

Contents

Welcome	3
Program of the day	5
Keynotes	13
Research and Industry Partner's Talk	17
Abstracts	
Oral Presentations	21
Poster Presentations	31
Travel Fellowships & Awards	81
Student Council Report	85
Organisers' Bio	93



Organising Committee

Chair	Katie Wilkins, Cornell University, United States of America
Co-Chair	Farzana Rahman, University of South Wales, United Kingdom
Finance Chair	Jakob Berg Jespersen, Massachusetts General Hospital, United States of America
Program Chair	R Gonzalo Parra, Buenos Aires University, Argentina
Web Chair	Mehedi Hassan, University of South Wales, United Kingdom
Travel Fellowship Chair	Margherita Francescato, German Center for Neurodegenerative Diseases, Germany
Communications Chair	Bart Cuypers, University of Antwerp, Belgium
Social Event Co-ordinator	Anupama Jigisha, University College Dublin, Ireland

Peers and Organising Volunteers

The following members of the ISCB Student Council community have acted as peers and volunteers in different phases of organising the event.

Alexandre Borrel	University of Helsinki, Finland and University Paris Diderot, France
Alexander Junge	University of Copenhagen, Denmark
Alexander Monzon	National University of Quilmes, Argentina
Arjun Ray	CSIR-Institute of Genomics & Integrative Biology, India
Aziz Khan	Tsinghua University, China
Carla P Franzotti	National University of Tucuman, Argentina
Dan DeBlasio	University of Arizona, United States of America
Ezequiel IJ Davies	The Bordeaux Bioinformatics Center, France
Inamul Hasan Madar	Korea University, Korea
Julien Fumey	Université Paris-Sud, France
Melissa Woghiren	University of Windsor, Canada
Nazeefa Fatima	University of Huddersfield, United Kingdom
Nicolas Palopoli	National University of Quilmes, Argentina
Pieter Meysman	University of Antwerp, Belgium
R. Taylor Raborn	Indiana University, United States of America
Salvador C Gutierrez	CBS Fungal Biodiversity Centre, The Netherlands.
Sayane Shome	VIT University, India
Zoe Dyson	La Trobe University, Australia

First print, July 2015

Copyright © 2015 ISCB Student Council and contributing authors. All rights reserved.

Published by the International Society for Computational Biology (the ISCB)

This booklet was designed and edited by Mehedi Hassan and R Gonzalo Parra.

Contents were prepared in June 2015.

For latest updates please visit the symposium page at www.scs2015.iscb.org

This work is licensed under the Creative Commons Attribution 4.0 International License. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>. This booklet may be reproduced without permission in its original form.

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.

Welcome to the 11th Student Council Symposium

The ISCB Student Council is pleased to welcome you to the 11th edition of the International Society for Computational Biology Student Council Symposium in Dublin, Ireland. After the success of previous Symposia, which took place in Boston (2014), Berlin (2013), Long Beach (2012), Vienna (2011), Boston (2010), Stockholm (2009), Toronto (2008), Vienna (2007), Fortaleza (2006), and Madrid (2005), we are thrilled to once again offer our delegates the opportunity to meet students and young scientists from all over the world. As usual, our primary goals are to promote the exchange of ideas and to foster the creation and strengthening of scientific social networks.

We are honored to have **Prof. Des Higgins** (University College Dublin, Ireland) and **Prof. Ruth Nussinov** (Tel Aviv University, Israel and National Cancer Institute, USA) 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 are also very excited to have Dr. Robert Davey (TheGenome Analysis Centre) and Mr. Michael Markie (Faculty of 1000) presentsome very pertinent topics to the symposium attendees.

We will start the Symposium with a session of “**Scientific Speed Dating**”, a chance to get to know your peers in an informal and friendly way. Throughout the day we will hear **oral presentations** from a selection of **12 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. Light refreshments will be offered during the poster reception session. Everyone involved in the organization of this Symposium contributed significantly to make this event happen. Our volunteers have spent months preparing all aspect of this Symposium ranging from the invitation of keynote speakers, fundraising, advertising and organizing the peer-review process to more technical aspect like designing the website.

We encourage you to make the most out of this opportunity and to be very active in engaging other delegates, asking questions, discussing ideas and showcasing your own research. You can make this Symposium a starting point for fruitful future collaborations and another step towards a successful career in computational biology.

