results in numerous epidemiological and/or genetic linkage studies.

9. ancGWAS: Graph-based Approach for Analyzing Sub-networks Underlying Ethnic Differences in Complex Disease Risk in Admixed Populations

Emile R. Chimusa, Mamana Mbiyavanga, Gaston K. Mazandu, Nicola J. Mulder, Eileen G.Hoal
University of Cape Town, South Africa

Despite numerous successful genome wide association studies (GWAS), detecting variants that have low disease risk still poses a challenge. If the effect of a gene polymorphism is small, and dependent on interaction with a second gene, GWAS may fail to detect a significant signal if the effect of a variant in the other gene is not considered. GWAS may thus generate false negative or inconclusive results. Examining the combined effects of genes increases the likelihood of fully characterizing the susceptible genes. Recent methods incorporate the association signal from GWAS into a human proteinprotein interaction network for testing combined effects of SNPs. However, although in many cases SNPs within a gene and genes within a pathway are correlated, most of these methods do not account for the dependency of p-values, which are assumed to be independent. Violation of independent assumptions in these methods may generate errors and inflate the results. We

present an algebraic graph-based method (ancGWAS) to identify significant sub-networks underlying ethnic differences in complex disease risk in a recently admixed population by integrating the association signal from GWAS data sets, the local ancestry and pair-wise linkage disequilibrium into the human protein-protein interaction network. Through simulation of interactive disease loci in the admixed population, we demonstrated that ancGWAS holds promise for comprehensively examining the interactions between genes underlying the pathogenesis of complex diseases and also underlying ethnic differences in disease risk. We conducted imputation on a GWAS dataset to study tuberculosis susceptibility in the South African Coloured population and applied ancGWAS to the data. Our results replicated previous tuberculosis loci and introduced novel genes and sub-networks underlying ethnic differences in tuberculosis risk. ancGWAS: http://web.cbio.uct.ac.za/ancGWAS/software.html

10. Oqtans: A Multifunctional Workbench for RNA-seq Data Analysis

Vipin T. Sreedharan, Sebastian J. Schultheiss, Geraldiné Jean, André Kahles, Regina Bohnert, Philipp Drewe, Pramod Mudrakarta, Nico Görnitz, Georg Zeller, and Gunnar Rätsch
Memorial Sloan-Kettering Cancer Center, USA

The current revolution in sequencing technologies allows us to obtain a much more detailed picture of transcriptomes via deep RNA Sequencing (RNA-Seq). In considering the full complement of RNA transcripts that comprise the transcriptome, two important analytical questions emerge: what is the abundance of RNA transcripts and which genes or transcripts are differentially expressed. In parallel with developing sequencing technologies, data analysis software is also constantly updated to improve accuracy and sensitivity while minimizing run times. The abundance of software programs, however, can be prohibitive and confusing for researchers evaluating RNA-Seq analysis pipelines. We present an open-source workbench, Oqtans, that can be integrated into the Galaxy framework that enables researchers to set up a computational pipeline for quantitative transcriptome analysis. Its distinguishing features include a modular pipeline architecture, which facilitates comparative assessment of tool and data quality. Within Oqtans, the Galaxy's workow architecture enables direct comparison of several tools. Furthermore, it is straightforward to compare the performance of different programs and parameter settings on the same data and choose the best suited for the task. Oqtans analysis pipelines are easy to set up, modify, and (re-)use without significant computational skill. Oqtans integrates more than twenty sophisticated tools that perform very well compared to the state-of-the-art for transcript identification, quantification and differential expression analysis. The toolsuite contains several tools developed in the R*ä*tsch Laboratory, but the majority of the tools were developed by other groups. In particular, we provide tools for read alignment (bwa, tophat, PALMapper, ...), transcript prediction (cuffinks, trinity, mTIM, ...) and quantitative analyses (DESeq, rDiff, rQuant, ...). In addition, we provide tools for alignment filtering (RNA-geeq toolbox), GFF file processing (GFF toolbox) and tools for predictive sequence analysis (easySVM, ASP, ARTS, ...). See http://oqtans.org/tools for more details on included tools. Oqtans is integrated into the publicly available Galaxy server http://galaxy.raetschlab.org which is maintained by the R*ä*tsch Laboratory. It is also available as source code in a public GitHub repository http://bioweb.me/oqtans/git and as a machine image for the Amazon Web Service cloud environment (instructions available at http://oqtans.org). Oqtans sets a new standard in terms of reproducibility and builds upon Galaxys features to facilitate persistent storage, exchange and documentation of intermediate results and analysis workflows.

Poster Presentations

Bioinformatics of Disease and Treatment

1. Statistical Data Analysis Methods Applied on Wisconsin Diagnostic Breast Cancer Data Set

Gokhan Ersoy
Middle East Technical University, Turkey

In statistics, to understand the data we have, we should analyze it using special methods. One of these methods is visualizing the data in 2 or 3 dimensions. But when data has more than 3 features, we cannot place it on a human-perceptible space. So we need to apply some dimension reduction techniques on high-dimensional data to reduce its features into at most 3. Another method is clustering. When we apply clustering algorithms on data, we can easily see the groups of related samples. The last but not least, we can build a trained model to predict the class of unlabeled samples using labeled samples. In this paper I propose the numerous methods that can be applicable on such a high-dimensional data and their results after the application on the Wisconsin Diagnostic Breast Cancer data set. Almost all methods which is applied on the Wisconsin Diagnostic Breast Cancer data set are successful to make the data understandable for human perception. The classes of data set becomes clearly visible on 2D and 3D projections. And validation methods prove the correctness of clustering results.

2. Mechanistic analysis of prospective natural drugs for checking Alzheimer's plaque pathology

Abhinav Grover
Jawaharlal Nehru University, India

Alzheimer's is a neurodegenerative disorder that results in loss of memory and decline in cognitive abilities. Accumulation of extra cellular β amyloidal plaques is one of the major pathology associated with this disease. β-Secretase or BACE performs the rate limiting step of amyloidal pathway. Inhibition of this enzyme offers a viable prospect to check the growth of these plaques. Numerous efforts have been made in the recent past for generation of BACE inhibitors, however many of them failed during the preclinical and clinical trials owing to drug related/induced toxicity. In the present work, we have used computational methods to screen a large dataset of natural compounds to search for small molecules having BACE inhibitory activity with low toxicity to normal cells and studied their detailed mechanistic behaviors. Preliminary experimental bioactivity assays of these compounds have shown favorable results. The structure of human BACE was prepared using Schrödinger. A data set consisting of 1,69,109 natural compounds from 10 different suppliers of ZINC database was prepared using LigPrep. Clustering of probe sites in BACE based on their spatial proximity and total interaction energies confirmed the binding cavity. Prepared data-set of natural compounds was then virtually screened against BACE using Glide's HTVS docking. The compounds above threshold of 6 HTVS score were selected and subjected to high-precision Glide's XP protocol for refinement. The top scoring compounds above a cutoff of 11 XP docking score were analyzed for structural and thermal stabilities using MD simulations on Desmond with OPLS all-atom force field 2005. The individual probe site related closely to the favored high-affinity binding site of BACE, thus validating the ligand binding pocket of the enzyme. Out of several hundred compounds screened using HTVS, six compounds showed significant binding affinity with high-precision Glide XP docking. The two top scoring natural compounds ATAET and DHED were studied for their detailed interactions with BACE. High ligand-efficiency and glide-Emodel scores obtained for these compounds suggested significant binding affinity for BACE. MD simulations carried out to mimic bodily environment and to study the dynamical behavior of binding revealed steady and low-value RMSD trajectories of ligand-bound enzyme complexes, indicating

stabilization of these ligands in the BACE functional cavity. Our analysis showed that the first proteolytic cleavage step APP - the rate limiting step of BACE, will be inhibited by binding of ATAET while DHED will restrict flexibility of the flap and modulate the functionality of the enzyme leading to non-formation of the intermediate complex, thus preventing Alzheimer's. Preliminary bioactivity assays of these natural compounds have confirmed computational findings, thus providing experimental evidence to the study.

3. An integrative rare diseases research portal

Pedro Lopes, Paulo Gaspar, José Luís Oliveira
University of Aveiro, Portugal

The latest advances regarding modern life sciences hardware and software technologies brought rare diseases research back from the sidelines. Whereas in the past these diseases were seldom considered relevant, in the era of whole genome sequencing the direct connections between rare phenotypes and a reduced set of genes are of vital relevance. The increased interest in rare diseases research is pushing forward investment and effort towards the creation software in the field, and leveraging on the wealth of available life sciences data. Alas, most of these tools target one or more rare diseases, including only the most relevant scientific breakthrough in its specific niche. Hence, there is a clear interest in new strategies to deliver an holistic perspective over the entire rare diseases research domain. This is Diseasecard's reasoning, to build a true knowledge base for all rare diseases. Built using the latest semantic web technologies included in the COEUS framework (http://bioinformatics.ua.pt/coeus/), the Diseasecard portal delivers unified access to a comprehensive rare diseases network for researchers, clinicians, patients and bioinformatics developers. Data are aggregated in a semantic knowledge base, enabling the creation of an advanced web application, and empowering the usage of collected data in external systems. This latter point, an interoperability API, is a clear added value to this field, featuring a SPARQL endpoint capable of answering advanced rare diseases inference and reasoning queries. Connecting over 20 distinct heterogeneous resources, Diseasecard's web workspace provides a direct enpoint to the most relevant scientific knowledge regarding a given disease, through a navigation hyper-tree, LiveView interactions and full-text search, enabling in-context browsing.
Diseasecard is publicly available online at http://bioinformatics.ua.pt/diseasecard/.

4. Homology model of the30S ribosomal subunit from Mycobacteria tuberculosis

Peterson Gitonga Mathenge
University of Nairobi, Kenya

Mycobacterial tuberculosis, the causative agent of the tuberculosis disease, has infected more than a third of the world population to date. It has been known to be a very aggressive strain that is highly resistant to current drugs that target Tuberculosis. Antibiotics such as viomycin and capreomycin have been shown to bind to functionally important regions of the bacterial ribosome and thereby inhibiting protein synthesis process and thereby affecting the bacterial cells viability. Since current methods for obtaining three dimensional structures of the macromolecules are slow and tedious, we demonstrate a faster and inexpensive way of generating structural models in silico by employing both de novo and homology modeling methods In this thesis, we report modeling of the three dimensional structure of the 30S ribosomal subunit from Mycobacteria tuberculosis through the structure prediction methods mentioned above. We report a high resolution ribosomal structure comparable in quality to experimentally determined crystal structures. We hypothesize that, the structure will open a new door in the approach on drug target, and will be important in the development of a new class of anti-bacterial compounds. It will provide a structural scaffold on which structure based drug design studies can be performed. In silico screening of ligands can be carried out to identify compounds that show binding potential on ribosome, ribosomal RNA (16S rRNA) or the ribosomal proteins. Compounds

identified this way can be further studied for antibacterial activity. We hypothesize that the generation of the 30S ribosomal subunit from Mycobacteria Tuberculosis will provide a structural scaffold that will allow In-silico structure based drug design.

5. Predicting bone diseases risk based on the phenotypic comorbidity network approach

Mohammad Ali Moni, Pietro Liò
University of Cambridge, United Kingdom

A comorbidity relationship between two diseases exists if both of them appear simultaneously in a patient. Phenotypic information can be exploited to build a model that predicts disease risk by studying these comorbidity relationships between different diseases. Here we introduce a phenotypic bone diseases network based on the correlations obtained from the patient diseases medical records data and develop a prediction model to determine the risk of individuals to develop future bone diseases. Here we applied network and association analysis on a data set of patient medical records to construct disease networks and build a predictive model to assess disease risk for individuals based on medical history. At first we constructed a phenotypic disease comorbidity network based on the relative risk and phi-correlation and studied their structural properties in order to better understand the relationships between diseases. We built network whose nodes are the diseases and a link between two nodes occurs when a comorbidity relation appears. The edges weights were assigned with the calculated weight values based on the number of patients showing both diseases. Then the model is generated by using the set of frequent bone diseases and the diseases related to the bone metabolism that contemporarily appear in the same individual. The diseases the patient could likely be accepted in the future are obtained by considering the items induced by the high confidence rules generated by recurring disease patterns. The medical record of a patient is then compared with the patterns discovered by the model, and a set of diseases is predicted. Moreover, our result indicates that the progression of the different bone diseases is different with different genders patients. This combination of population-level data and phonotype comorbidity network information could help build novel hypotheses about bone disease mechanisms. Our findings show that this proposed approach is a promising method to predict individual risk disease by taking into account only the diseases a patient had in the past.

6. Structure of the 40s ribosomal subunit from Plasmodium falciparum By Homology and De novo modeling

Harrison ndungu Mwangi
University of Nairobi, Kenya

Generation of the three dimensional structures of macromolecules using in silico structural modeling technologies such as homology and de novo modeling has improved dramatically and increased the speed in which tertiary structures of organisms of interest can be generated. This is especially the case if a homologous crystal structure is already available. High resolution structures can be rapidly created using only their sequence information as input and thus increasing the speed of scientific discoveries. In this study, a host of homology modeling and structure prediction tools such as RNA123 and SWISS – MODEL among others, were used to generate the 40S subunit from Plasmodium falciparum. This structure was modeled using the published crystal structure from Tetrahymena Thermophila, a homologous eukaryote X-ray structure. In the absence of any information from the solved Plasmodium falciparum 40S ribosomal crystal structure, the model accurately depicts a global topology, secondary and tertiary connections, and gives an overall RMSD value of 3.9 Å relative to the templates crystal structure. The model accuracy is even better than prior hypothesis, though deviations are modestly larger for areas that had no homology between the templates. These results lay ground work for using this approach for larger and more complex eukaryotic ribosomes, as well for still larger RNAs, RNA-protein complexes and entire ribosomal subunits. The model created will provide a scaffold onto which in silico ligands screening

can be performed with the ultimate goal of developing new classes of anti-malarial compounds.

7. Screening the Prostate Cancer Susceptibility Loci at 2q37.3 and 17q12-q21 for Novel Candidate Genes in Finnish Prostate Cancer Families

Tommi Tapani Rantapero, Virpi Laitinen, Tiina Wahlfors, Daniel Fisher, Johanna Schleutker
University of Tampere, Finland

According to several studies, genetic risk factors have been shown to be associated to prostate cancer susceptibility. Several chromosomal loci have been shown to be associated to familial prostate cancer. In a recent genome-wide linkage study strong signals coming from 2q37 and 17q21-22 were discovered in Finnish population. To study these loci in detail we performed a targeted high-throughput DNA sequencing on 21 families including 65 cases and 5 controls. In addition, RNA-sequencing was performed for 33 of these cases from purified RNA from whole blood. The aim of this study was to identify variants that associate to prostate cancer susceptibility. Variant calling from sequencing was done using Samtools and variants were subsequently annotated using information from UCSC genome browser database. Three pathogenicity prediction tools Polyphen-2, Pon-P and Mutation taster were used to elucidate the possible phenotypic effects of variants located within genes. As an alternative variant priorization approach we compiled a list of prostate cancer associated genes within the regions of interest from literature, Cancer Gene, Gene-Ontology and pathway databases. To study the intergenic variants in more detail an eQTL-analysis was conducted applying two statistical models: Linear and a non-parametric directional test based model. As a result of the pathogenicity prediction a ranked list of 152 variants with putative effect on protein function was obtained. 38 of these variants as well as 20 additional variants from prostate cancer associated genes were chosen for further validation in a larger population. Validation of these and eQTL targets is currently ongoing.

8. Molecular Modeling study of 4-Phenyl-1H-Imidazole and its derivatives as potential inhibitor of indoleamine 2,3-dioxygenase (IDO)

R. Rathna, R. Aarthi, T. M. Vinodhini, A. Kumar, A. Pon, A. J. Kumar, S.A.H. Naqvi
BioDiscovery - GroupSolutions for Future, India

Science world has little knowledge about immune escape which is a crucial feature of cancer progression. Many human tumors express indoleamine 2, 3- dioxygenase (IDO), an enzyme which mediates an immune-escape in several cancer types. An approach for creating new IDO inhibitors by computer-aided structure-based drug design was created. Molecular docking approach using Lamarckian Genetic Algorithm was carried out to elucidate the extent of specificity of IDO towards different classes of 4-PHENYL-1H-IMIDAZOLE. Combining a novel algorithm for rapid binding site identification and evaluation with easy-to-use property visualization tools, the software has provided an efficient means to find and better exploit the characteristics of ligand binding site. Total molecules of 3000 were virtually screened from different databases on the basis of the structural similarity & Substructure of 4-PHENYL-1H-IMIDAZOLE. The docking result of the study of 3000 molecules demonstrated that the binding energies were in the range of -11.28 kcal/mol to -2.35 kcal/mol, with the minimum binding energy of -11.28 kcal/mol. We report molecule AP-1 which showed 4 H- Bonds with active site residue and lowest binding energy of -11/28 kcal/mol. The molecule AP-1 showed Drug Likeness score of 0.92 with Mol PSA as 42.10 A2 and MolVol as 322.23 A3. The MolLogS was -7.41 (in Log(moles/L)) 0.01 (in mg/L with Drug Likeness Score of 1.26, Drug-Score of 0.34 and Solubility of -8.93. The molecule showed no indication for mutagenicity, & tumorigenicity. Also, no indication for irritating & reproductive effects found. Further in-vitro and in-vivo study is required on this molecule as the binding mode and in-silico ADMET study including LD 50 value provided hints for the future design of new derivatives with higher potency and specificity.

9. Network information improves cancer outcome prediction

Janine Roy, Christof Winter, Zerrin Isik, Michael Schroeder
Technische Universität Dresden, Germany

Disease progression in cancer can vary substantially between patients. Yet patients often receive the same treatment. Recently, there has been much work on predicting disease progression and patient outcome variables from gene expression in order to personalize treatment options. Despite first diagnostic kits on the market, there are open problems such as the choice of random gene signatures or noisy expression data. One approach to deal with these two problems employs protein-protein interaction networks and ranks genes using the random surfer model of Google's PageRank algorithm. In this work we created a benchmark dataset collection comprising 25 cancer outcome prediction datasets from literature and systematically evaluated the use of networks and a PageRank derivative, NetRank, for signature identification. We show that the NetRank algorithm performs significantly better than classical methods such as foldchange or *t*-test. Despite an order of magnitude difference in network size, a regulatory and protein-protein interaction network perform equally well. Experimental evaluation on cancer outcome prediction in all of the 25 underlying datasets suggests that the network-based methodology identifies highly overlapping signatures over all cancer types, in contrast to classical methods that fail to identify highly common gene sets across the same cancer types. Integration of network information into gene expression analysis allows the identification of more reliable and accurate biomarkers and provides a deeper understanding of processes occurring in cancer development and progression.

10. Development of a Moroccan Database for Cancer Care (MD2C)

Oussama Semlali, Adil El Yamine, Fadoua Haoudi, Housna Arrouchi, Ahmed Moussa, Azeddine Ibrahimi

In Morocco women's Breast Cancer constitutes a major public health problem. According to the Central Cancer Registry RCCR, the disease's incidence increased during the period of three years to 39.9 new cases per 100.000 women. Breast cancer is a heterogeneous disease with different morphologies, molecular profiles, clinical behavior and disparate response to therapy. However, the increasing understanding of molecular carcinogenesis has begun to change paradigms in oncology from traditional single-factor strategy to a multi-parameter systematic strategy. The classic therapeutic model for breast cancer treatment has changed from adopting radical surgery, conservative surgery, radiotherapy, chemotherapy and hormonotherapy to more personalized strategy. In this paper, we describe the development of the Moroccan Database for Cancer Care (MD2C). As a first step this platform will integrate all the information relevant to Moroccan breast cancer patients in a database. A query interface is developed using open source technologies, allowing easy secure access to the breast cancer database. The second step is to generate experts systems to assist in decision making. Our MD2C database includes all patient's personal and socio-economical data, family and personal disease history, clinical and paraclinical diagnosis, genetic and genomic data. This work, and during all the development phases, was done by our bioinformatics team in a multidisciplinary setting including oncologists, pathologists and pharmacists. This database will help Moroccan doctors in making precise decisions concerning risks, diagnosis and therapeutic protocols to use and will allow us to extract of knowledge to generate the first Moroccan breast cancer therapeutic model.

11. A Novel Time Series Analysis of Transcriptional Response to Targeted Treatment of Diffuse Large B-Cell Lymphoma (DLBCL)

Arjan G Van der Velde, Heather Selby, Adam Labadorf, Bjoern Chapuy, Margaret Shipp, Stefano Monti
Boston University, USA

Diffuse large B-cell lymphoma (DLBCL) is the most common non-Hodgkin lymphoma in the United States. Forty percent of patients with DLBCL succumb to the disease, and new therapeutic approaches are needed. One such therapy is currently in clinical trials; however, the detailed biological mechanisms governing the response to this treatment in DLBCL are not well understood. Characterization of the transcriptional response to treatment is essential to understand the biological mechanisms of action of a drug. The focus of our project is the analysis of a large gene expression dataset consisting of a panel of DLBCL cell lines profiled at five time points after treatment. We developed a novel time series analysis approach to quantify the dynamic evolution of gene expression, and applied it to our dataset to carefully characterize the response to the pharmacological perturbation. The time series analysis identifies differential expression of genes, and enrichment of biologically relevant gene sets and pathways from publicly available repositories. We created a custom visualization tool to explore the various dimensions of our results at multiple levels of detail that is biologically intuitive. The combination of the time series analysis pipeline and the visualization tool identified both novel and previously known mechanisms of actions of the therapeutic treatment on DLBCL cell lines.

Comparative Genomics

Simulation of Gene Family Histories

Maribel Hernandez-Rosales, Nicolas Wieseke, Marc Hellmuth, Peter F. Stadler
University of Leipzig, Germany

The reconstruction of the evolutionary history of large gene families has remained a hard and complex problem, which amounts to disentangling speciation events from gene duplication events. The evaluation of reconstruction algorithms is hampered, furthermore, by the lack of well-studied cases that could serve as a gold standard. We present here a method for simulation of gene family histories. Starting with the generation of a species tree S with the "Age Model" which reflects the topology of real data species trees, followed by a Poisson Process for the generation of small and/or large scale duplications: gene, cluster or genome duplications with gene order rearrangements. A special rule is then applied to recently duplicated genes to account for the deletion of redundant gene copies before they can be stabilized by sufficient functional divergence or subfunctionalization. This model in particular accounts for the increased loss rates in the wake of multiple gene duplications and in particular for genome duplications. The result of this simulation is a gene tree Gi for each family i together with a true reconciliation map to the species tree S. The known reconciliation provides us with a labeling of the internal nodes of Gi with duplication or speciation events. This in turn determines the true orthology relation for all genes residing in the leaves of S. In addition to that, the gene orders within their respective genomes is obtained as well. We proposed an algorithm that simulates gene family histories akin to real data. This will allow reconstruction algorithms to measure their accuracy and performance. Given a certain reconstruction method one might ask if the orthology matrix could be deduced from the inferred reconciled tree or if the homology relation between the genes was predicted correctly. Furthermore it could be analysed if the method was able to infer the gene duplications and losses. A method that is able to detect large scale duplications will then identify the cluster and genome duplications generated by our algorithm.

12. Determinants of protein evolutionary rates in light of ENCODE functional genomics

Nadezda Kryuchkova, Marc Robinson-Rechavi
University of Lausanne, Switzerland

The influence of different parameters, from gene size to expression levels, on the evolution of proteins has been previously studied mostly in yeast and *Drosophila*. The main feature which has been found to explain protein

evolutionary rate was the level of gene expression, especially in yeast.
Here we investigate these relations further, and extend them to mammals, especially taking in account gene expression and chromatin organization in different organs and different developmental stages. For expression we used a microarray experiment over zebrafish development as well as the RNA-seq data from ENCODE for 22 different tissues of mouse. We used ENCODE data to define which transcript is used as reference to compute gene length, intron number, etc. We use partial correlation to take into account dependencies between gene features. We find strong differences between tissues or developmental stages in impact of expression on evolutionary rate. Over all tissues, an interesting result is that evolutionary rate is better correlated with maximal expression in one tissue then with average expression value over all tissues. Dependencies between gene features need to be taken into account for an unbiased view of gene evolution. Overall results are consistent with those in *Drosophila*. We find important differences between tissues in the relation between expression and evolutionary rate, especially for the central nervous system and testis.

13. A Comparative Bioinformatics Study of the A-domain Binding Pockets of the Nonribosomal Peptide Synthetases of Bacillus atrophaeus UCMB 5137(63Z)

Candice N. Ryan, Svitlana Lapa, Lilija Avdeeva, Rainer Borriss, Oleg Reva, Özlem Tastan Bishop
Rhodes University, South Africa

Due to increased plant resistance to the existing antibiotics produced, there is a need to develop alternatives. Nonribosomal peptides (NRPs) are important plant phytopathogens synthesized by nonribosomal peptide synthetases (NRPSs). In this study, a newly sequenced Bacillus strain Bacillus atrophaeus UCMB5137 (63Z), found to have increased phytopathogenic activity, was investigated to gain insights to the possible reasoning behind this activity. NRPS modules were identified using a novel script that can act on unannotated, raw DNA sequences. The Structure Based Sequence Analysis Webserver was used to identify the amino acids incorporated into the final NRP, which were compared to the NRP database. Five NRPSs were found within the strain; fengycin, mycosubtilin, surfactin, bacillibactin and bacitracin. Some of the modules usually present for these NRPSs were not present in the test strain. A phylogenetic study was carried out and the topologies of the trees showed that genes were not transferred horizontally. Which lead to the hypothesis that different NRPS genes are under different adaptive evolutionary pressures. 3D structures of certain domains of the fengycin synthetase and mycosubtilin synthetase from the test strain were constructed. A docking study was performed using the amino acid predicted to bind within the A-domain binding pocket as the ligand and the active site regions inferred from literature. The NRPS from which the A-domain originates also influences substrate specificity as well as the module in which the A-domain occurs within the NRPS. Future work will include other strains within the Bacillus subtilis group.

14. Comparative co-expression: pairing transcriptomics and metabolomics data from Solanaceous species reveals genes mediating the Biosynthesis of anti-nutritional glycoalkaloids

O. Tzfadia, M. Itkin, U. Heinig, A. Bhide, Y. Chikate, J. Beekwilder, A. P. Giri and A. Aharoni
Weizmann Institute of Science, Israel

Steroidal glycoalkaloids (SGAs) found in *Solanaceae* food plants *viz.* potato, tomato and eggplant are well-known anti-nutritional factors in humans. Nearly 200 years since the first report on the main potato SGA, namely, *α*-solanine, the biosynthetic pathway of these molecules remains largely unknown. We took advantage of the extensive transcriptome data available in the related potato and tomato plants, both producing SGAs, to identify conserved genes that are co-expressed in both species. This approach appeared to be most valuable as we could identify multiple genes that likely participate in SGAs metabolism and its control. Detailed characterization of one of

these candidates, GLYCOLAKLAOID METABOLISM 4 (GAME4), encoding a cytochrome P450 protein, revealed that it performs a key step in the cholesterol-derived SGA pathway. The *in silico* comparative co-expression approach used in this study could be highly effective for gene discovery in the case of other, related plant species, that produce analogous specialized metabolites.

15. Environmental Pressure Imprinted in Genomes through Protein Disorder

Esmeralda Vicedo, Avner Schlessinger, Burkhard Rost
Technical Universiy Munich, Germany

Many organisms have been able to adapt to extreme habitats. Such prokaryotes, referred to as extremophiles, need to change their genetic code with respect to their non-extremophile relatives to survive in the extreme. Here, we reveal that differences between organisms from distinct habitats are imprinted upon a single feature of protein structure, namely the fraction of proteins with long regions that are predicted to be disordered. We ran various prediction methods on 46 entirely sequenced genomes representing organisms from diverse habitats and taxonomy and found that the overall composition of proteins with long disordered regions is linked to extreme conditions. In fact, the overall percentage of proteins with disordered regions was more similar between organisms of similar habitats than between organisms of similar taxonomy. For example, proteins from archaean and bacterial halophiles that survive high salt conditions tend to have substantially more disordered regions than their taxonomic neighbours. More generally, our finding that a microscopic feature as coarse-grained as the overall content in proteins with disordered regions correlates with such a complex macroscopic variable, as the environment remains surprising and will have to be investigated through future case-by-case studies of the underlying molecular mechanisms.

Computational Aspects

16. Computational Analysis Of Anopheles Gambiae Metabolism To Facilitate Insecticidal Target Discovery

Marion Adebiyi, Segun Fatumo, Ezekiel Adebiyi
Covenant University, Nigeria

Insecticide resistance is an inherited characteristic involving changes in one or more insect gene, and a major public health problem hindering the control of malaria. Biochemical research has elucidated a complete image of the metabolic architecture of organisms that included that of *Anopheles gambiae*, analyzing the metabolic network of *A. gambiae insilico* represents a smart way to identify essential reactions that determines the survival of this organism, which systematically gives insight to identification of potential metabolic insecticide target. We developed a graph based model that analyzed the topology of the metabolic network of *A. gambiae* and identified essential enzymes in the network. We extracted, re-annotated and re-constructed all metabolic reactions of *A. gambiae* from BioCyc (AnoCyc version 1.0) database. This yielded a network with 1328 metabolites and 1215 reactions and each reaction was considered to be reversible. Choke point metabolites/ reactions and essential reaction were extracted by knock-out. We tested qualitatively if these products could be generated by reactions that serve as potential deviations of the metabolic flux using breadth first search-method. 198 essential reactions was identified, their corresponding genes, sequences and E.C number was extracted and blasted (BlastP) against the human protein genome to further extract its human homologue, modeled structure, %identity and E-value. To select our insecticidal targets, we compared this list with a comprehensive list of 22 gold standard targets of approved Insecticide, we further analyzed their sequences with ParCrys Algorithm to estimate and predict the crystallization propensity of these proteins and validated with ExtalPred and PPCpred. Finally we presented a refined list of 13 new potential candidate targets for *A. gambiae*, most of which have

reasonable evidence to be valid targets against other arthropods, pests and insect vectors.

17. Stemming Algorithms and Applications in Bioinformatics

Reinaldo Alvares, Rubem Mondaini, Nicolas Carels
Universidade Federal do Rio de Janeiro, Brazil

Stemming algorithms reduce graphical forms of related words to a compact representation - called stem - which represents their base meaning. For example, the words computer, computing and computed, generically point to the meaning associated with the computer. The comput substring is a good candidate to be the stem for this group. Such algorithms can be designed depending on the language, through the use of lists of prefixes and suffixes. These solutions known as rule-based stemmers. Furthermore, statistical methods can be used with the advantage of not depending on any knowledge of the morphological structure of the target language. These solutions known as statistical stemmers. In the last decade, there has been significant accumulation of biological sequence data on the Web. Several studies have been conducted with the aim of finding interesting patterns, which may somehow contribute to scientific development. When it comes to proteins, there are thousands of sequences available for consultation with information on various characteristics, e.g. the motif thereof. We believe there is an intersection between statistical stemmers and bioinformatics. The search for this intersection is the objective of this research. Two linguistic methods were explored: the first one uses the hierarchical clustering. The second one works with the conception of statistical rule to estimate the stem of words. The solutions were applied in the process of stemming and in bioinformatics, in the context of the problem to search for protein motifs in amino acid sequences. In the second case, as motivation, we use the metaphor that the protein sequences can be analyzed as if they were words of that protein language. Thus, as the stem represents the base meaning of the words, the motif captures the base meaning of the protein. Two experiments were conducted: in the first, 5000 Portuguese words were tested in the process of stemming. The best results came from hierarchical clustering. In the second, 6000 protein sequences extracted from the PFAM database were used, in order to estimate the motif of them. The test results show the applicability of the approaches presented, particularly the one that makes use of statistical rule.

18. Computational phenotype prediction of ionizing-radiation-resistant bacteria with a multiple-instance learning model

Sabeur Aridhi, Haïtham Sghaier, Mondher Maddouri, Engelbert Mephu Nguifo
Blaise Pascal University, France

The bioremediation of nuclear wastes with pertinent bacteria and low cost is a challenging problem. The use of ionizing-radiation-resistant bacteria (IRRB) for the treatment of these radioactive wastes is
determined by their surprising capacity of adaptation to a variety of toxic molecules. To date, genomic databases indicate the presence of thousands of genome projects. However, only a few computational works are available for the purpose of phenotypic prediction discovery that rapidly determines useful genomes for the bioremediation of radioactive wastes. A main idea in this context is that resistance to ionizing radiation of IRRB is the result of basal DNA repair pathways and that basal DNA repair proteins in IRRB present a strong ability to effectively repairs damage incurred to DNA. In this work, we study the basal DNA repair protein of IRRB and ionizing-radiation-sensitive bacteria IRSB to solve the problem of phenotypic prediction in IRRB. Thus, we consider that each studied bacterium is represented by a set of DNA repair proteins. Due to this fact, we formalize the problem of phenotypic prediction in IRRB as a multiple instance learning problem (MIL) in which bacteria represent bags and repair proteins of each bacterium represent instances. Information on complete and ongoing IRRB genome sequencing projects was obtained from the GOLD database. We initiated our analyses by retrieving orthologous proteins

implicated in basal DNA repair in eight IRRB and six IRSB with fully sequenced genomes. Protein sequences were downloaded from the FTP site of the curated database SwissProt (http://www.uniprot.org/downloads). We proposed a novel MIL-based approach for predicting IRRB using proteins implicated in basal DNA repair in IRRB. We used a local alignment technique to measure the similarity between protein sequences to predict ionizing-radiation-resistant bacteria. To the best of our knowledge, this is the first work which proposes an *in silico* approach for phenotypic prediction in IRRB. The proposed system is available online at http://home.isima.fr/irrb/. The first results provide a MIL-based prediction system that predicts whether a bacterium belongs to IRRB or to IRSB. We have shown that our MIL-based approach allows good results in comparison with traditional setting of machine learning.

19. Inferring clonal evolution of tumors from SNV frequency data

Wei Jiao, Shankar Vembu, Amit G. Deshwar, Lincoln Stein, Quaid Morris
University of Toronto, Canada

High-throughput sequencing allows the detection and quantification of frequencies of somatic single nucleotide variants (SNV) in heterogeneous tumor cell populations. In some cases, the evolutionary history and population frequency of the subclonal lineages of tumor cells present in the sample can be reconstructed from the SNV frequency measurements, given gentle assumptions about how somatic SNVs arise. But automated methods to do this reconstruction are not available and the conditions under which reconstruction is possible have not been described. We describe the conditions under which the evolutionary history can be uniquely reconstructed from SNV frequencies from single or multiple samples from the tumor population and we introduce a new statistical model to infer the subclonal evolutionary structure of cancer cells from these frequencies. Our model, PhyloSub, uses a Bayesian nonparametric prior over trees that groups SNVs into major subclonal lineages. PhyloSub automatically estimates the number of lineages, their ancestry and the proportion of cells from each lineage. We sample from the joint posterior distribution over trees to identify evolutionary histories and cell population frequencies that have the highest probability of generating the observed SNV frequency data. When multiple phylogenies are consistent with a given set of SNV frequencies, PhyloSub is designed to explicitly represent the uncertainty in the exact phylogeny. Experiments on a simulated dataset and two real datasets comprising tumor samples from acute myeloid leukemia and chronic lymphocytic leukemia patients demonstrate the ability of PhyloSub to accurately reconstruct subclonal phylogenies. PhyloSub can successfully infer not only simple tree structures like chains but also tree structures with branching from single and multiple tumor samples.

20. Diagnostics of Biclustering Patterns in Gene Expression Data by Nonparametric Bootstrap

Tatsiana Khamiakova, Ziv Shkedy
Hasselt University, Belgium

Biclustering is a popular exploratory tool for high throughput data. Although a large number of biclustering methods has been proposed, diagnostics of the bicluster solution remains a relatively unexplored area, relying mostly on the biological interpretation of the biclusters. We propose a set of diagnostic procedures for biclusters based on non-parametric bootstrap and two-way ANOVA with one replicate per cell. The method provides scores for differential co-expression in terms of row (variable) and column (sample) structure, which can be used for the selection of the most differentially co-expressed biclusters. Applying these tools to a set of additive biclustering methods (Delta-biclustering, Plaid models, FLOC) and multiplicative (FABIA, ISA) run on Breast Cancer Data, we could detect that in terms of difference in co-expression patterns, Plaid and FABIA biclusters were the most distinct. Compared to the currently available method of Chia and Karuturi (2010) , our method has straightforward selection of

thresholds on the scores and is able to highlight non-maximal biclusters by the shape of bootstrap distribution. The proposed diagnostic procedure is applicable to the wide range of biclustering methods and highlights the most relevant biclusters in terms of differential co-expression for further interpretation.