Do not forget to visit the **Student Council Lounge** (Booth 25, July 11 - 14) to learn about the Student Council and its activities and how you can be part of this great effort. We encourage you to attend the ISCB Student Council's Career Central that stimulates discussions and opinions on careers in bioinformatics. This year the ISCB Student Council is collaborating with the Junior PI group and COBE COSI on an **Applied Knowledge Exchange Sessions** workshop dedicated on career development at ISMB/ECCB 2015 (Liffey Meeting Room 3, AKES 06 session, 1:30 PM on Saturday, July 11). We would also like to invite you to our **Open Business Meeting** (Room: Wicklow Hall 2A, 12:45 – 1:50 PM on Monday, July 13), the **Art and Science Exhibition** (Booth 12-13, July 12 -14) and, of course, our **Social Event** (Location: “Bar With No Name”, 7:00 PM onwards on Friday July 10)! We promise you that our social events are fun.

Enjoy your time in Dublin!

Thank you

On Behalf of the organisers,

Katie Wilkins, Student Council Symposium Chair

Farzana Rahman, Student Council Symposium Co-Chair

Pieter Meysman, Student Council Chair

Tweeting at the Symposium

Official Hashtag #SCSDublin2015

Follow us on  @iscbsc

Program of the Day

Time	Event/Activity	Page
08:00-08:30	Registration	
08:30-08:45	Icebreaker: Scientific Speed Dating	8
08:45-09:00	Welcome and opening	
09:00-09:40	Keynote: A Mechanistic Structural View of Ras Biology by <u>Prof. Ruth Nussinov</u> , National Cancer Institute, USA and Tel Aviv University, Israel	13
Oral Presentations - Session I		
09:40-10:20	Prioritizing a drug's targets using both gene expression and structural similarity , by <u>Griet Laenen</u> , KU Leuven, Belgium	28
	Organism specific protein-RNA recognition: A computational analysis of protein-RNA complex structures from different organisms by <u>Nagarajan Raju</u> , Indian Institute of Technology Madras, India	28
10:20-10:50	Coffee Break	
Oral Presentations - Session II		
10:50-11:50	Jabba: Hybrid Error Correction of Long Sequencing Reads by <u>Giles Miclotte</u> , University of Ghent, Belgium	29
	Genome-wide ceRNA networks by <u>Mario Flores</u> , University of Texas, USA	29
	Detection of Heterogeneity in Single Particle Tracking Trajectories by <u>Paddy Slator</u> , University of Warwick, UK	30
	<i>Research and Industry Partner Talk</i>	
11:50-12:00	Publishing in the digital age – introducing the ISCB Community Journal by <u>Mr. Michael Markie</u> , Faculty of 1000 Limited, UK	17
	Lunch and Poster Session	
12:00-13:40	<i>Research and Industry Partner Talk</i>	
13:40-14:00	Open science needs open scientists: an ever-increasing interdependence by <u>Dr. Robert Davey</u> , The Genome Analysis Centre (TGAC), UK	19

Program of the Day (*continued*)

Time	Event/Activity	Page
14:00-14:40	Keynote: Clustal Omega and multiple sequence alignments by <u>Prof. Des Higgins</u> , Conway Institute, University College Dublin, Ireland	15
Oral Presentations - Session III		
14:40-15:40	3D-NOME: 3D NucleOme Multiscale Engine for data-driven modeling of three-dimensional genome architecture by <u>Przemek Szalaj</u> , Medical University of Bialystok, Poland	30
	eFORGE: a tool for identifying tissue-specific signal in epigenetic data by <u>Charles Edmund Breeze</u> , University College London, UK	31
	Co-regulation of human paralog genes in the three-dimensional chromatin architecture by <u>Jonas Ibn-Salem</u> , Johannes Gutenberg University Mainz, Germany	31
15:40-16:00	Coffee Break	
Oral Presentations - Session IV		
16:00-17:20	A novel feature selection method to extract multiple adjacent solutions for viral genomic sequences classification by <u>Giulia Fiscon</u> , Sapienza University of Rome, Italy	32
	Systematic integration of molecular signatures identifies novel components of the antiviral RIG-I-like receptor pathway by <u>Robin van der Lee</u> , Radboud University Medical Center, The Netherlands	32
	Understanding <i>leishmania</i> development and drug resistance using an integrative 'omics compendium by <u>Bart Cuypers</u> , University of Antwerp, Belgium	33
	Unravelling signal regulation from large scale phosphorylation kinetic data by <u>Westa Domanova</u> , University of Sydney, Australia	33
17:20-17:40	Closing remarks	
17:40-18:40	Final Poster Session & Refreshments	