21. A hybrid approach to complex transcriptome assembly

Marco Moretto, Mirko Moser, Riccardo Velasco, Duccio Cavalieri, Azeddine Si-Ammour
Fondazione Edmund Mach, Trento

A transcriptome assembly is usually performed de novo or using a reference genome. Each methodology was used successfully to assemble transcripts by aligning reads generated from various next generation sequencing (NGS) platforms. However, using a reference based approach to assemble transcriptomes of newly sequenced polyploid and heterozygous organisms such as fruit trees is tedious. This is mainly due to the presence of ambiguous regions containing a high number of repetitive regions or highly fragmented assembly of these genomes. We will report a hybrid approach used for transcriptome assembly that exploits both methodologies and uses reads generated by 454 and Illumina NGS technologies. A first de novo assembly step is performed using the longest, unique and high quality 454 reads. Assembled transcripts together with shorter 454 and Illumina reads are then aligned to the reference draft genome in order to detect protein coding genes. Assembled transcripts are then iteratively extended using neighbor reads. Finally, to assess the quality of the transcripts we performed an ORF prediction based on protein similarity obtained from public protein databases. Our strategy enabled the rapid identification of proteins with unknown domains and noncoding RNAs with high confidence.

22. BioSeq.jl : A package for bioinformatics in Julia

Diego Javier Zea, Kevin Squire
Universidad Nacional de Quilmes, Argentina

Bioinformatics is a broad field with diverse needs in informatic tools. With the explosion of bioinformatics data available, performance is becoming a key ingredient of bioinformatic workflows. While high level dynamic languages languages like R, Perl and Python are commonly used for daily bioinformatics tasks and workflows, bioinformaticians frequently move to compiled languages like C or C++ when performance is crucial. In these situations, we believe Julia to be an excellent candidate language for bioinformatics tasks. Julia is a high level dynamic language focused on performance, which runs close to the speed of C. Like Python, Perl, and R, it contains many high level features, including perlcompatible regular expressions, integration with bestofbreed mathematical libraries, strong support for running external programs, and a growing collection of packages. In order to use this Julia potential for bioinformatics, we have started to implement basic types and functionality for working with nucleotide and amino acid sequences. These tools are in the package BioSeq.jl and now cover from classical 8 bit ASCII coding scheme nucleotides and amino acid to 2 bit nucleotide sequences and an alternative 8 bit bitlevel coding scheme. Julia support for regex and metaprogramming makes it possible to match with PROSITE patterns or create regular expressions directly using IUPAC ambiguities. In Julia, sequences objects are mutable arrays, but many classic strings methods are defined for them. This allows the exploration of Julia's other capabilities, including parallel computation and memory-mapping. BioSeq.jl design was inspired by BioPython's Bio.Seq module and by Bioconductor's Biostrings functionality, with a great focus on performance and flexibility. We are continuing to add new sequence types and functionality, and we hope this flexibility and performance provides the fundamental basis for a huge amount of Bioinformatic tools written in Julia.

Functional Genomics

23. Improving microarray data analysis by handling missing values with nonlinear clustering- based method

Chia-Chun Chiu, Shih-Yao Chan, Wei-Sheng Wu
National Cheng Kung University, Taiwan

Microarray technology is commonly used in many biological experiments over past years. However, the fact that microarray data are peppered with missing values is often underestimated. Lots of methods of estimating missing values have been proposed, but the correlation structures behind expression data are mistaken as linear in these studies. Thus euclidean distance or Pearson's correlation coefficient are usually to measure similarity between genes, which leads into erroneous interpretation. In this work, we proposed a robust method incorporating cluster analysis and nonlinear relationship to estimate missing values. Maximal information coefficient and mutual information were used as similarity metrics in cluster analysis, and k-means clustering was used to classify genes into various groups. The missing values of a gene were imputed using the genes in the same group, and the accuracy was estimated by comparing the results of cluster analyses between complete data and simulation data. We applied this nonlinear-clustering based method on a variety of expression data and compared it with commonly used methods. The results show that our method outperforms other methods and improves the outcomes of cluster analysis. Adequate handling of missing value is critical to microarray data analysis. Incorporating nonlinear similarity measurement and clustering-based selection of similar genes are helpful to elevate the performance of imputation methods on incomplete data. Besides, comparing with the data which containing missing values or the data imputed by existing methods, the data imputed by our method provides more reliable and robust results in downstream analysis.

24. A Hidden Markov Model for gene mapping based on whole genome sequencing data

Jurgen Claesen
Hasselt University, Belgium

The analysis of polygenic, phenotypic characteristics such as quantitative traits or inheritable diseases remains an important challenge. It requires reliable scoring of many genetic markers covering the entire genome. The advent of high-throughput sequencing technologies provides a new way to evaluate large numbers of single nucleotide polymorphisms (SNPs) as genetic markers. Combining the technologies with pooling of segregants, as performed in bulked segregant analysis (BSA), should, in principle, allow the simultaneous mapping of multiple genetic loci present throughout the genome. The gene mapping process, which we consider in this presentation, consists of three steps: (i) First, a controlled crossing of parents with and without a trait is performed. (ii) Second, selection based on phenotypic screening of the offspring, followed by the mapping of short offspring sequences against the parental reference, is done. (iii) Finally, genetic markers such as SNPs, insertions, and deletions are detected with the help of next generation sequencing (NGS). Markers in close proximity of genomic loci that are associated with the trait have a higher probability to be inherited together. Hence, these markers are very useful for discovering the loci and the genetic mechanism underlying the characteristic of interest. Several statistical approaches for genetic mapping have been proposed over several decades. These methods range from simple single marker tests to computationally very demanding multiple-gene-loci models. Thus far, few methods have been proposed that are specifically developed for high-throughput sequencing data. Most of the existing methods are two-stage approaches. First, a p-value from parametric tests or a linkage probability is determined for each marker. Subsequently, those values are combined in a sliding window approach. The resulting averaged probabilities are used to select the most probable regions containing gene loci. None of these approaches directly map gene loci while considering all relevant marker data, including information from

nearby markers. We propose a hidden Markov model (HMM) to analyze the marker data obtained by the BSA-NGS process. The model includes several states, each associated with a different probability of observing the same/different nucleotide in an offspring as compared to the parent. The transitions between the SNPs implies transitions between the states of the model. After estimating the between-state transition probabilities and state-related probabilities of nucleotide (dis)-similarity, the most probable state for each SNP can be selected. The most probable states can then be used to indicate regions in the genome with a high probability of nucleotide (dis)-similarity, i.e., which may be likely to contain trait-related genes. The application of the model is illustrated on the data from a study of ethanol tolerance in yeast (Swinnen et al., 2012).

25. Identification and Classification of ncRNAs in Trypanosoma cruzi: A Multistep Approach

Priscila Grynberg, Mainá Bitar, Glória Regina Franco
Universidade Federal de Minas Gerais, Brazil

Non-coding RNAs (ncRNAs) prediction studies have become increasingly essential, since several classes of these molecules with different regulatory, catalytic and structural function have been discovered. The estimative of the number of ncRNAs in different genomes is a permanent question nowadays. In the last years, several kinetoplastids genomes have been finalized, and ncRNAs prediction studies in Leishmania major and Trypanosoma brucei had been published. Based on these previously results, we intended to predict and classify into families ncRNAs in Trypanosoma cruzi complete genome. For this purpose, we used EQRNA, an algorithm for comparative analysis of iological sequences that extracts probabilistic inference about the functionality of a conserved region. The methodology can be divided in three main steps. The first one is mainly a BLAST run between the query sequence and a previously chosen reference genome (in this case, T. brucei genome). The subsequent step comprises EQRNA run itself, identifying all ncRNAs candidate sequences. The last step includes a thorough comparison between the candidate sequences and a set of databases containing experimentally determined ncRNA sequences. The entire genomes of T. brucei and T. cruzi were used to generate the initial alignments submitted to eQRNA, and 4195 ncRNA candidate sequences equal to or longer than 30 nucleotides were found. The candidate sequences were used for blastx search (e-value = 10e-05) against T. cruzi annotated proteins. 2816 candidates matched protein-coding sequences and the remaining 1382 candidates were submitted to a pipeline that included search against 25 different ncRNA databases, ab initio RNA tools and structural analysis. 1301 candidates had no evidence to be classified as ncRNAs and 49 candidates are tRNAs or rRNAs. Twenty-nine candidates presented similarity with ncRNAs from several databases. Our next goal is to identify putative regulatory ncRNAs that may be directed to UTR elements by matching the 29 ncRNAs to a catalog of 5' and 3' UTR sequences of T. cruzi transcripts retrieved from dbEST. In silico approaches concerning energy parameters will be employed to test the validity of these findings.

26. Music-listening regulates innate immune response genes in human peripheral whole blood

Chakravarthi Kanduri, Minna Ahvenainen, Li Tian, Liisa Ukkola-Vuoti, Pirre Raijas, Irma Järvelä, Harri Lähdesmäki
University of Helsinki, Finland

The rewarding effect and the physiological benefits of music-listeningon human health are well acknowledged, but the underlying molecular mechanisms and biological pathways triggered by music-listening remain largely unknown. Here, using Illumina Human HT-12 v4, we analyzed the gene expression profiles in the peripheral whole blood of 41 subjects before and after music-listening to understand its effect on bodily functions. Statistical analyses using linear models identified the differential expression of 23 unique genes that have functions crucial for cellular innate immune responses. Functional

annotation analyses demonstrated the beneficial effect of music-listening on tissue homeostasis. Firstly, network analysis showed that musiclistening dampens peripheral immune responses by down-regulating a closely interacting immune-related network that is associated with functions such as cell death and survival, cell-to-cell signaling and interaction, and inflammatory response. Secondly, biological pathway analysis showed the dampening effect of musiclistening on peripheral immune response-related pathways such as natural killer cellsignaling. Thirdly, biological regulator analysis suggested that music-listening may reduce inappropriate immune responses by balancing pro- and anti-inflammatory agents. Another striking evidence of our study provides reasonable explanation for the modulation of immune responses through music-listening. In conjunction, all these findings suggest that music-listening restrains immune over-activation. These findings provide the primary evidence for the effect of musiclistening on human gene expression and immune responses. A balanced immune homeostasis after music-listening substantiates the benefits of music-listening on human well-being.

27. The characteristics of the DNA structural profile of prokaryotic promoter regions

Pieter Meysman, Kristof Engelen, Julio Collado-Vides, Kris Laukens
University of Antwerp, Belgium

The genomic promoter regions, which control the expression levels of their downstream genes, exhibit different characteristics in their local DNA molecular structure than the remainder of the genome. These characteristics of the DNA structure contain valuable insight into transcription processes and are indeed already frequently used in promoter prediction tools. In this study, the structural patterns present in the molecular DNA structure were compared across a variety of prokaryotic organisms based on accurate transcription start site information that is publically available. Promoter regions were found to be on average less stable, more rigid and more curved than the genomic DNA across all studied prokaryotes. Further sets of promoters could be grouped based on similar structural properties, with each set displaying a unique structural profile. These sets could then be related to regulation by specific sigma factors or to certain expression behaviors of the downstream genes. However comparison between different organisms revealed large differences in the found structural profiles, with larger evolutionary distances resulting in greater differences. It could be concluded that there is great variety in the structural DNA properties of promoter regions, which is likely related to the functionality of the promoter.

28. Coordination of gene expression during the development of Drosophila melanogaster evidenced by K-means clustering

Flavio Pazos, Ramón Álvarez, Gustavo Guerberoff, Rafael Cantera
Instituto de Investigaciones Biologica Clemente Estable, Uruguay

Temporal series of genomic expression data can be used to characterize relationships between genes, its regulation and coordination. Several studies have indicated the existence of clusters of genes with coordinated expression during the development of *Drosophila melanogaster*, but so far this has only been correlated with very general aspects of body morphogenesis. We are searching coherent waves of genetic expression with special interest in the development of the nervous system. Our hypothesis is that there is a temporal correspondence between the stages of nervous system development and the expression of clusters of genes during each of these stages. Our final objective is to obtain a catalogue of genes with biological relevance to the synapse assembly. We used RNA-seq poliA data, covering 27 time points along embryonic, larval and pupal stages. We applied hierarchical and non-hierarchical clustering methods, and the more homogenous clusters were those obtained applying the K-means algorithm, clustering the data in 5 groups. Here we show these results and the biological characterization of the obtained clusters. To do so, we performed functional enrichment

analysis of Gene Ontology terms and classified the genes according to their level of expression in five different tissues, using data from FlyAtlas. Three of the five clusters are highly enriched in terms associated to different and consecutive steps of the nervous system development, showing coherent changes in gene expression along time. Using k-means we have found that clustering a temporal series of gene expression in *Drosophila mellanogaster* results in 5 clusters of genes with coherent changes in gene expression along time. Each cluster is enriched in terms associated to several biological processes, showing that clustering genes by its temporal expression profiles results in functionally enriched clusters. This results support the hypothesis that there is a relationship between the temporal expression profile of a gene and its biological function. Our next steps include further clustering of each group of genes and improving the biological characterization of each cluster.

29. Proposed Curation of the database of Clusters of Orthologous Groups of Gene (COG)

Farzana Rahman, Mehedi Hassan, Denis Murphy, Alexandar Bolshoy, Tatiana Tatarinova
University of South Wales, United Kingdom

In recent years, rapid emergence of drug-resistant bacterial strains became a global issue. This is generating speculations like „end of the antibiotic era", „crisis of modern medicine" and the application of chaos theory to medicine. Emergence of Next Generation Sequencing methods opened new possibilities for bacterial analysis and presented new challenges due to the enormous volume of data. Applying NGS to a variety of environmental samples provides information on thousands of bacterial species, which has led to the discovery of new Taxa. We intend to predict bacterial reaction to new antibiotics. We intend to do so by observing their genomic patterns and trend of evolution and adaptation to a hostile environment. Therefore, we are considering curation of bacterial families by improving the Clusters of Orthologous Groups of Proteins (COG). We developed a non-parametric Bayesian method for validation of consistency of the database of Clusters of Orthologous Groups of gene. We found, in at least 35% clusters, the distribution of gene length cannot be approximated using hierarchical Poisson-Gamma distribution but as a mixture of two or more distributions. In the framework of Nonparametric Bayesian approach, model parameter distributions consist of multiple discrete "support" points, up to one per subject in the population. Each "support point" is a set of point estimates of each model parameter value, plus the probability of that set. We set-up a website at http://193.63.130.241/cog to demonstrate the findings. This provides the status of any COG in terms of modality (unimodal/multimodal) and also provides a health check for the COG.

30. Global regulation of alternative splicing by adenosine deaminase acting on RNA (ADAR)

Oz Solomon, Shirley Oren, Michal Safran, Naamit Deshet-Unger, Pinchas Akiva, Jasmine Jacob-Hirsch, Karen Cesarkas, Reut Kabesa, Ninette Amariglio, Ron Unger, Gideon Rechavi, Eran Eyal
Bar-Ilan University, Israel

Alternative mRNA splicing is a major mechanism for gene regulation and transcriptome diversity. Despite the extent of the phenomenon, the regulation and specificity of the splicing machinery are only partially understood. Adenosine-to-inosine (A-to-I) RNA editing of pre-mRNA by ADAR enzymes has been linked to splicing regulation in several cases. Here we used bioinformatics approaches, RNA-seq and exon-specific microarray of ADAR knockdown cells to globally examine how ADAR and its A-to-I RNA editing activity influence alternative mRNA splicing. Although A-to-I RNA editing only rarely targets canonical splicing acceptor, donor, and branch sites, it was found to affect splicing regulatory elements (SREs) within exons. Cassette exons were found to be significantly enriched with A-to-I RNA editing sites compared with constitutive exons. RNA-seq and exon-specific microarray revealed that ADAR knockdown in hepatocarcinoma and myelogenous leukemia cell lines leads to global

changes in gene expression, with hundreds of genes changing their splicing patterns in both cell lines. This global change in splicing pattern cannot be explained by putative editing sites alone. Genes showing significant changes in their splicing pattern are frequently involved in RNA processing and splicing activity. Analysis of recently published RNA-seq data from glioblastoma cell lines showed similar results. Our global analysis reveals that ADAR plays a major role in splicing regulation. Although direct editing of the splicing motifs does occur, we suggest it is not likely to be the primary mechanism for ADAR-mediated regulation of alternative splicing. Rather, this regulation is achieved by modulating trans-acting factors involved in the splicing machinery.

31. Overexpression of Beta-Galactosidase in Chickpea and its role in the treatment of Lactose intolerance

Sunita Yadav
Indian Institute of Technology Kharagpur, India

Chickpea (Cicer arietinum) is the rich source of protein with other crucial nutritional content. In India 90 percent of rural and urban population utilize the Chickpea as a source of their routine diet. From the recent studies it was reported that soaked Chickpea contains fair amount of Beta-Galactosidase (BGAL) and its enzymatic nature is helpful in the digestion of lactose into glucose and galactose. Due to the BGAL activity in Chickpea, we realized its role for the treatment of Lactose intolerance in which children and adult are unable to digest the lactose present in food products specially milk, which is again good source of vitamins and other nutritional content. Children below five years are mainly affected by lactose intolerance. In order to find out the cure of this serious problem we have designed construct of BGAL from Chickpea under CaMV35S constitutive promoter and transformed in Chickpea using Agro bacterium-mediated transformation. We have checked the activity of BGAL in T0 transgenic line using artificial substrate ONPG and found that activity was increased approximately double in comparison to BGAL from control plant. Its activity still to be checked in homozygous T1 transgenic line before the animal test. The result of animal test may guide as whether the over expression of BGAL in Chickpea and its supplement with food products rich in lactose may be proved as best and reliable cure of Lactose intolerance or it would provide us resource to further study the BGAL enzymatic property in the future.

32. Techniques and tools for marker-assisted breeding in potato

Konrad Zych, Ritsert C. Jansen
University of Groningen, Netherlands

The potato is the fourth most important food plant. It is a cheap and nutritious staple food for more than a billion people from 125 countries, and also very important for starch production. The Netherlands are leading in production of seed potatoes, breeding for new varieties and potato research. Unfortunately, potato breeding is still time-consuming and laborious. So-called marker-assisted or precision breeding would potentially solve this. It requires high quality molecular detection platforms to be developed. This is the aim of our project that combines research on a population of commercial varieties (Genome Wide Association Study) and carefully designed crosses (Quantitative Trait Loci mapping). Tetraploidy and tetrasomic inheritance create the grand challenges to genomics and bioinformatics. We provide a comprehensive review of data generation techniques applicable to potatoes: whole-genome sequencing; genotyping using SNP-arrays and DNA-seq; molecular phenotyping using high-density microarrays and RNA-seq; high-throughput classical phenotyping using robots. We also provide an extensive overview of computational analysis tools and software platforms: genotype calling, Genome Wide Association Study, Quantitative Trait Loci mapping (including general tools such as Genome Studio or R/qtl and specific tools such as fitTetra or TetraploidMap); structured data storage enviroments (e.g. Potato Pedigree Database or CropQTL).

Genetic Variation Analysis

33. Comparative homology modeling and protein sequence analysis of p53 tumor suppressor gene

Arora D, Rana P, Dubey Kashyap
GB Pant Engineering College Uttarakhand Technical University, India

P53 is a well-known tumor Suppressor Protein which is present in ~50% of human cancers. With having functions like receiving upstream signals in response to cellular stress, Arrests cell growth if there is repairable DNA damage, Triggers apoptosis if DNA damage is irreparable, second site mutations can rescue cancer mutants. The three-dimensional (3D) structures of the p53 tumor suppression gene retrieve from (Protein data Bank) PDB was constructed based on the homology modeling approach using the Modeller 9.11 program. The molecular dynamics method was utilized to refine the model and it was further assessed by ProSA, three-dimensional structural superposition (3d-SS) and PROCHEK in order to analyze the quality and reliability of the generated model. An appropriate template for p53 was identified based on the e-value and sequence identity. The template and the target sequences were then aligned using ClustalW. The outcomes of the modeled structures were ranked on the basis of an internal scoring function, and those with the least internal scores were identified and utilized for model validation. Discrete Optimized Protein Energy (DOPE) is a statistical potential used to asses homology model in protein structure prediction. MATLAB tool was further used for analysis of closely related these protein structure and to analyses of changes in these structure at amino acids level to map the evolutionary change pattern occurs due to mutations. Predicted model can be useful to develop new inhibitor against p53 tumor suppressor gene. CASTp can be used to study surface features and functional regions of proteins.

34. Improving mutation-induced stability changes prediction in unseen non-homologous proteins with feature-based multiple models

Lukas Folkman, Bela Stantic, Abdul Sattar
Institute for Integrated and Intelligent Systems, Griffith Univeristy, Australia

Even a single amino acid substitution, a mutation, may result in significant changes in protein stability, structure, and therefore in protein function as well. Hence, reliable prediction of stability changes induced by protein mutations is an important aspect of computational protein design. Several machine learning methods capable of predicting stability changes from the protein sequence alone have been introduced. However, their performance on mutations in previously unseen non-homologous proteins is relatively low. Moreover, the performance varies for different types of mutations based on the secondary structure, accessible surface area, or magnitude of the stability change. We explored how designing multiple models, each trained for a different type of mutations, can be beneficial for prediction. We considered three different secondary structure types (α-helix, β-sheet, and coil), two relative accessible surface area assignments (exposed and buried), and three intervals for the magnitude of the stability change (< -1, [-1, 1], and > 1). Thus, this methodology produced three different methods, each with multiple models. By employing careful feature selection, we identified specific features to make each model highly specialised. Support vector machines were used for implementation. Experimental evaluation shows that our approach of combining feature-based multiple models increases prediction performance when compared to currently available methods. In particular, a method based on different secondary structure types employing three models for mutations in helical, sheet, and coil residues yielded the best performance. We found that each of these models included features describing evolutionary conservation and accessible surface area of the mutated residue. However, the remaining features were model-specific. Our method achieved a classification accuracy of 75.42% and

Matthews correlation coeffcient of 0.51. These results represent an absolute improvement of 3.25 percentage points in terms of accuracy and a relative improvement of 16% in terms of Matthews correlation coeffcient when compared to the best currently available method. Our results support a presumption that different interactions govern protein stability in residues with different types of secondary structure. Another interesting aspect of this study is that the concept of feature-based multiple models could be extended to the related areas predicting the impact of a protein mutation.

35. Biological pathways of musical aptitude

Oikkonen J., Onkamo P., Huang Y., Vieland V., Järvelä I.
University of Helsinki, Finland

Humans have developed the perception, production and processing of sounds into the art of music. Even newborn infants can recognize familiar melodies: the neuronal architecture is already set to process music. Musical aptitude has long been recognized to be inherited; indeed, heritabilities have been estimated to be as high as 0.7. Here, we evaluate the genetic background of musical aptitude. We have measured musical aptitude as a skill of auditory perception: abilities to discriminate pitch, duration and sound patterns in tones Genome-wide linkage and association scans were performed for 76 informative families ranging from trios to extended multigenerational families. We used Bayesian approach KELVIN, which supported these large families and quantitative phenotypes. Several genes were identified. Importantly, most of the identified genes are involved in the development of cochlear hair cells or inferior colliculus (IC), both of which belong to the auditory pathway. We also confirmed previous findings of chromosome 4 being linked to musical aptitude. Notably, all of the best associations were located at gene promoters. To study the biological meaning of the sites, we performed promoter analysis for the best-sociated genes. We hypothesize that genes affecting the development of auditory pathway constitute the ground for musical abilities and that differences in gene regulation cause the variation in these skills.

36. FaSD: a efficient model to detect SNPs for NGS data

Weixin Wang, Feng Xu, Panwen Wang, Mulin Jun Li, Pak Chung Sham, Junwen Wang
The University of Hong Kong , Hong Kong

Various methods have been developed for calling single-nucleotide polymorphisms from next-generation sequencing data. However, for satisfactory performance, most of these methods require expensive high-depth sequencing. Here, we propose a fast and accurate single-nucleotide polymorphism detection program that uses a binomial distribution-based algorithm and a mutation probability. We extensively assess this program on normal and cancer next-generation sequencing data from The Cancer Genome Atlas project and pooled data from the 1,000 Genomes Project. We also compare the performance of several state-of-the-art programs for single-nucleotide polymorphism calling and evaluate their pros and cons. We demonstrate that our program is a fast and highly accurate single-nucleotide polymorphism detection method, particularly when the sequence depth is low. The program can finish single-nucleotide polymorphism calling within four hours for 10-fold human genome next-generation sequencing data (30 gigabases) on a standard desktop computer.

Genome Organization and Annotation

37. biomvRhsmm: Genomic segmentation and copy number variation analysis with Hidden-semi Markov model

Yang Du, Eduard Murani, Siriluck Ponsuksili, Klaus Wimmers
Leibniz Institute for Farm Animal Biology, Germany

With high throughput experiments like tiling array and massively parallel sequencing, large scale genomic data are growing at an

unforeseeable velocity. Researchers applying these experiments frequently look at these genome-wide data searching for continuous homogeneous segments or signal peaks, which would represents either chromatin state, methylation ratio, transcript or genome region of deletion or amplification. The objective of these investigations could be generalized as the segmentation problem of partitioning the genome into non-overlapping homogeneous segments. Various models have been proposed to handle either general segmentation problem or particular types of partitioning task. Some of the most addressed areas are copy number analysis with array based comparative genomic hybridization (aCGH) or SNP array, transcript and protein-binding site detection with tiling array. While in recent years, growing efforts have been devoted to the development of computational tools to deal with read count data generated from Next-generation sequencing (NGS). In the R/Bioconductor package biomvRCNS, we implement a novel Hidden-semi Markov model (HSMM), biomvRhsmm, which is specially designed and tailored to serve as a general segmentation tool for multiple genomic profiles, arising from both traditional microarray based experiments and the recent NGS platform. As a generalization of hidden Markov models (HMM), HSMM allows the sojourn distribution (probability distribution of staying in the same state) to be specified other than the Geometric distribution implicitly used in common HMM. In the package, several types of sojourn distribution are implemented. Other than the flat prior commonly used in Bayesian inference, prior information for the sojourn density could be estimated from annotation or previous studies, thus be effectively utilized together with positional information of features to guide the estimation of the most likely state sequence. With its full probabilistic model, various emission densities are provided, enabling the model to handle normally distributed data from traditional array platform as well as count data from sequencing experiment. The proposed model has been tested against well studied aCGH dataset from Coriell cell lines and RNA-seq data generated by ENCODE project to show both its functionality and reliability.

38. Contracted repertoires of novel putative chemosensory gene in Glossina morsitans morsitans
genome

Obiero G.F.O, Nyanjom S.G, Mireji P.O, Christoffels A, Masiga D
International Center of Insect Physiology and Ecology, Kenya

Tsetse flies detect ecological chemical stimuli using chemoreceptor-related proteins like odorant receptors (ORs), gustatory receptors (GRs), ligand-gated ionotropic receptors (IRs), odorant binding proteins (OBPs), chemosensory proteins (CSPs), and sensory neuron membrane proteins (SNMPs). The interaction of these proteins trigger physiological behaviors leading to locating obligate vertebrate host for blood-meal, con-specific mates and larviposition sites. While feeding, the flies transmit various trypanosome species that cause deadly African trypanosomiases, whose remedies are yet unknown. The study aimed at elucidating putative chemoreceptor repertoires in *Glossina morsitans morsitans* and offer an understanding of their possible use in designing novel tools for managing the vectors of infectious diseases. Published protein coding chemosensory genes in fruit fly and mosquitoes downloaded from public databases were used as queries into tsetse pre-annotated genome assembly v0.5 Yale strain. Many bioinformatic tools and databases were used to annotate the tsetse chemoreceptors. We recovered 46 ORs, 14 GRs, 18 IRs, 32 OBPs, 5 CSPs and 2 SNMPs in tsetse genome, revealing an overall reduction compared to other Dipterans. A significant reduction in OBPs, GRs, ORs and IRs in tsetse relative to *D. melanogaster* was identified. Significantly, tsetse genes share strong one-to-one orthology to *D. melanogaster* kins. Post annotation analysis revealed six-fold expansion of pheromone-like ORs and CO2 sensitive GRs with no sweet taste sensors, and highly conserved antennal IRs for common odors. Observed chemoreceptome contraction in

tsetse matches their less complex ecology, obligate vertebrate blood-feeding behavior, and limited host range, thus eliminating the need for expanded chemical sensors. Further more, tsetse combine visual and tactile cues to olfaction in finding their specific hosts, thus relieving over-reliance on olfaction. The *G.m.morsitans* genome encode lower chemoreception genes compared to other diptera, probably explaining their peculiar habitat, feeding and reproduction style. The results are already integrated into vectorbase repositories to increase available resources for the nagana vector, *G. m. morsitans.* In addition to revealing molecular basis of tsetse chemoreception, results lay foundation for future functional and comparative studies with other related species. Once their regulatory features are established, these genes will improve search for olfactory-based management tools against bloodfeeding vectors.

39. Evaluation of the Results of Hidden Markov Model Based Gene Finding Algorithms Applied on Unfinished de novo Genome Sequence

Zeynep Ozkeserli, H. Gokhan Ilk, Hilal Ozdag
Ankara University Biotechnology Institute, Turkey

One of the most important steps in a *de novo* genome project is the gene finding step. In this study, several hidden Markov model based gene finding algorithms are applied to a bacterial *de novo* genome sequence, consisted of 13 scaffolds, and the results are compared. The study is designed to investigate the unique effects of the algorithms on the process. For this purpose both pipelines and various algorithms are used. The pipelines used are NCBI PGAAP and RAST. All algorithms used are from the GeneMark series and part or extention of each other. From the GeneMark Series, GeneMark.hmm, GeneMarkS, and Heuristic Approach for Gene Prediction are used. Comparisons are done using an overlap investigation software with bp resolution designed by the researchers. Comparisons made are: 1- PGAAP vs. GeneMarkS + GeneMark.hmm, 2- GeneMarkS Gene Finder vs. GeneMarkS Combined (GeneMarkS with Heuristic Approach) Gene Finder 3- Heuristic Approach + GeneMark.hmm vs. GeneMarkS + GeneMark.hmm, 4- GeneMarkS Combined vs. GeneMarkS + GeneMark.hmm, 5- PGAAP vs. RAST. After making the comparisons, sequence properties of regions where gene finding is found to be differentially performed are investigated. The results are evaluated to observe the differences on a quantitative and qualitative basis. As a pipeline PGAAP and as a gene finding algorithm GeneMarkS + GeneMark.hmm found to be performed better than others on our data when the results interpreted altogether.

40. COIF: Using assembly metadata to improve prediction and scaffolding of mobile genetic elements in unfinished genomes

Mitchell J Sullivan, Nico K Petty, Scott A Beatson
University of Queensland, Australia

Mobile genetic elements (MGEs) enable the lateral genetic transfer (LGT) of DNA between bacterial cells. LGT contributes to the rapid adaptation of bacteria to differentenvironmentsand plays an important role inbacterial virulenceand antibiotic resistance. MGEs (includingplasmids, phage,transposonsand genomic islands) make up a significant proportion ofintraspecies variation and present a serious informatic challenge due to their repetitive nature, whichleads to fragmentation in draft genome assemblies. We present COIF,a tool that creates and uses a contiguous sequence adjacency graph to improve identification and scaffolding of MGE contigs (http://coif.sourceforge.net). Adjacencies are identified using a De Bruijn graph.MGE contigs are predictedby traditional methods (such as sequence annotation). Contigs are then added or removed from the predicted set based on proximity in the graph to other predicted contigs. Plasmid contigs are usually identified in draft genome data by identifyingplasmid specific annotationsand by identifying contigs with higher than average coverage. Whereas the former methodresultsin a large number of false negative predictions as plasmid specific

genes are only present on a small number of contigs, coverage-based methods arenot always feasible for large plasmidsthat may be low copy numberwithin the genome. We show how using contig adjacency metadata with COIF significantly improves both the sensitivity and specificity of finding plasmid sequencesin draft genome data. We also show how COIF has been used to completely assemble large (>100kilobase) antibiotic resistance plasmids in Illumina paired-end assemblies of Escherichia coli genomes.

Metagenomics

41. Effect of imbalanced training data on testing and generalization performance

Müşerref Duygu Saçar, Jens Allmer
Izmir Institute of Technology, İzmir

MicroRNAs (miRNAs) are single-stranded, non-coding RNAs, which control gene expression at the post-transcriptional level. Although hundreds of miRNAs have been identified in various species, many more still remain unknown. Therefore, discovery of new miRNA genes is an important step for understanding the miRNA mediated post-transcriptional regulation mechanism. Due to limitations of experimental approaches to identify miRNA genes, many sophisticated computational methodologies have been proposed to identify miRNAs. Recently, there have been many attempts to predict miRNAs by employing machine learning. However, the biggest weakness of existing machine learning based miRNA gene identification methods is the imbalance of positive and negative examples used for training. This problem originates from the fact that the exact number of real miRNAs in any genome is unknown and the number of positive examples is significantly smaller than that of negative examples. In this study we used negative data with differing number of examples, ranging from 800 to 50000 while the number of positive examples remained constant at 1600 examples. We employed four different classifiers: Logistic Regression, Naïve Bayes, Random Forest, and SVM. When we compared the best and worst precision and recall values among the results, we observed a difference of up to 50%. Moreover, when we applied the models obtained from these classifications on a different data set to test the generalization, the precision and recall values showed even bigger differences. Consequently, our findings suggest that the training example imbalance is an important factor influencing the quality in machine learning based miRNA prediction methods.

Pathogen Informatics

42. Analysis of amino acid usage pattern for the emerging trends in the pandemic potential of hemagglutinin genes of Influenza A (H1N1) Virus

Rachana Banerjee
University of Calcutta, India

In April 2009, a new swine-origin influenza A (H1N1) virus was discovered in United States and Mexico that spread rapidly across the world by human-to-human transmission. Hemagglutinin (HA) is one of the major glycoproteins of influenza virus that helps in viral entry into the host cells by evading the host immune response. In the present study, a comprehensive comparison of the amino acid composition of all the HA gene sequences of H1N1, specific for human host, available in the Flu Database (NCBI) from 1918 to 2009 has been performed. Phylogenetic tree constructed using all the HA amino acid sequences mentioned above, shows significant genomic diversity of the HA amino acid sequences of May-December, 2009, which corresponded to the pandemic strains, forming a separate clade. Correspondence Analysis on amino acid usage of all the above HA gene sequences generates an entirely separate cluster for the HA genes from May-December 2009, establishing the amino acid usage pattern of 2009 pandemic HA genes to be characteristically different from that of the non-pandemic genes. In order to assess the extent of divergence between different clusters, we have used Mahalanobis distance. The Mahalanobis distance of a cluster comprising of HA gene of the year 1918 from

the non-pandemic cluster is almost half of the distance of the pandemic 2009 strains from the non-pandemic cluster. Conversion of Hotelling's T2 distribution for the two clusters to F distribution followed by calculation of p-values proves that the distance between these clusters is statistically significant ($p<0.001$). Present study indicates the possibility of identifying 2009 pandemic HA genes on the basis of their amino acid composition alone. It also explains that the pandemic characteristics of the HA gene sequences of 2009 are significantly different compared to that of 1918. There is a hint that if the 1918 sequence would have re-emerged in the present decade, the impact of pandemic perhaps would not be so devastating. This is further supported by the fact that the cluster comprising of the 1918 HA genes also contains a HA gene from the year 2005, which caused only local outbreak in Thailand.

43. Insights into the B-cell response in a natural human infection: high-throughput mapping of epitopes using next-generation peptide chips

Santiago J. Carmona, Claus Schafer-Nielsen, Juan Mucci, Morten Nielsen, Fernán Agüero
Universidad Nacional de San Martín, Argentina

The full set of specificities in a human antibody response to a natural infection remains largely unexplored. Here, we used next-generation high-density peptide microarrays to demonstrate for the first time that it is feasible to identify and map hundreds of B-cell epitopes from a complex natural human infection. In this work, we have analyzed the B-cell immune response in humans with Chagas Disease, caused by a protozoan parasite. The chip consists on a tiling array of ~200K 15-mer peptides synthesized using a maskless photolithographic technique, which in concert cover >500 individual proteins. This represents a coverage of ~5% of the proteome, including known antigens, previously uncharacterized proteins selected using a recently published bioinformatic method (Carmona SJ, et al 2012,), randomly selected proteins, and random sequences following parasite's proteome di-peptide composition. Antibody pools from healthy individuals and infected patients were assayed in a single chip, and data processed to obtain disease-specific signal for each peptide. These were used to reconstruct full-length protein antigenicity profiles (Figure 1, left panel). A smoothing procedure showed significant improvement on signal to noise ratio. A testing set of previously known Chagas antigens with fine mapped epitopes was used to assess our performance on linear B-cell epitope identification. Performance on this task was excellent, with an area under the ROC curve of 0.92 (Fig1). Discrimination of antigens from nonantigens is a more challenging task, however. Using a threshold of 20u (1u = background interquartile range) and a setup with an antigen-non-antigen ratio of 1%, we were able to detect 20 out of the set of 45 known antigens (44.4%) with a Positive Predictive Value (PPV) of 91% corresponding to 2 false positive predictions. Applying this threshold to the complete set of proteins analyzed on the chip, we detected 78 novel potential antigens, with an average of 1.5 epitopes per protein. In this work we show that high-density peptide chips allow rapid, highthroughput identification of B-cell epitopes from a natural infection, caused by a complex pathogen. These findings open the door to complete B-cell response maps of complex human infections.