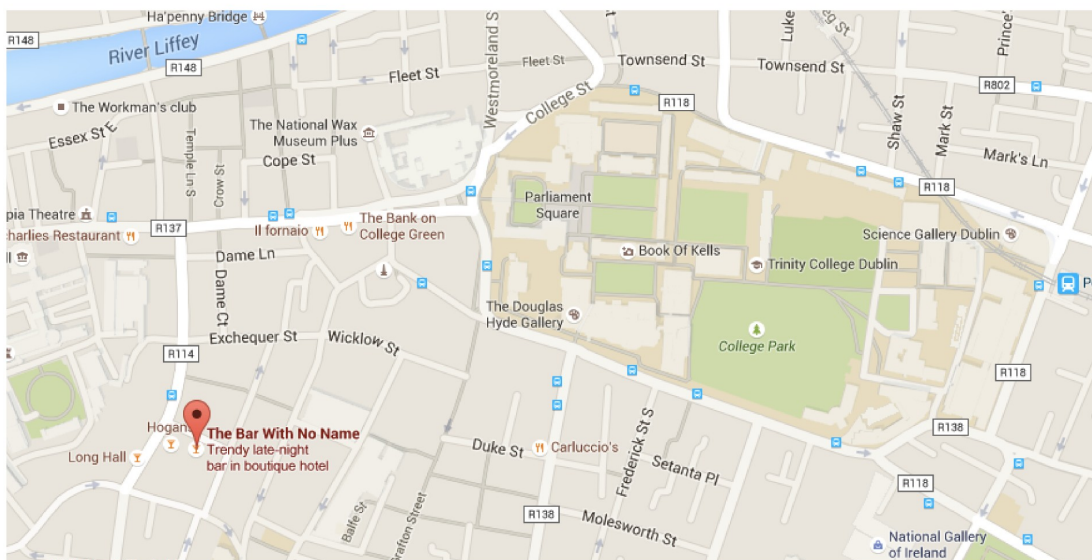
Join us at our social event!



“Bar with no name” is located on 3, Fade street, Dublin 2.

The bar does not have a name literally! Look out for a large wooden **snail hanging** on the outside (see picture above)!

We will be in the ‘red room’ of the bar **7pm onwards on July 10th!** See you there.



We thank our social event sponsor Bina technologies

bina

Scientific Speed Dating

We are continuing the tradition of 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!



Interactive Session at SCS2014, Boston, USA

Art and Science Exhibition

ISMB/ECCB Dublin, 2015 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.

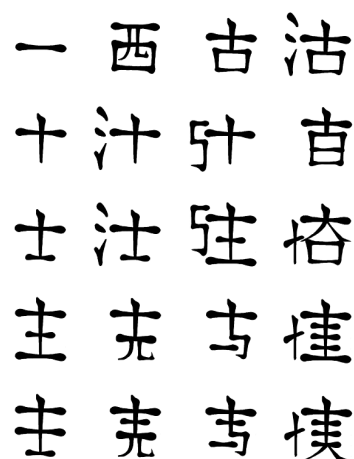
Often these visual aids are limited and provide nothing more than a small clue to the solution of the problem. But then there are special ones that make the whole more than the sum of its parts. Ones that combine outstanding beauty and aesthetics with deep insight that perfectly proves the validity of 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. Organized by a team headed by Dr. Milana Frenkel-Morgenstern, the Arts and Science exhibition allows you to witness the union of art and science.