44. Machine Learning Approach for the Identification of Effector Proteins in Pathogenic Bacteria

Tatyana Goldberg, Burkhard Rost, Yana Bromberg
Technical University of Munich, Germany

Many Gram-negative bacteria with pathogenic or parasitic lifestyles exert their function by secreting a set of proteins, termed effectors, directly into the cytoplasm of a host cell. These effectors modulate then the cellular environment such that the optimal growth of pathogens is assured. The primary goal of this study was to identify effector proteins in a genome scale to better understand the molecular mechanisms of bacterial pathogenesis. We have developed a two-step computational approach for the prediction of

effector proteins. In the first step, we build evolutionary profiles to detect those effectors whose sequences are highly conserved. Notably, using this information alone our method reached a high accuracy rate of 96% in identifying over half of the bacterial effectors. In the second step, we use a set of sequence-based features to identify also those effectors whose sequences are poorly conserved. The features include sequence length, composition of 30-N terminal residues, predicted secondary structure, and predicted sub-cellular localization in the domains of Bacteria (to identify secreted proteins) and Eukaryota (to identify proteins mistaken by Eukaryotes as their own and localized in their cell's interior). Overall, our method achieved sustained levels of 86% accuracy and 80% coverage when evaluated on a non-redundant test set. We rigorously benchmarked our method in comparison to the best alternative methods for effector proteins prediction. In our hands, our method outperformed all other predictors. After development, we used our method to identify pathogenic determinants in more than 2,000 publicly available bacterial genomes. We found the majority of known effector proteins in annotated organisms and suggest novel candidates for further experimental validation in newly sequenced ones.

45. Computational identification of miRNAs in Vericella zoster virus (VZV) and it's target determination

Rezwan-ul-Haque, Auditi Purkaystha, Jahed Ahmed
Shahjalal University of Science & Technology, Bangladesh

MiRNAs(miRNA) are a class of short(~22 nt) endogenously expressed noncoding RNA molecules that regulate gene expression through binding to the mRNA molecules. To date, thousands of miRNAs have been identified whereas a large portion includes in different viral family. Here, we identified miRNA of Vericella zoster virus(VZV) and also determine their targets through computational analysis. This study shows only one mature miRNA is present in Vericella zoster virus(VZV) among 16 pre-miRNAs and surprisingly this miRNA confers no target neither on its host, nor on itself, but we found its target on the other viruses of its own family namely, Alcelaphine herpes virus 1, Ateline herpes virus 3, Bovine herpes virus, Felid herpes virus 1, Gallid herpes virus 1, Macacine herpes virus. The predicted miRNA can be used to regulate the viral replication through direct targeting of key viral replication genes or through manipulation of host pathways which is a breakthrough of traditional research.

Population Genetics, Variation and Evolution

46. On the expansion of dangerous gene families in vertebrates

Severine Affeldt, Herve Isambert, Param Priya Singh, Giulia Malaguti
Institut Curie, France

We report that the expansion of "dangerous" gene families, defined as prone to dominant deleterious mutations, can be traced to two rounds of whole genome duplication dating back from the onset of jawed vertebrates some 500MY ago. We argue that this striking expansion of "dangerous" gene families implicated in severe genetic diseases such as cancer is a consequence of their susceptibility to deleterious mutations and the purifying selection in post-whole-genome-duplication species. Our data mining analyses, based on the 20; 506 human protein coding genes, first revealed a strong correlation between the retention of duplicates from whole genome duplication (so-called "ohnologs") and their susceptibility to dominant deleterious mutations in human. It appears that the human genes associated with the occurence of cancer and other genetic diseases (8,095) have retained significantly more duplicates than expected by chance (48% versus 35%; 48% : 3,844=8,095; $P = 1.3 \times 10^{-128}$; X^2). We also investigated an alternative hypothesis frequently invoked to account for the biased retention of ohnologs, namely the "dosage-balance" hypothesis. While this hypothesis posits that the ohnologs are retained because

their interactions with protein partners require to maintain balanced expression levels throughout evolution, we found that most of the ohnologs have been eliminated from permanent complexes in human (7.5% versus 35%; 7.5% : 18/239; $P = 1.2 \times 10^{-18}$; X^2). Our results also show that the gene susceptibility to deleterious mutations is more relevant than dosage-balance for the retention of ohnologs in more transient complexes. To go beyond mere correlations, we performed mediation analyses, following the approach of Pearl, and quantified the direct and indirect effects of many genomic properties, such as essentiality, expression levels or divergence rates, on the retention of ohnologs. Our results demonstrate that the retention of human ohnologs is primarily caused by their susceptibility to deleterious mutations. All in all, this supports a non-adaptive evolutionary mechanism to account for the retention of ohnologs that hinges on the purifying selection against dominant deleterious mutations in post-whole-genome-duplication species. This is because all ohnologs have been initially acquired by speciation without the need to provide evolutionary benefit to be fixed in these populations.

47. An Application for Geographic Population Structure Prediction

Eran Elhaik, Mehedi Hassan, Tatiana Tatarinova
University of South Wales, United Kingdom

The question of ancestry has kept us busy for millenia. For centuries, curious minds dug through the historical books, scriptures and census records to find roots of their ancestors. With molecular biology at disposal, modern scientists have tried to find a link between biological data and ancestry. In Europe, science achieved some success; moern bio-geographical algorithms achieved accuracy of 700 km in finding the point of origin, within Europe. However, these methods were inaccurate elsewhere in the world. We developed an online interface to the GPS algorithm (Elhaik et al 2013)cwhich accurately infers the biogeography of worldwide individuals down to their village of origin. The users can enter their admixtures data obtained from DNA based commercial ancestry tests. Upon entering the anonymous data, the tool analyses and plots the prediction on Google map. The user can also upload comma seperated text files for a group. We present the new tool to convert genomic data to geographic coordinates and determine the country of origin.

48. Molecular Evolution and Phylogenomics of the Anopheles gambiae Complex

Mofolusho O. Falade, Benson Otarigho
University of Ibadan, Nigeria

The six species of mosquitoes comprising the *Anopheles gambiae* complex include malaria vectors that have most stable and deadly vectorial capacity in the world, minor vectors and non vector species. The evolutionary history of this species group was inferred using publically available DNA sequence data. The most morphologically, ecologically and behaviourally similar species, *Anopheles gambiae* and *Anopheles arabiensis* (major vectors) were found to evolve from a common ancestor. All the members of this complex were all AT rich. *A. gambiae* and *A. arabiensis* had the highest AT composition, while *A. merus* had the least among the complex. The evolutionary divergence estimates show that these two major vectors are genetically similar. *A. quadriannulatus* (non vector) and *A. melas* (minor vector) were also found to evolve from the ancestor.

49. Coevolution pattern analysis of centrosomal proteins CEP63 and CEP152 using Bayeisan Algorithm

Mrinal Mishra
University of Turku, Finland

Human genome commonly contain gene complexes that code for conserved centrosomal proteins CEP63 and CEP152.Together, they form ring like structure. This CEP63-CEP152 complexed ring is essential for maintaining normal centrosomal numbers in cells. The disruption in co-localisation of these two proteins is the root cause of number of diseases, particularly microcephaly. The co-

ordinated functionality of CEP63-CEP152 ensures proper neurodevelopment, particularly human cerebral cortex growth. The co-ordinated functioning suggests that they might coevolved during their evolutionary history. The aim of this work is to get deep insight into the coevolution aspect of these two proteins in their evolutionary history. Coevolution pattern across 51 different species has been observed with the help of sequence alignment data obtained from clustal x tool. Aligned data was further analysed using TNT to find the most parsimonious tree for coding sequences of both genes. MEGA was used to find the most optimal model for the dataset. Analysis was run for 1 million generation using the optimal model for coding sequence data for both genes and Summary tree was obtained by applying Bayesian algorithm through MrBayes. Most parsimonious trees and Bayesian analysis tree analysis with human CEP63 and CEP152 perspective shows that Human CEP 63 is closely related to Chimpanzee, Gorilla, Macaque and Gibbon CEP63 while Human CEP152 is distantly related to Chimpanzee, Gorilla, Macaque and Gibbon CEP152.So,Human CEP63 is not coevolved with CEP152 but as Chimpanzee, Gorilla, Macaque and Gibbon CEP63 and CEP152 are closely related as shown in most parsimonious tree as well as in Bayesian analysis summary tree, they might have coevolved during their evolutionary history. Similarly, Mouse and kangaroo rat CEP63 and CEP152 might have coevolved and as both are closely related in both Most parsimonious tree as well as Bayesian analysis tree. The resulting coevolution pattern obtained in the group of species might be the result of the parallel events occurring in evolutionary time in both genes which are related to each other. The further understanding and close tracking of the coevolution path will be very beneficial for the novel understanding of missing link enables the formation of CEP63-CEP152 complex ring formation whose understanding will be very important for gaining insight into the diseases associated

Protein Structure and Function Prediction and Analysis

50. Assessing quality of competing structural alignments: An objective measure based on information content.

James H Collier, Arun Konagurthu, Lloyd Allison, Maria Garcia de la Banda, Arthur Lesk
Monash University, Australia

The field of protein structural comparison lacks consensus on how to assess the quality of structural alignments. The difficulty stems from the opposing objectives of maximising the number of equivalent residues whilst minimising loss of fidelity under optimal structural superposition. This has resulted in a number of ad hoc quality measures that aim to obtain a reasonable trade-off between these objectives. We propose a new method to evaluate alignment quality using the natural and objective measure of information content. We treat structural alignments as hypotheses that explain (or, losslessly compress) the coordinate data of one structure in terms of another structure. The quality of any pairwise alignment can therefore be quantified by the total length of the message required to describe the coordinates of the two structures using the alignment hypothesis { the shorter the message length the better the alignment. This framework yields a natural Null hypothesis test: if the alignment hypothesis results in a longer explanation message than it takes to explain each structure independently, without an alignment, the alignment is rejected. Four main useful properties emerge from our information-theoretic measure: Firstly, the negative logarithm posterior probability of a given alignment varies according to its explanation message length. Secondly, the difference in explanation message lengths of two competing alignments is their log-odds posterior ratio. Thirdly, our measure achieves an objective trade-off between the alignment complexity and the ability of the alignment to concisely explain the coordinate data. Finally, our measure easily handles shifts and hinge-

rotations commonly observed in protein structures. We rigorously evaluated our measure on large alignment benchmarks and compared its performance against several popular measures of alignment quality. The results clearly demonstrate a superior discriminative power of our information-theoretic measure compared to other measures. Importantly, our measure, but not others, varies consistently according to the relationship of structures defined by the SCOP hierarchy. A web server implementing the proposed measure is available at: http://lcb.infotech.monash.edu.au/I-rate.

51. Alternative Conformation Prediction of Vibrio Cholerae Concentrative Nucleoside Transporter

Nicholas Giangreco, Timothy Lezon
University of Rochester, USA

Secondary Transporters couple the uptake of substrate with ion transport. Crystallization of these proteins in the past decade allowed for investigation into the molecular mechanism of transport. These structural analyses revealed internal symmetries, hypothesized as integral components of their transport domains. Alternate conformations of secondary transporters suggest that they function through the alternating access mechanism, allowing substrate access from only one side of the membrane, depending on the transporter's conformation. Experimental evidence conducted alongside these analyses confirms the existence of these alternate conformations in human homologues in vivo. With the recent crystallization of Vibrio cholerae concentrative nucleoside transporter, structural and computational analysis predicting the molecular mechanism of transport can now be conducted using coarse-grained modeling to predict the protein dynamics of transport.

52. Coarse-Grained Simulation: Fast And Accurate Calculation Of Protein Binding Affinity

Qingzhen Hou, Jaap Heringa, K. Anton Feenstra
VU University - Amsterdam, Netherlands

Protein-protein complexes are involved in many biological process. A thorough knowledge of Protein-protein interaction is critical in revealing how two proteins interact with each other and form a complex. Although many efforts have been devoted to the development of methodology for this purpose, both prediction from sequence and protein docking methods all have particular limitations. Furthermore, current scoring functions in protein-protein docking are very much limited in their ability to predict binding affinity. In our former study, we have successfully applied coarse-grained simulation to calculate binding free energy (May et al.). In the current work, we try to use this method to estimate the binding affinity of protein complexes from the Protein-Protein Docking Benchmark 3.0. We calculate bind free energy of 43 proteins from the Protein-Protein Docking Benchmark 3.0, and finish all simulations for 43 proteins in 7 days. The value of the binding affinity (pKd) was correlated to the score calculated for each complex. The correlation (r-value) is 0.234 based on all 43 proteins, which is better than most of the docking methods. This open up an opportunity for us to calculate binding affinity.

53. Mechanism-based molecular docking guided by prosthetic groups

Francois Martz, Sylvie Cortial, Jamal Ouazzani, Bogdan Iorga
Centre National de le Recherche Scientifique, Institut de Chimie des Substances Naturelles, France

Presently, traditional docking methods are based on the energetic evaluation of interactions between ligand and protein surfaces in the active site. However, some proteins contain in their active sites small, non-protein chemical entities called prosthetic groups that are essential for the enzymatic reaction. In these cases, the interaction between the ligand and the prosthetic group are difficult to evaluate and generally this leads to an incorrect positioning of the ligand in the binding site. To circumvent this problem, we are currently developing a new method of molecular docking (ProDock) based on enzymatic mechanism and guided by the

presence of a prosthetic group in the active site. This method involves the addition of a new term in the classical scoring function, which will be calculated by quantum methods from potential energy surfaces corresponding to each complex prosthetic group / ligand. This new term accounts for the interaction energy between the ligand and the prosthetic group, whereas all other interactions (protein / ligand and protein / prosthetic group) are evaluated using the existing terms of the scoring function. Etot = ECoulomb + EVDW + EHbond + Etorsion (+ Equant). We have also developed an automatic procedure (PredFace) for the prediction of the stereochemistry of enzymatic reactions catalyzed by prosthetic groups. The current version of this procedure is compatible with proteins containing prosthetic groups from the flavin family, with the isoalloxazine core as the common scaffold. This procedure was validated by the analysis of all PDB structures containing these prosthetic groups and will be soon is available via a web interface. Our ProDock method should be more accurate than the traditional docking approaches, and the PredFace procedure should facilitate the prediction of enzymatic reactions stereochemistry based on structural information. Currently, these methods are used in our laboratory for the study of enzymatic reactions involving nitroreductase NfrA1 and cytochromes P450.

54. CoDNaS: a database of Conformational Diversity of Native State in proteins

Alexander Monzon, Ezequiel Juritz and Gustavo Parisi
National University of Quilmes, Argentina

The native state of a protein is represented by an ensemble of conformers in equilibrium. The presence of different factors, such as a ligand, posttranslational modifications or a change in pH, could shift this equilibrium towards a given conformer. Following conformational selection theory the change in the dynamic landscape and re-distribution of conformer population is a key feature to understand protein function. It has been shown that different structures for the same protein obtained under different conditions represent snapshots of protein dynamism and then characterize putative conformers. CoDNaS database (from Conformational Diversity of the Native State [http://www.codnas.com.ar]) is a redundant collection of PDB files for the same protein, obtained from different experimental protocols to obtain the structure of a protein. CoDNaS is extensively linked with physicochemical and biological information (such as presence of ligand, change in pH and temperature, presence of mutations, change in oligomeric state, loopy and disorder extension, presence of posttranslational modifications) allowing the user to explore how these information could modulate conformational diversity in proteins. Currently CoDNaS have 9398 proteins with an average of 6.14 conformers per protein which involves the analysis of 54364 structures. Using an all vs. all structural alignment between the corresponding conformers of each protein we defined the extension of conformational diversity as the maximum RMSD registered, but other measurements of local and global structural changes are also available. We think that CoDNaS database is a novel tool to relate conformational diversity of native state with different parameters and properties allowing us to increase our knowledge in such important feature of proteins.

55. Tracing the similarities between enzyme folds through binding-site similarity networks

Richa Mudgal, Narayanaswamy Srinivasan, Nagasuma Chandra
Indian Institute of Science, India

Exponential growth in protein sequence and structural databases significantly increases the need for computational approaches for comparing proteins and for obtaining accurate functional annotations.
Comparison of molecules is carried out at different levels, that of whole sequences, domains, sequence motifs, structural folds and sub-structures. The problem is complex even for enzymes, despite being the best characterized type of proteins, due to one fold – many functions and many folds – one function association types. Understanding sequence-fold-site relationships and their

evolutionary implications is thus a challenging task. In this study, we address this problem using a network approach that links enzyme functions through similarities in their binding sites, folds and ligands. We first systematically annotate the domain(s) in a protein structure involved in reaction catalysis using information from bound cognate ligands, cofactors, or known catalytic site residues. A bipartite network of functions and folds as nodes shows that 1683 enzyme functions are associated with 395 structural folds of which 191 functions are associated with more than two folds. A second network of functional associations using similarity in the binding-sites and domain superfamilies identifies 285 functions and 1209 shared interactions. This network reveals about 20 clusters of functions that could be rationalized by similarities either in their ligands or in their catalytic mechanisms. A consolidated network is then constructed and used for elucidating probable paths in evolution from one function to another. This study is directly useful for obtaining accurate functional annotations of proteins and also in poly-pharmacology and protein engineering.

56. Comparative Analysis of MMPs IN Anopheles gambiae

Jacqueline(Koko) Mutai, Mark Wamalwa, Ramadhan Mwakubambanya, Paul Mireji
International Livestock Researh Institute, Kenya

Human malaria is the most important disease in tropical countries in terms of morbidity and mortality. Malaria transmission involves complex interactions between *Plasmodium falciparum* and Anopheles gambiae. For successful invasion/infection the parasite must overcome the immune responses of the vector. Matrix metalloproteinase (MMPs) are a family of zinc metalloendopeptidases, which are known to disrupt sub-endothelial membranes, destroy tight junctions and shed active cytokines, chemokines and other MMPs through cleavage from their precursors. The latter function putatively explains the great parasite loss during invasion of the *Anopheles gambiae* midgut. The objective of this thesis was to study MMPs in *Anopheles gambiae* as a potential for drug target or transmission blocking vaccine (TBV). *Drosophila melanogaster* was used a model and BLAST was used to find MMP in *Anopheles gambiae*. The domains of these proteases were determined through the InterProScan server found at the European Bioinformatics Institute (EBI) website. The 3-D structure was determined using MODELLER and the template file used was 1SU3.pdb, the structure was validated using MetaMQAPII, ProSA and SAVeS. These proteases were then classified into superfamily and family through the identification of conserved domains in a multiple sequence alignment (MSA). The MSA was generated using ClustalX. The results showed that 2 proteins similar to *Drosophila melanogaster's* MMP were found in *Anopheles gambiae*. These proteins were found in the National Centre for Biotechnology Information (NCBI) repository and have the following accession numbers gi | 157020567 | gb | EDO64756.1 |, gi | 33346953 | gb | EAL39361.4 | . These proteases were shown to have the following characteristics: 3 main domains; the prodomain which contains which conserved cysteine residue in the consensus sequence PRCVxPD, a metalloproteinase domain (catalytic domain) and has the conserved consensus sequence HEbxHxbGbxHz where b is a bulky amino acid, x a variable amino acid and z a family specific amino acid. The three hsitidine coordinate a Zn2+ ion which is important for activation of the proteinase and a hemopexin domain which is a 4-bladed propeller with 4 β-sheets in each propeller. Lastly, it falls under the metzincins superfamily due to a 'Met-turn' found a few residues downstream of the catalytic site and matrixins family due to a conserved serine found after the third histidine in the consensus sequence. This study therefore shows the importance of using bioinformatics to study proteins whose structures have not yet been elucidated by X-ray crystallography and Nuclear Magnetic Resonance Spectroscopy (NMR).