If you are interested in helping out or want news about the 2015 Exhibition, check out:
<http://www.iscb.org/ismbecb2015-submission/ismbecb2015-artscience>

A glimpse from past exhibitions



The Fly (Insulin Dimer) by Maja Klevanski,
University of Heidelberg, Germany
Winner of The winner of Art & Science
at the ISMB 2011



Twenty Characters by Kristian Rother,
IIMCB Warsaw, GERMANY.
Winner of The winner of Art & Science
At the ISMB 2009

Blast from the past



During coffee break at SCS2014, Boston, USA

Career Central

Every year at its annual symposium, the ISCB Student Council organises 'Career Central' to present discussions, opinions and talks related to careers in bioinformatics. This year at ISMB/ECCB 2015, we are collaborating with the Junior PI group and COBE COSI on an AKES workshop dedicated to career development (AKES 06). This workshop will be of interest to a wide audience, from students and junior PIs, looking for new ways to build their careers, to senior faculty, looking for advice on how to mentor your students and get the best out of them.



The Student Council is proud to be organising the first session “Improving your elevator pitch” presented by Dr Mick Watson, Director of ARK-Genomics, The Roslin Institute, University of Edinburgh, UK.

Learn more about Dr. Mick Watson on his website (<http://www.roslin.ed.ac.uk/mick-watson/>) or at his blog ‘opiniomics’ (<https://biomickwatson.wordpress.com>).

You’ve just bumped into the person who could decide your next career move. They are about to head to their next meeting and you have two minutes to convince them that your science is worthy. Can you do it?

In this session you will learn the theory behind how to deliver a great elevator pitch. You will then get the opportunity to put this theory to the test, honing and practicing your pitch with other audience members so that the next time you hop in an elevator, you will be ready.

Mark the date and time - July 11, 1:30pm.

Hope to see you all there!

Blast from the past



At SCS2013, Berlin, Germany

Keynote Speakers



Prof. Ruth Nussinov

*Cancer and Inflammation Program,
National Cancer Institute, USA and
Medical School, Tel Aviv University, Israel*

Our first keynote speaker of the day is Prof. Ruth Nussinov. She has an intense career with many achievements. After completing her Ph.D. in Biochemistry from Rutgers University, she became a fellow at the Weizmann Institute. She was a visiting Scientist at the University of California, Berkeley and Harvard University. In 1985, she joined Tel Aviv University as Associate Professor and became Full Professor in 1990. Her association with the NCI also initiated in 1985.

Nussinov's 1978 paper proposed the dynamic programming algorithm for RNA secondary structure prediction which is to date the leading method in this field. She also pioneered DNA sequence analysis in the early 1980s. In 1999, her NCI group proposed the model of 'conformational selection and population shift' as an alternative to 'induced fit' to explain molecular recognition. This paradigm has impacted the scientific community's views and strategies in drug design, biomolecular engineering, and molecular evolution.

Ruth Nussinov was a recipient of the 2011 Biophysical Society Fellow Award for her extraordinary contributions to advances in computational biology on both nucleic acids and proteins. She has coauthored over 440 scientific papers and is highly cited. Her research mainly focuses on protein structure, dynamics, function, protein–protein interactions, and cellular signaling.

Keynote Title: A Mechanistic Structural View of Ras Biology

Abstract

Ras proteins are small GTPases that act as signal transducers between cell surface receptors and several intracellular signaling cascades. KRas4B is among the frequently mutated oncogenes in human tumors. Ras proteins consist of highly homologous catalytic domains, and flexible C-terminal hypervariable regions (HVRs) that differ significantly across Ras isoforms. We have been focusing on key mechanistic questions in Ras biology from the structural standpoint. These include whether Ras forms dimers, and if so what is their structural landscape; how do Ras dimers activate Raf, a key Ras effector in a major signaling pathway; what is the role of calmodulin in Ras signaling, what is the potential regulatory role of the hypervariable region and its membrane anchoring and what are the mechanisms of oncogenic mutations. We believe that structural biology, computations and experiment, are uniquely able to tackle these fascinating and important questions.

Blast from the past



At SCS2012, Long Beach, California, USA

Keynote Speakers



Prof. Des Higgins

*School of Medicine & Medical Science,
Conway Institute, University College Dublin, Ireland*

Our second keynote speaker of the day is Prof. Des Higgins. He has a PhD in Zoology from Trinity College Dublin, Ireland. His lab currently maintains and develops the Clustal package for multiple sequence alignment in collaboration with groups in France, Germany and the UK. He wrote the first version of *Clustal* packages in 1988 and then moved to the EMBL Data Library group, in Heidelberg, in 1990 as a post-doc and later, staff scientist. In 1994, he moved to the EBI, Hinxton with the Data Library and stayed there for two years. This coincided with the release of Clustal W, and later, Clustal X which became extremely widely used and cited. Currently his team is working on Clustal Omega which is designed for making extremely large protein alignments.