57. Detecting repetitions and periodicities in proteins by tiling the structural space

R. Gonzalo Parra, Rocío Espada, Ignacio E. Sánchez, Manfred J. Sippl, Diego U. Ferreiro
Buenos Aires University, Argentina

The notion of energy landscapes provides conceptual tools for understanding the complexities of protein folding and function. Energy Landscape Theory indicates that it is much easier to find sequences that satisfy the "Principle of Minimal Frustration"' when the folded structure is symmetric. Similarly, repeats and structural mosaics may be fundamentally related to landscapes with multiple embedded funnels. The mere existence of repetitions does not guarantee that the system will be symmetric as these should arrange in particular ways and coalesce into higher order patterns. Detecting repeated units and patterns is a first step towards an understanding of their assembly in complete structures and the emergence of symmetry. We present analytical tools to detect and compare structural repetitions in protein molecules. By an exhaustive analysis of the distribution of structural repeats using a robust metric we define those portions of a protein molecule that best describe the overall structure as tessellation of basic units. Patterns produced by such tessellations provide intuitive representations of the repeating regions and their association towards higher order
arrangements. We developed concepts and methods to structurally compare repetitions and patterns in protein structures. We find that some protein architectures can be described as nearly periodic, while in others clear separations between repetitions exist. Since the method is independent of amino acid sequence information we can identify structural units that can be encoded with different primary elements. The methods can be applied to various topological families and resolve fine geometrical differences. Moreover, we define a metric that allows for a crude comparison of the symmetrical dispositions of repetitions between proteins of different size, topology and quaternary arrangement on the same grounds.

58. Attacking Mycobacterium Tuberculosis in the dormant phase: A Combination of expression data with structural druggability and nitrosative stress sensitivity

Leandro G. Radusky, Lucas A. Defelipe, Marcelo A. Marti , Adrian G. Turjanski
University of Buenos Aires, Argentina

It is estimated that onethird of the world population is infected with Mycobacterium tuberculosis (Mt), resulted in 1.8 million deaths worldwide. (World Health Organization, 2011) The host immune response to tuberculosis (TB) infection relies in phagocytosis of the bacilli by the macrophages resulting in the formation of a granuloma which stops bacterial replication. Inside the granuloma the bacteria faces a particular stressing condition characterized by hypoxia, inducible Nitric Oxide (NO) synthase derived NO and nutrient deprivation, and in response switches to a non replicative state, usually called the dormancy phase, where it can remain hidden and alive for decades. Reactivation of latent Mt is a high risk factor for disease development particularly in immunocompromised individuals. Common treatment of TB involves a long treatment with the front line drugs, isoniazid, rifampicin, pyrazinamide and ethambutol. However, the emergence of multi and extensivelydrugresistants (MDR and XDR) Mt strains, and the negative drugdrug interactions with certain HIV (or other disease) treatments, show the urgent need for new antiTB drugs. In the present work we have performed a proteome scale analysis of Mt potential drug targets specific for the dormant phase. For this sake, for all Mt protein domains with available structure, we have first the determined their i) sensitivity to RNOS based upon aminoacidic composition of the active site, ii) pocket druggability using fpocket and different pocket properties. This information was then combined with essentiality, offtarget and microarray derived data in a target prioritization pipeline. Using all the information cited above we performed a weighted search using Sensitivity of RNOS, Druggability, Essenciality, Offtargeting against Human targets and Upregulation in RNOS conditions as criteria for selection. Three new

putative targets have been chosen to follow a virtual screening protocol.

Proteomics

59. Towards an intrinsic protein disorder ontology: generating a controlled vocabulary for the MobiDB database

Tomás Di Domenico, Silvio C. E. Tosatto
University of Padua, Italy

Intrinsically disordered proteins (IDPs) have become an important research topic in structural proteomics. The main sources of IDPs annotations are indirect methods (e.g. PDB structures) and predictions. Several approaches have been made to classify disordered proteins from various points of view. At the same time, a number of databases and information sources on IDPs have been created. We argue, however, that the lack of a uniform way to represent the available knowledge concerning IDPs, apart from not allowing the existing resources from easily interacting, causes the field to become increasingly fragmented. Controlled vocabularies provide the means to organize knowledge in a consistent manner, thus making it more manageable and more easily accessible. This makes it easier for projects using the vocabulary to interact between them. Here we propose a controlled vocabulary for intrinsic protein disorder, based on our analysis and classification of data from the MobiDB database. We hope it will clear the path for the future creation of an intrinsic protein disorder ontology.

60. Insights of HydA1 in enhancing of Hydrogen Production

M.V.K.Karthik
Birla Institute of technology, India

Identification of conserved binding site residues, molecular dynamics simulation of HydA1:PetF complex in Chlamydomonas reinhardtii towards understanding photobiological hydrogen production is reported in the present study. This study indicates that conserved amino acids of HydA1 viz. Arg344, Arg353, where found in the binding site of HydA1-PetF complex. The Z-score of the modeled HydA1 were significant and they were found to be 1.219 which is nearly equal to 1.0 confirms the accuracy of the modeled structure. The docking results show greater extent of binding interaction in HydA1:PetF complex with E-total value of -463.35 Kj/mol. Furthermore the studies validated the stability of this complex through molecular dynamics simulation where Arg23 of PetF showed the hydrogen bonding interaction with Glu151, Glu154 and His156 for HydA1. Finally, the validation of thermodynamics studies revealed temperature value of 311.245 °K and total energy of 4387.020 Kcals/mol which are sufficient to prove a meticulous stability of our model.

61. Utility of prior information such as RNASeq and GPMdb protein observation frequency for improving MS/MS based protein identifications

Avinash Kumar Shanmugam, Alexey Nesvizhskii, Anastasia Yocum
University of Michigan, USA

Tandem mass spectrometry (MS/MS) based shotgun proteomics has become the method of choice for protein identification in most studies. The method employs spectral matching algorithms and statistical models to identify the proteins present in the sample based on the MS/MS spectra generated. However these methods do not, in general, take into account any prior information available about the sample in their protein inference step. Since for most biological systems there is often a wealth of prior information available, algorithms that can incorporate and utilize such information could help to improve the sensitivity of protein identification from shotgun proteomics data. In this study, we have explored the utility of RNASeq abundance values and GPMdb protein observation frequencies for improving MS/MS based protein identification. We have developed a statistical method for adjusting the identification probabilities of proteins, initially computed by the Trans-Proteomic Pipeline analysis suite, to account for the

GPMdb and RNASeq information. But utilizing these adjusted probabilities we were able to confidently identify proteins that would have otherwise fallen below the confidence threshold. In moderate and low depth datasets, (< 2000-3000 proteins) the method allowed improvements of 2-10% in number of protein identifications at 1% FDR. The method described can allow us to obtain more usable information out of available MS/MS data. The method described is general enough that it can be adapted to integrate other kinds of information into proteomic analysis pipelines too.

62. DNMSO, an Ontology to Representation of De Novo Sequencing Results

Savaş Takan, Jens Allmer
İzmir Institute of Technology, İzmir

Proteomics is the study of proteins that can be derived from a genome. For the identification and sequencing of proteins, mass spectrometry has become the tool of choice. In seconds, a tandem mass spectrometer is capable of ionizing a mixture of peptides measure their respective parent mass to charge ratios, selectively fragment peptides and measure the fragment ions. The peptide sequencing problem is then to derive the sequence of the peptides given this measurement. Since there is currently no accepted standard to represent de novo sequencing results, other, not suitable formats are used to store de novo predictions. PEAKS, commercial de novo sequencing software, for instance, stores results in pepXML format. Other de novo sequencing algorithms are using custom formats which may not provide adequate information for further downstream analyses and knowledge mining. We recently proposed DNML, a representation for the representation of de novo sequencing results based on XML. XML is inherently tree-based but de novo sequencing results call for a graph representation due to many to many relationships. Here, we present DNMSO which uses ontology to represent the de novo sequencing results. In addition to overcoming the aforementioned problem, the application programming interface (API) we provide offers reading, validation, writing, creating, and conversion facilities and is downward compatible with our DNML format. We believe that providing a comprehensive API which further supports reading of existing standards for spectra representation makes our representation very attractive since it removes all file handling tasks from the developers of de novo sequencing algorithms.

Sequence Analysis

63. Why Charlie Chaplin came third in a Charlie Chaplin look-alike contest? The importance of objective function in evaluating your alignments

Punto Bawono, Arjan v.d.Velde, Sanne Abeln, Jaap Heringa
VU University - Amsterdam, Netherlands

Charlie Chaplin once entered a look-alike contest in disguise, and won the third prize. This story illustrates the importance of the objective function (OF) in optimization as well as benchmarking procedures. In Bioinformatics, the accuracy of a MSA tool is typically assessed by benchmarking it using reference alignments and the quality of a particular alignment is commonly determined by calculating its sum-of-pair alignment score. The similarity between a query and reference MSA is commonly measured using sum-of-pairs (SP) and column (CS) scores. However, in general these scoring schemes are less meaningful for larger alignments containing long indels. We investigated how well the commonly used alignment score represents the alignment quality since it has been shown that the true biological alignment is not always the highest scoring one. So, how good is the alignment score as an OF? We developed SPdist, a new scoring scheme for MSA tool benchmark and alignment score calculation. SPdist takes the extent of amino acid shifts in query MSAs into account, resulting in a benchmarking scoring scheme which is more able to evaluate alignment quality, particularly for alignments which contain long stretches of indels. The effect of structural diversity on the sequence alignment is studied by comparing the alignment scores to the GDT_TS (main

structural similarity measure used in CASP) over a large number of protein structural alignments. In SPdist the alignment score is calculated in a more biologically-aware fashion in which structurally/functionally important regions are up-weighted while ambiguous regions are down-weighted. We show that SPdist benchmark scoring scheme leads to a better discrimination of the accuracy of MSA tools benchmarked. For example, methods that are being criticized in the user community, despite their relatively high SP/CS scores, indeed yield lower SPdist scores. From our benchmarking studies one can observe that many reference alignments (especially larger ones) attain lower alignment scores than the query alignments. Thus the commonly used alignment score is not so ideal as an OF, and a more biologically-aware alternative such as the SPdist score is more preferable. Several test cases are also presented to show the benefit of SPdist.

64. A new method for spectral characterization of protein families from sequence information using Fourier Transform

Magali Berland, Bernard Offmann, Philippe Charton, Frederic Cadet, Magali Remaud-Simeon, Isabelle Andre
Université de la Réunion

The Resonant recognition model (RRM) is a physico-mathematical model that tentatively links the long-range electron charge transfer along the protein backbone to the biological properties of proteins. The model provides a framework to identify, from sequence information, original features extracted through their numerical encoding and subsequent treatment by Fourier transform. These have been linked to functional properties of protein families through the calculation of family-based consensus spectra. We provide a new method for calculating and characterizing the consensus spectra in the RRM method. Our methodology proceeds by calculating the geometrical mean of the energy spectra of sequences from a functionally characterized family instead of calculating a cross spectrum by simply multiplying the spectra. The amplitude of the recurrent peaks are, in our method, independent of the number of sequences in the initial dataset. Henceforth, to evaluate the significance of a peak, we propose that the signal-to-noise (s/n) cut-off value used to determine characteristic peaks of the consensus spectrum is based on the standard deviation of the s/n values. To determine the statistical significance of each of these peaks, we set up a bootstrap strategy consisting in shuffling the amino acid order for each sequence of the family. We illustrate this new canvas on the analysis of a highly divergent odorant binding protein family and of a more conserved sialidase enzyme family. The proposed canvas provides a generalized and rational framework for a statistically robust spectral characterization of protein families using Fourier transform.

65. Visualseq: A Script for Visual Comparison of Sequence Conservation

IR Costa, F Prosdocimi
Universidade Federal do Rio de Janeiro, Brazil

With the exponential growth of biological sequences databases, the need to analyze sequences automatically and in a user-friendly way becomes more evident. Visual inspection of a large number of sequences comparisons can help achieving a higher level of understanding and insight and enables our well-developed pattern recognition capabilities to reach its full potential. We created a python script to compare amino acids sequence conservation between up to 3 groups of orthologs, called visualseq. This program uses multi-fasta aligned protein files as input, and loops over the alignments averaging the scores for each amino acid position. The scores are generated using a substitution matrix, such as BLOSUM (BLOcks SUbstitution Matrix) or the identity matrix, which defines values for every amino acid comparison. The data generated for all the amino acids positions is then smoothed with a low-pass filter and plotted using the matplotlib python library. We used this program in three clade comparisons involving distant homologs from fungi, plants and metazoans. This result enabled the successful identification of domains conserved only in plants and fungi in essential amino acid

biosynthesizing enzymes. Our work brings practical solutions to some complex problems as multiple alignment comparisons and conserved domain identification in distant clades. While there is still room for improvement such as adding more customization and other smoothing filters, by allowing the user to see his data, this simple program can improve the quality of his analysis and inspire new discoveries. Visualseq is an open sorce program and can be downloaded at: https://github.com/igorrcosta/visualseq . Supported by: FAPERJ.

66. Facet: a feature-based accuracy-estimation tool for protein multiple sequence alignments

Dan DeBlasio, John Kececioglu
University of Arizona, USA

Selecting an aligner - and parameter values for the aligners scoring function - to obtain a quality alignment of a specificc set of sequences can be challenging. Different aligners and dfferent parameter values can produce vastly different alignments of the same sequences. In principle, a user could simply try various aligners and parameter settings, and choose the one that yields the most accurate alignment { except that in practice, the accuracy of an alignment cannot be measured (since the correct alignment is not known). We overcome this obstacle by combining efficiently-computable, real-valued features of an alignment into an estimator of its accuracy that is suitable for choosing both aligners and parameter settings. Facet (Feature-based Accuracy Estimator) is an easy-to-use, open-source utility for estimating the accuracy of a protein multiple sequence alignment, available at http://facet.cs.arizona.edu. Facet can be readily applied to both parameter advising (choosing good parameter values) and aligner advising (choosing a good aligner). The tool provides optimized default coefficients for its linear estimator that are best on average (coefficients may also be specified manually), and can be run as a stand-alone tool, or included in any pre-existing Java application. The Facet website also provides pre-computed parameter sets (substitution matrices and affine gap penalties) that are optimal for boosting aligner accuracy via parameter advising. We show experimental results on applying Facet to parameter advising and aligner advising that improves alignment accuracy by as much as 27% on the most challenging sets of sequences.

67. Nucleosomal TATA-switch: competing orientations of TATA on the nucleosome

Jan Hapala, Edward N. Trifonov
Masaryk University, Czech Republic

Transcription is known to be affected by the rotational setting of the transcription response elements within nucleosomes. We studied the rotational positioning of the TATA box, the most universal promoter motif. We applied a computational nucleosome mapping technique with single nucleotide resolution to eukaryotic promoters. Our results show that the nucleosome DNA sequence harboring the TATA box encodes alternative rotational positions for the same piece of DNA. This may serve for switching the gene activity on and off.

68. Discovery of alternatively regulated transcriptional start sites by cap analysis of gene expression technology

Hiroko Ohmiya, Morana Vitezic, Martin Fith, Masayoshi Itoh, Piero Carninci, Alistair R R Forrest, Yoshihide Hayashizaki, Timo Lassmann
RIKEN, Japan

The production of specific mRNAs by RNA polymerase II is regulated in most phases of homeostasis, growth, differentiation, and development in eukaryotes. Therefore, measuring the transcription initiation events comprehensively will enable us to characterize the transcriptome, and ultimately lead to the discovery of causative genes for human diseases and their regulators. For the analysis of transcription start sites (TSSs) of genes, tag-based methods have been developed, including cap analysis of gene expression (CAGE). This technology is based on sequencing a read starting with the most 5' end of mRNAs, and a challenge for the data analysis is selecting regions harboring

biologically relevant events at multiple scales, including alternative TSSs and promoters, TSS usage, and the genomic distribution of the usage. We developed an approach that firstly constructs the set of reproducible regions among related samples and in a second step detects significant changes between the different samples. we applied our approach to the widely used HeLa and THP-1 cells, and revealed complex patterns of alternative usage at promoters. In particular we detected novel differentially expressed alternative start sites for large number of genes. We conclude that our method is an effective tool to automatically discover alternative promoters and reveal that genes have complicated structures of transcription initiation events. Therefore, it is suggested that the method prompt the resolution of the transcriptome.

69. MISTIC: a Mutual Information Server to Infer Coevolution

Franco Lucio Simonetti, Elin Teppa, Ariel Chernomoretz, Morten Nielsen, Cristina Marino Buslje
Fundacion Instituto Leloir, Argentina

Changes in protein sequences do not occur randomly, there are functional and structural constraints that shape the way in which protein sequences diverge and evolve. These are reflected as conservation patterns in multiple sequence alignments. Besides residue conservation, another information rich approach is to examine the correlated mutational pattern between columns of an alignment. Mutual information (MI) from information theory can be used to analyze this type of covariation. We have previously developed an MI algorithm that corrects several known biases in the MI signal, such as phylogeny, sequence redundancy and low counts. Here we present MISTIC, a publicly available Mutual Information Server To Infer Coevolution that brings covariation analysis closer to non-computational biologists with a friendly-user interface and an interactive view of the results. MISTIC uses a multiple sequence alignment as the basis for calculating mutual information, while structural data can be added by choosing a PDB structure and a reference sequence. Results can be separated in two visually rich formats. First, essential information obtained from the alignment is displayed in a circular layout called MI circos. This provides an integrated view of the multiple sequence alignment in terms of i) the mutual information between residue-pairs, ii) sequence conservation, and iii) the residue cumulative and proximity scores. Also, an interactive view allows analysis of the coevolution network in simultaneous with the structure. Each network node is a position of the multiple sequence alignment and lines (edges) are mutual information values. Selection of nodes and edges shows associated data and automatically maps it to its corresponding position in the PDB structure using a Jmol applet. Several network-oriented tools are also available. Finally, results can be exported in different formats, including images, tables and network files. MISTIC offers an interactive platform to analyse MI and distance networks, as well as different options for graphical representation of the different information signals. We believe the server will provide a powerful tool for non-bioinformatics end-users to analyse the information contained within protein families and guide the search for residues essential for protein function. MISTIC is available at http://mistic.leloir.org.ar.

70. High-throughput sequencing reveals influence of miRNA isoforms on the outcome of qPCR validation

Tomasz Stokowy, Michał Świerniak, Bartosz Wojtaś, Kornel Labun, Knut Krohn, Michał Jarząb, Barbara Jarząb, Ralf Paschke, Markus Eszlinger, Krzysztof Fujarewicz
Silesian University of Technology, Poland

Small RNA expression profiling platforms differ significantly in the sensitivity of detection for small RNA isoforms. This fact implies frequent lack of results reproducibility obtained from high-throughput sequencing (HTS) with quantitative polymerase chain reaction (qPCR). The aim of the study was to design and implement software that indicates and evaluates miRNA isoforms potentially interfering with the main miRNA isoform

validation. 20 follicular thyroid tumor samples were analyzed with Illumina hiScan sequencing. 3 statistically differentiating malignant and benign tumors miRNAs were selected from reads per million and DESeq normalized data. Identified malignancy markers were tested with isoform specific (custom) and standard Qiagen qPCR. The validation was carried in 20 samples library (method confirmation) and 84 independent samples. The implemented R/Bioconductor software – miRNA FMG allowed to identify interfering isoforms that perturbed standard qPCR but did not influence custom validation. Furthermore, implemented software included miRNA seed handling, t-test p value, false discovery rate, mean and median based fold changes of isoform specific analysis. qPCR technique is limited in validation of small RNA expression. Application of proper analysis and FMG software strategy leads to proper design of validation experiments, dedicated for follicular thyroid tumors differentiation and other miRNA HTS studies. Funding: FNP MPD Program "Molecular Genomics, Transcriptomics and Bioinformatics in Cancer" (TS, BW).

Systems Biology and Networks

71. Identification of genetic determinants of colony Morphology switch in natural saccharomyces cerevisiae strain Using a systems biology approach

Cappelletti V, Berná l, Ramazzotti M, Stefanini I, Lee W, Romualdi C, Cestaro A, Kapushesky M, Cavalieri D

Fondazione Edmund Mach, Research and Innovation Centre, Italy

Colony morphology is a fascinating phenotype described in unicellular organisms as a possible step towards multicellularity. The spreading of filamentous structures is used by some pathogenic fungi, as Candida albicans, to invade human tissues. Albeit the genetic determinants of the colony morphology response are poorly understood, several connected pathways, in particular the Ras2/cAMP and the MAPK pathways, have been shown to regulate colony development. Cavalieri and colleagues described a specific type of yeast colony morphology, called filigreed morphology, as present in heterozygosis in M28 *Saccahromyces cerevisiae* strain. The mendelian inheritance observed in M28 meiotic segregants makes this natural strain a good model to elucidate genetic regulation of filamentous growth and deepen into the ecological role of multicellular structures. We analyzed cellular and colony morphology of M28 meiotic segregants in several different carbon sources. The addition of ethanol as the only carbon source lead to an increase in filamentation: in this perspective the stable and uniform morphotype, induced by ethanol, could reflect an adaptation to stress. Transcriptional analysis by mean of Microarrays on cells grown in fermentable and not-fermentable carbon sources and Functional Enrichment Analysis identified the genes involved in the regulation of colony morphology switch and allowed to dissect the filamentation process. Whole genome comparative analysis on 12 M28 sporal derivatives of three different M28 tetrads, whit Next-Generation Sequencing approach, allowed to investigate the mendelian inheritance of filamentous morphotype. The SNPs calling procedure has been used to identify variants able to explain the morphotype differences between M28 meiotic segregants. Our results support the hypothesis of an ecological function of filamentous phenotype in creating a community adaptable to changes of the environmental conditions. We demonstrate that a number of three tetrads is sufficient to map a genetic trait with mendelian inheritance from NGS data. Moreover, by comparing M28 to the S288c fully annotated genome we have also found some putative new genes in M28 natural strain. Finally, RNA-seq based transcriptome analysis on the all M28 sequenced genome allowed to identify a gene expression profile associated to the filamentous morphotype and to confirm the candidate morphotype regulators genes.

72. Graph based Identification of Structural Repeats in Proteins

Broto Chakrabarty, Nita Parekh
International Institute of Information Technology-Hyderabad, India

Repetition of super secondary structure is a common phenomenon, especially in higher eukaryotic organism. The copy number and assembly of these repeating units are responsible for diverse functions mediated through protein-protein interactions, and consequently, defects in repeat proteins have been linked to a number of human diseases. The variation within the repeat units makes their identification difficult at the sequence level and structure based approaches are desired. However, most of these employ structure-structure alignment, which is computationally intensive. Here we propose a computationally efficient structure-based approach for the identification of structural repeats in proteins using concepts from graph theory. The three-dimensional topology of protein structures is known to be well captured by protein contact graphs. The connectivity information in a graph is represented in the adjacency matrix and the eigenspectra of the adjacency matrix depicts the topological importance of each node to the connectivity of the graph. In our earlier work on comparative analysis of graph centrality measures for the identification of ankyrin repeats, we observed that the principal eigenspectra of the adjacency matrix well captures the tandemly repeated structural motifs. Here we propose an algorithm for the identification of any tandemly repeated structural motif using graph properties and secondary structure information from STRIDE database. The algorithm begins by first identifying the length of the repeat motif by analyzing the periodicity of peaks in the eigenvector centrality and repeat boundaries are identified by superposing the contiguous repeats and extending on either side of the peak regions to the start/end of the secondary structure elements and checking for periodicity of the secondary structure architecture in the identified repeat regions. Thus, using the secondary structure annotation helps in refining the boundaries of the repeat regions and to discard false positives. W have tested the algorithm for identifying various structural repeats such as HEAT, WD, Ankyrin (ANK), Tetratricopeptide repeat (TPR), Leucine rich repeat (LRR), etc. which different super secondary structure motif, ranging from all alpha, to all beta to a mixed topology such alpha-turn-alpha, beta-alpha, etc. The predictions are in agreement with annotation in UniProt database. The graph based analysis of protein structures, along with domain information such as the organization of the secondary structure elements provides a computationally efficient approach for the identification of structural repeats.

73. Boolean modeling of the septation initiation network in fission yeast

Anastasia Chasapi, Ioannis Xenarios, Paulina Wachovicz, Viesturs Simanis, Anne Niknejad, Daniel Schmitter, Daniel Sage
University of Lausanne , Switzerland

The goal of our project is to create an inferred regulatory network of the septation initiation (SIN) in *S.pambe,* which will be used to represent current knowledge and aid in predictive experiments. The model contains all interactions specific to the SIN described in the literature to data, and extends a previously published model by Csikasz-Nagy. The produced model represents the largest SIN model to date. The model was simulated using qualitative Boolean modeling, whic resulted in a number of steady states. To evaluate the steady states, a scoring method was developed, using experimentally determined phenotypes of knock-out (KO) and over-expressed (OE) genes. The scoring was used to test thousands of mutant combinations *in silico*. An optimization procedure was followed to increase the score for our SIN model. At each step, the best scoring models were undergoing a regulatory refinement. This included introduction or suppression of regulatory rules; models were then scored by performing KO and OE experiments. Models with increased score were retained as candidates for further evaluation. Using this scoring method we were able to evaluate 110

different models. after selecting the best scored model candidate, we performed a full screening for all possible double perturbations (overexpression and expression loss) of the model genes, a total of 3362 *in silico* experiments. This steady states produced were scored for their similarity to the multiseptated and no-septated phenotype of *S. pombe*. A counter-intuitive prediction was generated using this approach, according to which in a sid4KO-byr4KO we should observe multiseptated phenotype. This prediction was experimentally tested but could not be validated. The observed experimental phenotype, however, was used to modify our model scoring and help us assess the area of improvements for our model. Our current work is focused on capturing the SIN coordinate regulation with cell cycle events and implementing quantitative information. This example shows, even with non validated predictions, the virtuous cycle of modeling and assessment which is an essential component of modeling and simulation. Rarely the failed predictions are described as we tend to always consider positive cases of prediction and validation.

74. Network-based gene prioritization using DTProbLog

Lore Cloots, Sergio Pulido-Tamayo, Dries De Maeyer, Joris Renkens, Aminael Sánchez-Rodríguez, Luc De Raedt and Kathleen Marchal
Ghent University, Belgium

The recent revolution on sequencing technologies makes it possible to obtain the complete DNA sequence of a genome, including humans. Having data on polymorphisms at a genome-wide scale makes the challenge of identifying the genes that lead to quantitative phenotypic traits (i.e., quantitative trait loci or QTLs) amenable. However, a genetic study for such purpose (e.g., GWAS, eQTL) typically results in one or more regions on the DNA that associate to the phenotype of interest. Due to linkage disequilibrium, the associated regions often contain several candidate genes, complicating a direct biological interpretation. Several computational approaches have been developed to identify the most promising genes from these associated regions. This task is referred to as gene prioritization. We developed a network-based gene prioritization approach that also extracts the subnetworks involved in transducing the signal from the gene predicted to be causal to the target genes (genes that are differentially expressed under the phenotype of interest). In our setting the problem of gene prioritization and subnetwork extraction was solved in a decision theoretic version of ProbLog, a novel probabilistic programming language based on logic programming and Prolog. We benchmarked our method on a yeast single gene knockout expression compendium and also applied it on a real dataset resulting from a pooled segregrant-pooled expression analysis. As result we identified novel causative genes responsible for ethanol production capacity in yeast. Next-generation sequencing and interaction networks allow us to pin-point better the responsible genes that influence a Quantitative Trait Loci. Network analysis is a great tool for gene prioritization; future networks will contain more and better information and the results obtained by methods like this one will only improve.

75. A comparative analysis of novel complex disease pathways

Nick Dand, Benjamin Lehne, Nikolaos Barkas, Russel Sutherland, Christopher Tebbe, Frauke Sprengel, Volker Ahlers and Thomas Schlitt
Kings College London, United Kingdom

In recent years a number of tools have been developed to study disease-specific genetic data in the context of interaction networks, with the aim of identifying novel molecular pathways underlying common complex diseases. Typically such tools are validated using data from a well-studied disease so that results can be compared to known pathway genes. Beyond this, although the aim of using the network to propose novel disease mechanisms is clear, the interpretation of results is difficult. We have developed a framework which allows a comparative analysis between disease-related subnetworks. This can be used to compare subnetworks

derived from different data sources for the same disease, or subnetworks corresponding to different diseases. Region Growing Analysis (RGA) is a tool we have developed, whose simple input requirements provide the flexibility to identify regions (proposed disease pathways) based on DNA sequence data, GWAS results, expression data or other experimental output. Using RGA we compare putative disease pathways for different diseases based on Wellcome Trust Case Control Consortium GWAS results. We test the hypothesis that diseases of the same physiological system will show greater pathway similarity than unrelated diseases. We suggest that studying the pathway components that differentiate one disease from another will lead to important inferences regarding the contribution of constituent genes to the disease process.

76. Modeling the Wnt/b-catenin Signalling

Annika Jacobsen
Vrije Universiteit - Amsterdam, Netherlands

The Wnt/β-catenin signaling pathway is important for cell development and stem cell maintenance. Dysregulation of the signaling pathways can lead to tumor formation and colon cancer. Because of the low number of drugs available for effective treatment new insights into the signaling mechanism of the pathway can lead to new targets for treatment. In the mature cell there is a shortage of Wnt leading to β-catenin degradation in the cytoplasma and a limited concentration of nuclear -catenin. When Wnt is present it activates the signaling pathway resulting in nuclear β-catenin accumulation, which in turn activates Tcf/Lef transcription of various Wnt target genes. In colon cancer this dysregulation is mainly caused by mutations in the adenomatous polyposis coli (APC) or β-catenin. A non-deterministic concentration dependent Petri Net model has been developed for the Wnt/ β-catenin signaling pathway. Simulations of the model were able to recapitulate the increased levels of Tcf/Lef seen in experiments with Wnt signaling, APC mutation and β-catenin, respectively. In addition the model also recapitulated a number of experiments with overexpression, knockdown and knockout of different proteins involved in the signaling pathway. The next step will be to make a refined version of the current model and an extended model and use these models to perform predictions of various experiments on the Tcf/Lef levels to be further experimentally validated.

77. Computing Multi-Level Clustered Alignments of Gene-Expression Time Series

Deborah Muganda-Rippchen, Mark Craven
University of Wisconsin, USA

Identifying similarities and differences in expression patterns across multiple time series can provide a better understanding of the relationships among various normal biological and experimentally induced conditions such as chemical treatments or the effects induced by a gene knockout/suppression. We consider the task of identifying sets of genes that have a high degree of similarity both in their (i) expression profiles within each condition, and (ii) changes in expression responses across conditions. Previously, we developed an approach for aligning time series that computes clustered alignments. In this approach, an alignment represents the correspondences between two gene expression time series. Portions of one of the time series may be compressed or stretched to maximize the similarities between the two series. A clustered alignment groups genes such that the genes within a cluster share a common alignment, but each cluster is aligned independently of the others. Unlike standard gene-expression clustering, which groups genes according to the similarity of their expression profiles, the clustered-alignment approach clusters together genes that have similar changes in expression responses across treatments. We have now extended the clustered alignment approach to produce multi-level clusterings that identify subsets of genes that have a high degree of similarity both in their (i) expression profiles within each treatment, and (ii) changes in expression responses across treatments. We evaluate this method by considering the stability of resulting

clusters and their agreement with extrinsic data sources.

78. Cross-species alignment of gene-gene coexpression networks

U.K. Nandal, M. El-Kebir, M. van der Wees, J. Heringa, A.H.C. van Kampen, G.W. Klau, P.D. Moerland
Academic Medical Center - Amsterdam, Netherlands

Animal models have been useful for improving our knowledge of molecular interactions underlying human diseases. However, often animal models fail to mimic human disease adequately. One way of validating the similarity of a model organism to its human counterpart is to integrate gene expression profiles from different studies and then identify conserved co-expression subnetworks across species via network alignment. However, gene-gene coexpression networks are densely connected and require the alignment of millions of weighted edges making the alignment problem computationally demanding. We implemented a network alignment algorithm Natalie, to find the optimal alignment of coexpression networks. However, NP-hardness of the optimization problem complicates the search for the best scoring alignment. Natalie uses Lagrangian relaxation approach in order to obtain lower and upper bounds to the solution. Natalie calculates the best scoring alignment of two coexpression networks by optimizing an objective function that incorporates (i) similarity between nodes (genes) based on all-against-all BLAST and (ii) conservation of the degree of coexpression between pairs of genes. We assess Natalie's ability to align two coexpression networks constructed using two large human and mouse liver gene expression datasets and by calculating empirical P-values for the alignment using a permutation test. The conserved subnetworks derived from the network alignment of the liver datasets showed strong concordance in terms of biological processes involved. The results derived from Natalie show that the method scales well and produces meaningful conserved co-expressed clusters.

79. COLOMBOS: an ever expanding collection of bacterial expression compendia

Paolo Sonego, Pieter Meysman, Qiang Fu, Daniella Ledezma, Marco Moretto, Kris Laukens, Julio Collado-Vides, Kristof Engelen
Fondazione Edmund Mach, Italy

COLOMBOS is a publically available access portal to comprehensive organism-specific cross-platform expression compendia for bacterial organisms. It provides a suite of tools for exploring, analyzing, and visualizing the data within these compendia. The expression compendia themselves are built based on a propriety methodology that is unique in directly combining the data from different technological platforms. COLOMBOS also incorporate extensive annotations for both genes and experimental conditions; these heterogeneous data are functionally integrated in the analysis tools to interactively browse and query the compendia not only for specific genes or experiments, but also metabolic pathways, transcriptional regulation mechanisms, experimental conditions, biological processes, etc. Several improvements have been made. Content wise, we have invested in the development of a compendia creation and management system that has enabled us to greatly expand existing compendia (*Escherichia coli, Bacillus subtilis,* and *Salmonella* Typhimurium) as well as add compendia for other species. Additionally, the current version supports the inclusion of RNAseq
data. Functionally, we have revamped the interface with new interactive visualization and analysis tools, a bicluster tree algortihm for discovering complex coexpression patterns around a set of query
genes, and inclusion of noise models for measurement errors, enabling analysis of differential expression with measures of statistical significance. This work is relevant to a large community of microbiologists by facilitating the use of publicly available genome-wide expression data to support their research, as well as providing a useful resource for top-down systems biology applications.

Other

80. COMPARTMENTS: Using text-mining approach to unravel protein localizations

Janos Binder, Reinhard Schneider, Lars Juhl Jensen
EMBL, Germany

Function of proteins are heavily depend on their subcellular localization, and it is known that many diseases are caused by inefficient, or wrong localization of the proteins. Various biological experiments depend on where in the cell a protein is expressed, but there is no single resource that collects evidence on subcellular localization and provide an overview. We have developed an allinone
web resource for subcellular localization information by combining curated knowledge, literaturemining and software predictions. We collected all annotated knowledge from SwissProt, SGD, FlyBase, WormBase and MGI and we used prediction softwares like YLoc and PSORT to process 1,684,376 unique protein sequences. As a new source of information, we used an inhouse textmining
tool to find pairs of comentioned proteins and subcellular localizations in 23 million PubMed abstracts. Upon a query we show an overview figure, where proteins are categorized into 12 organelles of the cell and different shades of green correlates with the strength of underlying evidence. On the page detailed tables of different channel are also provided with linkouts for the above mentioned databases, and showing the relevant PubMed abstracts in order to support literaturemining results. COMPARTMENTS is available at: http://compartments.jensenlab.org

81. Statistical analysis of molecular pathways

Hu Chen, Joshua Yuan
Texas A&M University, USA

Quantitative pathway analysis is essential for understanding the molecular mechanisms for biological processes and identifying key biomarkers for clinical applications. However, quantitative analysis of molecular pathways is complicated by the multidimensional feature of the problem, where defined statistical modeling and interpretation are challenging. Unlike the single gene analysis, the pathway analysis outcomes are often biased by various modeling assumptions. In order to address these challenges, we hereby proposed, evaluated, and implemented an alternative approach to use multivariate analysis for molecular pathway quantification, comparison, and analysis. The approach clearly defines the variables, p value, and parameter estimations that can be interpreted to classify differentially regulated pathways and identify key genes involved in the classification. The approach is implemented by Python and R into a software package named as Statistical Analysis of Molecular Pathways (SAMP) for the three-step analysis. First, gene expression data from RNA-seq was processed and mapped to specific molecular pathways. Our statistical model here specified the features of genes, i.e. gene expression, as independent variable, and the pathway outcome or classification as dependent variables. Second, MANOVA test was carried out to screen the molecular pathways for differential regulation. Third, principal component analysis (PCA) was carried out as the exemplary multivariate analysis to weigh each genes in variation contributing pathways. Using the new approach, we analyzed several breast cancer test datasets. The novel approach overcomes the limitations of assumptions in previous approaches and provided an effective approach for quantitative pathway analysis with both scientific and clinical application potentials.