Keynote Title: Clustal Omega and multiple sequence alignments

Abstract

Multiple Sequence Alignments (MSA) are used to take a set of related DNA or protein sequences and line them up so as to make them easy to compare to each other. Most MSAs are made using a range of related heuristics that involve clustering the sequences and building an alignment that follows the clusters. These methods have served us well for the past 25 years but are now starting to creak. I will describe a new program called Clustal Omega which can make alignments of very large numbers of sequences. It gives good quality alignments in reasonable times and has extensive features for adding new sequences to or for exploiting information in existing alignments. It is available for download from www.clustal.org in a command-line driven format (Linux style) for proteins only. It is also available for on-line use from the EBI.

ISCB Community Journal

Immediate & Transparent Publishing

Dedicated to the official and affiliated ISCB conferences
and the ISCB's Communities of Special Interest (COSIs)



ARTICLES



POSTERS



SLIDES

For more information, visit booth 21, or the channel page:

f1000research.com/iscbcommj

Research and Industry Partner Presentation



Mr. Michael Markie

*Associate publisher,
F1000Research, Faculty of 1000 Ltd, UK*

Michael Markie is an open science and open data advocate and is currently the associate publisher for *F1000Research*, an authored open science platform that offers immediate publication, transparent peer review (post-publication) and full data deposition and sharing. Michael's role at F1000 is to help build and innovate new digital solutions for sharing scientific research, and in particular spends a lot of time doing academic outreach and collaborating with the scientific community to identify new ways to improve how scientific research is disseminated and made publically available.

Talk Title: Publishing in the digital age – introducing the ISCB Community Journal

Abstract

Despite moving from paper to pixels, the publication of scientific research has not harnessed the full potential of the internet and remains remarkably similar to the first journal published 350 years ago. Journals at large still use a system that's rooted in the pre-web era of print and they are often criticised for stifling academic progress. The web provides the opportunity to communicate science much more efficiently and allows us to address some of the most talked about issues around journals such as anonymous peer review, delays in publication, and reproducibility. *F1000Research* is an online only open science platform that sets out to tackle these issues through immediate and transparent publishing, and is now collaborating with the ISCB to publish the *ISCB Community Journal*, which will support the open publication of research in a range of formats from select ISCB conferences and COSIs.

Research and Industry Partner Presentation



Dr. Robert Davey

*Group Leader, Data Infrastructure & Algorithms ,
The Genome Analysis Centre (TGAC), United Kingdom*

Dr. Davey joined TGAC in February 2010 as the lead software engineer on the MISO lab information management (LIMS) project, which was released to the community as an open source framework for tracking sequencing experiments in 2012. He went on to become the Core Bioinformatics Project Leader, managing a team of developers to advance MISO as well as new projects into data infrastructure and management, and the genomic data visualisation tool, TGAC Browser. Robert was appointed as Data Infrastructure and Algorithms Group Leader in late 2012. Robert's main interests are in enterprise-grade software development, data management and associated HPC infrastructure, sequence analysis and quality control pipelines, novel visualisation strategies for sequencing and biological data, metadata and the Semantic Web, and the open source ethos. Prior to joining TGAC, he was a post-doctoral researcher at the Institute of Food Research (IFR) in the National Collection of Yeast Cultures (NCYC) group, providing analytical tools and bioinformatics support to help drive this important national capability. He completed his degree in Microbiology and his PhD in Bioinformatics at the University of East Anglia (UEA), the latter developing algorithms for assessing the gene content of bacterial organisms using Comparative Genomic Hybridisation microarrays.

The Data Infrastructure and Algorithms group focuses on research into understanding how best to manage, represent and analyse data for open science, as well as exploring new hardware, algorithms and methodologies to develop tools to push the boundaries of data-driven informatics in the life sciences. The team applies their research expertise to develop infrastructure platforms for data and software dissemination and publication, assembly algorithms for viral and microbial metagenomics, large-scale data visualisation, and best practice and training in bioinformatics.