82. Reactome FI Cytoscape Plugin

Eric T Dawson, Guanming Wu
Ontario Institute for Cancer Research, Canada

Reactome is a highly-reliable, manually curated pathway-based protein functional interaction network covering nearly fifty percent of human proteins. A plugin for the open-source network visualization tool Cytoscape had been previously developed which allowed researchers to find network patterns related to cancer and other diseases using the functional interaction network. The latest Reactome FI

Cytoscape app brings compatibility with Cytoscape 3.x while allowing for future expansion of the already extensive feature set. Users can create FI subnetworks based on a set of genes, query the FI database for underlying evidence of the interaction, build and analyze network modules of interacting genes, annotate modules using functional enrichment analysis, expand the network by finding genes related to the experimental data set, display pathway diagrams, and overlay the network with a variety of information sources such as cancer gene index annotations. Such a tool has implications in uncovering the genetic roots of various diseases and providing insight into how network analysis may lead to future improvements in personalized medical treatment based upon an individual's functional interaction profile.

83. Stochastic algorithms for motif discovery: a comparison of sampling strategies

Alastair M. Kilpatrick, Stuart Aitken
The University of Edinburgh, United Kingdom

The EM algorithm is the basis of a number of algorithms for motif discovery. A stochastic version of EM has been shown to alleviate known limitations of EM in theory and has been implemented in a motif discovery context as the SEAM algorithm. SEAM removes the complex computation and maximisation of the likelihood function required by EM, replacing it by simply sampling each input sequence using a weighted roulette wheel method. However this sampling method requires considerable computation and may be inefficient. We propose Markov Chain Monte Carlo as a potential solution. We implement a version of SEAM using a Metropolis independence sampler and compare this modified algorithm with the original roulette wheel method to evaluate the potential performance benefits and computational cost. The effects of varying the number of Monte Carlo samples and EM iterations are also investigated. Both methods are tested on a collection of datasets containing previously characterised E. coli TFBS motifs extracted from the RegulonDB database. The Metropolis independence sampler is shown to give good recovery of motifs, based on site-level sen- sitivity and positive predictive value. Using large numbers of samples is shown to often return stronger motif models, based on motif energy. While the Metropolis independence sampler is a relatively simple sampling strategy, its performance in this study indicates the potential in exploring alternative more efficient sampling strategies.

84. PALMapper: Fast, Accurate and Variation-Aware RNA-Seq Alignments

David Kuo
Memorial Sloan-Kettering Cancer Center, USA

High Throughput Sequencing technologies have revolutionized genome and transcriptome sequencing, providing ever larger amounts of read data at exponentially dropping costs. RNA-sequencing (RNA-seq) has become a standard technique not only for the assessment of gene and isoform expression profiles, but also for the investigation of alternative splicing diversity and the identification of novel genes and isoforms. Various tools have been developed to align read data to a single reference genome. More recently, the dramatic increase in sequencing power has led to a large number of studies generating whole exome or whole genome data to account for variations between individuals. Usually, in a first step the variation relative to the reference genome is determined and an individual genome is imputed that can be subsequently used for alignment. As the alignment is still not a trivial process, this multi step approach comes at a very high computational cost.
To address this problem, we describe a variant-aware extension of PALMapper (Jean et al., 2010) that builds upon the highly efficient read mapper GenomeMapper (Schneeberger et al., 2009) with the spliced read aligner QPalma (De Bona et al., 2008). PALMmapper is a fast, accurate, and easy-to-use tool that is designed to accurately compute both unspliced and spliced alignments against a single reference genome. The new version of PALMapper aligns against arbitrary combinations of variants using an efficient dynamic programming

approach. Variant information can be provided in established variant call formats (vcf, sdi, ...). PALMapper supports both single nucleotide variants as well as insertions and deletions. Variants of several sources can be combined and tracked throughout the alignment processes, allowing for a concurrent alignment against several accessions at once and simplifying the analysis of allele specific expression. Through an efficient "seed and extend" algorithm and a banded semi-global alignment, PALMapper is able to align up to 10 million reads per hour per thread to a human genome. In order to resolve ambiguously mapped reads in the PALMapper output, we also provide an efficient strategy to map such reads to unique locations based on local coverage information.

The variant-aware spliced aligner PALMapper combines the benefits of an accurate and highly sensitive alignment algorithm with the versatility of a mapper applicable to multiple highly variable reference sequences at a time. While the speed of PALMapper is comparable to most other recent alignment tools, its variant capability makes it the method of choice for projects requiring concurrent alignment against multiple reference genomes, e.g., genome-wide association studies in population genetics or the identification of causal variants within a disease setting. We illustrate applications of PALMapper in several contexts ranging from plants to worms to cancer.

85. Structural coverage using X-ray crystallography for a current snapshot of the protein universe

Marcin J Mizianty, Xiao Fan, Jing Yan, Eric Chalmers, Christopher Woloschuk, Andrzej Joachimiak, Lukasz Kurgan
University of Alberta, Canada

Structural genomics aims at solving 3D protein structures and shifts focus from individual proteins to protein family-directed structure determination where a large number of selected targets is processed by standardized pipelines. The target selection is aided by in-silico methods that predict crystallization propensity from a given protein chain. Here we designed a method for Fast DEtermination of Targets' Eligibility for CrysTalization (fDETECT) which provides high-throughput and accurate means to select easier to crystallize targets. Utilizing our time-efficient design we analyzed crystallization propensity for 9,586,243 proteins from 1149 fully sequenced proteomes (64 archaea, 553 bacterias, 201 eukaryotes and 331 viruses) from release 2011_08 of UniProt, clustered at 30% sequence identity; this identity threshold identifies structures that can be solved through homology modeling. Empirical tests on a benchmark dataset, which shares low similarity with our training dataset, show that fDETECT is competitive when compared with a comprehensive set of existing crystallization propensity predictors. Analysis of the complete proteomes reveals that using current crystallization protocols (using a cut-off equal to the median of scores for a representative set of 44,671 X-ray structures from PDB), majority of bacteria and archaea proteomes and bacterial/archaea viruses could be covered (each cluster can be solved using homology modeling from one of its members) at >60%, while eukaryotes and eukaryotic viruses could be covered at between 15 and 70+%. We also show that at least one structure could be solved for all functional and cellular component-based annotations defined in GO. Assuming that at least half of the clusters for a given annotation has to be solved, the coverage drops to 7% in viruses, while it remains very high at 80% in eukaryotes, 90% for bacteria and 95% for archaea. However, full coverage can be obtained for only 10% of annotations for archaea, and almost none for the remaining organisms. We show that combining current crystallization pipelines and homology modeling would assure good structural coverage for majority of organisms, except for most of eukaryotes and eukaryotic viruses. Achieving a high coverage of functional annotations would require further progress in the structure determination and prediction pipelines.

86. Identification of Inhibitors Blocking Interactions between HIV-1 Integrase and Human LEDGF/p75: Mutational Studies, Virtual Screening and Molecular Dynamics Simulations

Karnati Konda Reddy, Sanjeev Kumar Singh
Alagappa University, India

The HIV-1 integrase (IN) mediates integration of viral cDNA into the host cell genome, an essential step in the retroviral life cycle. Human lens epithelium-derived growth factor (LEDGF/p75) is a co-factor of HIV-1 IN plays a crucial role in HIV-1 integration. Because of its crucial role in the early steps of HIV replication, the IN-LEDGF/p75 interaction represents an attractive target for anti-HIV drug discovery. In this study, the LEDGF/p75 binding pocket of IN interaction was studied by in silico mutational studies using molecular dynamics simulations. The results showed that the IN mutations (Q168A, E170A, H171A and T174A) in the α4/5 connector impaired the interaction with LEDGF/p75. All the crucial residues identified in mutational studies were identified in as the binding site residues. We screened ChemBridge database through three different protocols of docking simulations of varying precisions and computational intensities. We have selected six compounds analyzing the interactions with the important amino acid residues of IN, binding affinity and pharmacokinetic parameters. Finally, we performed MD simulations for a time scale of 10ns each, to examine molecular interactions between protein-ligand complexes. Results show the stable binding of compounds at the α4/5 connector of HIV-1 IN. These finding could be helpful for blocking IN-LEDGF/p75 interaction, provides a method of avoiding viral resistance and cross-resistance.

87. DNA Motif Elucidation using Belief Propagation

Ka-Chun Wong
University of Toronto, Canada

Protein Binding Microarray (PBM) is a high-throughout platform that can measure the DNA binding preference of a protein in a comprehensive and unbiased manner. A typical PBM experiment can measure binding signal intensities of a protein to all the possible DNA k-mers (k=8$\sim$10); such comprehensive binding affinity data usually need to be reduced and represented as motif models before they can be further analyzed and applied. Since proteins can often bind to DNA in multiple modes, one of the major challenges is to decompose the comprehensive affinity data into multimodal motif representations. Here we describe a new algorithm that uses Hidden Markov Models (HMMs) and can derive precise and multimodal motifs using belief propagations. We describe an HMM-based approach using belief propagations (kmerHMM), which accepts and preprocesses PBM probe raw data into median binding intensities of individual k-mers. The k-mers are ranked and aligned for training an HMM as the underlying motif representation. Multiple motifs are then extracted from the HMM using belief propagations. Comparisons of kmerHMM with other leading methods on several datasets demonstrated its effectiveness and uniqueness. Especially, it achieved the best performance on more than half of the datasets. In addition, the multiple binding modes derived by kmerHMM are biologically meaningful and will be very useful in interpreting other genome-wide data such as those generated from Chip-Seq.

Note: The above abstracts are ordered by topic and the presenting author's last name. The information in the above section has been printed as provided by the authors.

Awards

The outstanding poster and oral presentation of the 9th ISCB Student Council Symposium will be recognized and awarded with the support of our sponsors. All awards this year have been made possible by the sponsorship of Oxford University Press.

OXFORD
UNIVERSITY PRESS

We thank Oxford University Press for their continued support of the Student Council Symposium.

Best Presentation Award

This award will acknowledge the Best Oral Presentation by a student at the Symposium. All Symposium delegates will be asked to vote for their favorite presentation. The winner will be determined by the vote and the opinions of a jury composed of ISCB Student Council Leaders. The Best Presentation Award this year is sponsored by Oxford University Press journal *Bioinformatics*.

Value of the award: 500 USD

Previous winners:

2012 – Kyle Bemis

2011 – Amit Deshwar

2010 – Geoff Macintyre

2009 – Nils Gehlenborg

BIOINFORMATICS
WWW.BIOINFORMATICS.OXFORDJOURNALS.ORG

Best Poster Award

The best poster of the Symposium will be chosen through voting by the delegates at the Symposium and by the opinions of a jury of ISCB Student Council Leaders. In addition to the Best Poster Award, a Best Poster Runner-Up Award will also be given out. The Best Poster Awards have been made possible by Oxford University Press journal *Nucleic Acids Research*.

Value of the Best Poster Award: 350 USD

Value of the Runner-Up Award: 150 USD

Nucleic Acids Research

Previous winners:

2012 - Ignacio S. Caballero & Gracia Bonilla

2011 - Benjamin Kwan (1st), Emre Guney (2nd), and Mrinal Mishra (3rd)

2010 – Mark McDowall and Wouter Meuleman (Runner-Up)

2009 – Jose Caldas

2009 – Inken Wohlers and Nikolay Samusik (Runner-Up)

ISCB Student Council Travel Fellowships

The ISCB Student Council has teamed up with this year's sponsors to give several students the opportunity to attend the 9th ISCB Student Council Symposium and ISMB 2013 in Berlin, Germany. Thanks to generous support from our sponsors, we were able to award nine Travel Fellowships this year.

Don't miss the awards ceremony, to be held together with the ISCB Student Council OpenBusiness Meeting on Sunday, July 21, 12:45 p.m. - 2:00 p.m. - Room Hall 4/5. See page 75 for more information.

Congratulations to all the Travel Fellowship awardees!

Acknowledgements

The success of an event the size of the ISCB 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:

Without the logistical support and invaluable advice of ISCB Executive Director **Diane E. Kovats** and ISMB 2013 conference organizer **Steven Leard**, the 9th ISCB Student Council Symposium would not have been possible. We deeply appreciate their continued support of the ISCB Student Council and the Symposium.

We are also greatly indebted to ISMB 2013 conference chairs **Dr. Burkhard Rost, Dr. Anna Tramontano** and **Dr. Martin Vingron** for giving us the opportunity to have the 9th ISCB Student Council Symposium in Berlin. 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. Gonçalo Abecasis, Dr. Alex Bateman and Dr. Satoru Miyano.** 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 program committee, without them, there would be no Symposium! All of our reviewers did a fantastic job and it's due to them that we stayed within our set deadlines.

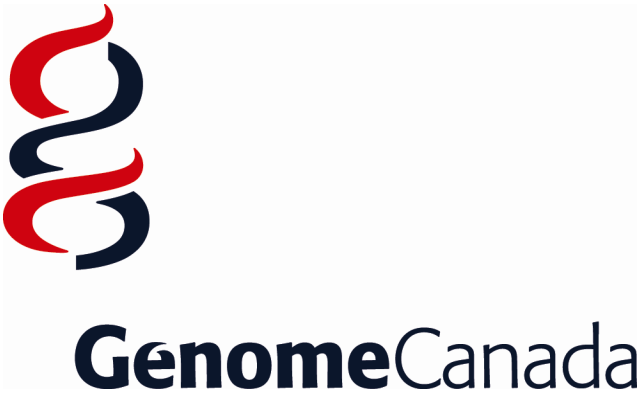
We are extremely grateful for the financial support that we received from our sponsors. Without their help many of the exciting opportunities that we offer to the delegates at the 9th ISCB Student Council Symposium would not have been possible.

Thank you all!

Sponsors

We thank our sponsors for sharing our vision and helping to make the 9th ISCB Student Council Symposium a success.

Specifically, we are grateful to:



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

Other Student Council Activities at ISMB/ECCB 2013

Art and Science Exhibition

ISMB 2013 brings together scientists from a wide range of disciplines, including biology, medicine, computer science, mathematics and statistics. In these fields we are constantly dealing with information in visual form: from microscope images and photographs of gels to scatter plots, network graphs and phylogenetic trees, structural formulae and protein models to flow diagrams; visual aids for problem-solving are omnipresent.

Often these visual aids are limited and provide nothing more than a small clue to the solution of the problem. But then there are special ones that make the whole more than the sum of its parts. Ones that combine outstanding beauty and aesthetics with deep insight that perfectly proves the validity of the scientific approach or goes beyond the problem's solution. Ones that surprise and inspire us through the transition from science to art, ones that open our eyes and minds to reflect on the work we are doing.

The Student Council is currently working closely with the founder of the Arts and Science Exhibition, Dr. Milana Frenkel-Morgenstern, to take over organization of the Arts and Science Exhibition. Come and witness the union of art and science! If you are interested in helping out or want news about the 2013 Exhibition, check out: http://symposium.iscbsc.org/content/art-science-exhibition

ISCB Student Council Open Business Meeting and Awards Ceremony

Sunday, July 21, 12:45 p.m. - 2:00 p.m. - Room Hall 4/5

We are pleased to invite you to the annual ISCB Student Council Open Business meeting. This meeting is for people who want to learn more about how the Student Council operates, how they can get involved, get updates on current Student Council activities, and find out where the Student Council will head in the future.

This year the Student Council Open Business meeting will be held in conjunction with the ISCB Open Business meeting. We strongly encourage you to come along to both meetings and learn how each of the organizations work.

Input and feedback from the community are very important to us and we hope that you will join us for this meeting. For those interested in getting involved in ISCB Student Council activities there will be mentors available who can answer your questions about the ISCB Student Council, and tell you how you can contribute to our community.

Agenda:

Report from the ISCB Student Council Leadership – Overview of recent developments within the ISCB Student Council

Contributing to the SC – Learn about contributing to the ISCB Student Council

Regional Student Group initiative – Presentations by Regional Student Groups members

9th ISCB Student Council Awards Ceremony - Presentation of the winners of the Best Poster and Best Presentation Awards as well as the recipients of the Student Council Travel Fellowships.

Outlook – Overview of the plans for the coming months and a preview of the 9th ISCB Student Council Symposium.

Feedback and Discussion – An opportunity to voice your opinion about the efforts of the ISCB Student Council and the direction it should be taking. The ISCB Student Council is counting on your input!

Student Council Career Central

The ISCB Student Council is committed to helping students develop successful careers in the field of Computational Biology and Bioinformatics. In addition to our Symposium and Regional Student Group initiative, the Student Council hosts a special booth on the exhibitor floor during ISMB that is dedicated specifically to helping students and post-docs with career development.

Please come visit us at **Booth 1**, where you will find a job posting board and tons of advice on how to navigate the job market in this exciting new field. You can also participate in our **interactive job posting board**, where you can sign-up to meet potential employers or supervisors who are also present at the conference. Don't forget to bring your CV!

Are you looking for a job, internship, PhD or post-doc? Then come along to the **Student Council Career Session on Monday, July 22, 5:40 pm - 6:40 pm**. The Student Council will welcome **Dr. Jaap Heringa** of Vrije Universiteit Amsterdam, The Netherlands, who will be giving a **talk on careers in computational biology and bioinformatics and factors to consider when choosing a career in these fields**. You will also have the opportunity to interact with early career scientists and to hear about their experiences. Join us and you can ask your questions on obtaining the ultimate job or research position! We look forward to seeing you there!

Student Council Symposium Organizing Committee

Tomás Di Domenico, Student Council Symposium Chair
BioComputing lab, Department of Biology
University of Padua, Italy

Tomasz Stokowy, Student Council Symposium Co-Chair
Automatic Control Department
Silesian University of Technology, Poland

Emre Guney, Program Chair
Structural Bioinformatics Labaratory
IMIM / UPF – GRIB, Spain

Esmeralda Vicedo, Travel Fellowships Chair
Department for Bioinformatics and Computational Biology
Technical University Munich, Germany

Cynthia Prudence, Finances Chair
Division of Biological and Medical Physics, Physics Department
University of Rhode Island, Kingston, USA

Malay Bhattacharyya, Keynotes Chair
Machine Intelligence Unit
Indian Statistical Institute, India

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

Jörgen Brandt, Social Event
Knowledge Management in Bioinformatics
Humboldt Universität zu Berlin, Germany

Canan Has, , Social Event
Molecular Biology and Genetics Department
Izmir Institute of Technology, Turkey

Himanshu Joshi, Social Event
Institute for Clinical Medicine
University in Oslo, Norway

Sayane Shome, Booklet
School of Biosciences and Technology
Vellore Institute of Technology ,India

Gokmen Zararsiz, Booklet
Department of Medical Statistics
Erciyes University, Turkey

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.

For a complete listing of ISCB conferences visit:
www.iscb.org/iscb-conferences