Talk Title: Open science needs open scientists: an ever-increasing interdependence

Abstract

In recent years, scientific research has experienced an interesting juxtaposition. There is increasing pressure from funding bodies to make research data accessible. Researchers also need the increasingly sensational track records published in high-impact journals to ensure continued project support and/or tenure.

Pressure to release data by funders, whilst obviously a step in the right direction, represents a formidably large stick but a depressingly small carrot which results in simply another tedious hurdle to getting research published rather than a vehicle to get recognition. The constant push for papers in journals with perceived impact and prestige, whilst still seen as a key assessment mechanism for a researcher's career, promotes a closed door approach and a touch of paranoia, with research becoming a competitive endeavour rather than a mutually beneficial collaborative one.

Thankfully, a new breed of researchers at all career stages, from graduate to PI, who can see these mutual benefits of sharing their work openly are becoming greater in number and more vocal by the day. Open source code, open data, powerful tools and infrastructure, social networks, and open access publishing all play a part in the ecosystems of the Open Science movement.

The Genome Analysis Centre

Building Excellence in Genomics and Computational Bioscience

Visit us at www.tgac.ac.uk



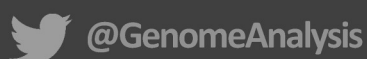
We are a world-class research institute focusing on the development of **genomics** and **computational biology**.

A UK hub for innovative bioinformatics through research, analysis and interpretation of multiple, complex data sets, TGAC offers:

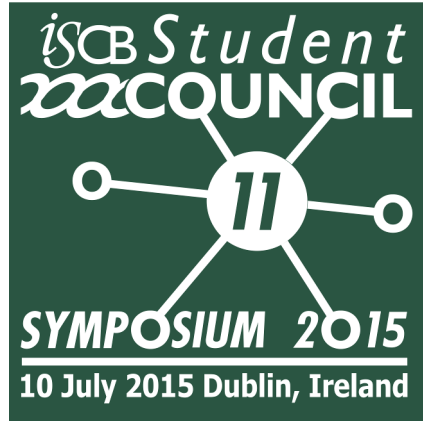
A state of the art DNA sequencing facility

Multiple complementary technologies for data generation

One of the largest computing hardware facilities dedicated to life science research in Europe



Contact us at tgac.enquiries@tgac.ac.uk

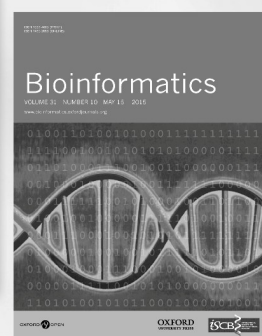
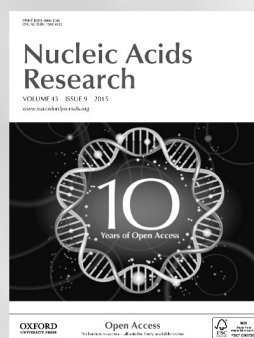


Oral Presentation Abstracts

Out of 93 works received from researchers from 23 countries,
12 works have been selected for oral presentation at SCS2015.

Bioinformatics and Nucleic Acids Research

are proud to sponsor
this year's ISCB
Student Council
best poster and best
presentation prizes.



Stop by OUP's booth for free journal
copies and promotional materials from
Bioinformatics and *Nucleic Acids
Research*, as well as *Database*, *Briefings
in Bioinformatics*, *Briefings in Functional
Genomics*, *Human Molecular Genetics*,
and *Molecular Biology and Evolution*.

OXFORD
UNIVERSITY PRESS

Oral Presentation Abstracts

Session I

1. Prioritizing a drug's targets using both gene expression and structural similarity

Griet Laenen, Sander Willems, Daniela Börnigen, Amin Ardeshtirdavani, Lieven Thorrez, Yves Moreau
KU Leuven

The pharmaceutical industry is facing unprecedented pressure to increase its productivity. Attrition rates in the later stages of drug development have risen sharply, with toxicity and lack of efficacy being the main bottlenecks. To address both these safety- and efficacy-related issues, a better understanding of the complex biological response to drug treatment is vital. Although many drugs exert their effects through the modulation of multiple targets, these targets are often unknown and identification among the thousands of gene products remains difficult. We propose a computational method to support this identification of putative targets of a drug. The first component of our method prioritizes proteins as potential targets by integrating experimental gene expression data with prior knowledge on protein interactions. More specifically, genes are ranked based on the transcriptional response of functionally related genes by diffusing differential expression signals following treatment over a protein interaction network. The second component of our method prioritizes proteins as drug targets based on the interaction with compounds structurally similar to the drug of interest. To this end compound-compound similarity scores are combined with compound-protein interaction scores. Our method has been evaluated on a test set of small molecule drugs for which the known targets were derived from ChEMBL. AUC values of up to

90% were obtained. These AUC values indicate the predictive power of combining gene expression data and structural information for a drug of interest with known protein-protein and protein-compound interaction information respectively, to identify the targets of that drug.

2. Organism specific protein-RNA recognition: A computational analysis of protein-RNA complex structures from different organisms

Nagarajan Raju, Sonia Pankaj Chothani, Ramakrishnan C, Sekijima M, Gromiha MM

Indian Institute of Technology Madras

Background: Understanding the recognition mechanism of protein-RNA complexes has been a challenging task in molecular and computational biology. In this work, we have constructed 18 sets of same protein-RNA complexes belonging to different organisms. The similarities and differences in each set of complexes have been revealed in terms of various sequence and structure based features such as root mean square deviation, sequence homology, propensity of binding site residues, conservation at binding sites, binding segments and motifs, preferred amino acid-nucleotide pairs and influence of neighboring residues for binding.

Description and Results: We found that the proteins of mesophilic organisms have more number of binding sites than thermophiles and the binding propensities of amino acid residues are distinct in *E. coli*, *H. sapiens*, *S. cerevisiae*, thermophiles and archaea. Proteins prefer to bind with RNA using a single residue segment in all the

organisms whereas RNA prefers to use a stretch of up to six nucleotides for binding with proteins. We developed amino acid residue-nucleotide pair potentials for different organisms, which could be used for predicting the binding specificity. Further, molecular dynamics simulation studies on aspartyl tRNA synthetase complexed with aspartyl tRNA showed specific modes of recognition in *E. coli*, *T. thermophilus* and *S. cerevisiae*. Conclusion: Based on the structural analysis and molecular dynamics simulations we suggest that the mode of recognition depends on the type of the organism in a protein-RNA complex.

Session II

3. *Jabba: Hybrid Error Correction of Long Sequencing Reads*

Giles Miclotte, Mahdi Heydari, Pieter Aude-naert, Piet Demeester, Jan Fostier
Ghent University

Background: Third generation sequencing techniques produce longer reads with higher error rates than second generation methods. While the improved read lengths can provide useful information for downstream analysis, the higher error rates can complicate the required mapping or alignment. Hybrid strategies have been proposed to correct the long reads using accurate short reads. Mapping short reads on long reads may eliminate up to 99% of all errors in bacterial datasets, however this requires significant amounts of computing resources. Mapping the long reads on a k-mer frequencies based de Bruijn graph is significantly more efficient, but loses some accuracy on larger genomes.

Description: We present Jabba, a hybrid method to correct long reads by mapping them on a corrected de Bruijn graph. First, accurate second generation reads are used to build a de Bruijn graph, which is then corrected based on standard topological graph correction methods. Finally a path in the

graph is then found by using maximal exact matches between a long erroneous read and the nodes of the de Bruijn graph. This path then dictates the corrected sequence.

Conclusions: Jabba achieves comparable gain to other available tools for bacteria and other small genomes. For larger genomes Jabba keeps performing well, while others either can not practically handle these at all, or only at significantly reduced gain.

4. *Genome-wide ceRNA networks*

Mario Antonio Flores
University of Texas

Postranscriptional regulation of gene expression can be modeled as a competitive endogenous RNA (ceRNA) network in which mRNAs compete for miRs binding. Previous research shows that this competition maintains and fine-tunes levels of protein coding genes and the disruption of the network contributes to phenotypic conditions like cancer. Based on our previous studies we provided a tool (TraceRNA) for reconstruction of ceRNA networks around a gene of interest (GoI). The approach used in TraceRNA although practical and useful for gene-based studies provides only a partial landscape of the ceRNA mechanisms and phenotypes. Besides TraceRNA offers an ad-hoc approach for the study of the ceRNA phenomenon. In this work we present a formal genome-wide approach for ceRNA networks study. This novel and formal treatment of the ceRNA phenomenon provides new perspectives in the study of ceRNA networks and its specific phenotype. We divide the study of genome-wide ceRNA networks in three main sections: network construction, analysis of network components by network perturbation and network stability. In the case of network construction we formalize the definition of ceRNA phenomenon. The construction of a genome-wide network of a specific phenotype (e.g. breast cancer) is modeled having as input

datasets of miR binding predictions as well as associated mRNA and miRNA expression datasets. In the case of the analysis of the main components (mRNAs, miRNAs) of the predicted ceRNA network we present an algorithm that is based on perturbation. For ceRNA network stability we present an approach using data subsampling and network stability indicators. This research was supported in part by the Intramural Research Program of the National Library of Medicine, NIH.

5. Detection of Heterogeneity in Single Particle Tracking Trajectories

Paddy Slator, Nigel Burroughs
University of Warwick

Background: Single particle tracking trajectories are fundamentally stochastic, which makes the extraction of robust biological conclusions difficult. This is especially the case when trying to detect heterogeneous movement of molecules in the plasma membrane. This heterogeneity could be due to a number of biophysical processes such as: receptor clustering, traversing lipid rafts, binding to the cytoskeleton, or changes in membrane diffusivity.

Description: Working in a Bayesian framework, we developed multiple models for heterogeneity, such as confinement in a harmonic potential well, and switching between diffusion coefficients. We analyse these models using Markov chain Monte Carlo algorithms, which infer model parameters and hidden states from single trajectories. We also calculate model selection statistics, to determine the most likely model given the trajectory. Our methodology also accounts for measurement noise. For LFA-1 diffusing on T cells we found 10-26% of trajectories display clear switching between diffusive states, depending on treatment. Analysis of the motion of GM1 lipids bound to the cholera toxin B subunit in model membranes showed transient trapping in harmonic potential wells. We have also demonstrated that allowing for measurement noise is essential,

as otherwise false detection of heterogeneity may be observed.

Conclusions: We have used Bayesian methodology to analyse single particle tracking trajectories. Rather than existing methods, which rely on generic properties of Brownian motions, our approach allows us to test which biophysical model best fits a trajectory. With the continuing improvement in spatial and temporal resolution of trajectories, these methods will be important for biological interpretation of experiments.

Session III

6. 3D-NOME: 3D NucleOme Multiscale Engine for data-driven modeling of three-dimensional genome architecture

Przemysław Szalaj, Zhonghui Tang, Oskar Luo, Paul Michalski, Yijun Ruan, Dariusz Plewczynski
Medical University of Bialystok, Poland

Human genome is folded into three-dimensional structures. The 3D organization of the genome is thought to facilitate compartmentalization, chromatin organization and spatial interaction of genes and their regulatory elements. Recently developed high-throughput ChIA-PET method allows us to capture the genome-wide map of physical contacts between distal genomic loci. We present a 3D-NOME, a multiscale computational engine we developed to model the 3D organization of the genome. First, we use a bottom-up approach to create a hierarchical model representing the nucleus based on the underlying data features. Then we use Monte Carlo simulations to sequentially reconstruct all the levels in a top-down manner, i.e. we reconstruct more general levels first and we use them to guide the simulation on the following levels. This approach allows us to efficiently model the chromatin folding on a level of whole chromosomes as well as single topological do-

mains. In our modeling we consider CTCF (which is long known to be responsible for chromatin weaving) and RNAPII (which activates genes transcription) interactions. Taken together these two protein factors provides a comprehensive map of human genome interactions. The specificity of the ChIA-PET data allows us to model the shape of individual chromatin loops and their mutual interactions. In this work we describe both the model construction and simulation steps of our algorithm. We do also highlight main advantages of our approach compared to existing methods for genome architecture modeling.

7. *eFORGE: a tool for identifying tissue-specific signal in epigenetic data*

Charles Edmund Breeze, Dirk S. Paul, Lee M. Butcher, Javier Herrero, Ewan Birney, Ian Dunham, Stephan Beck
University College London

Background: Epigenome-wide association studies (EWAS) provide a novel means of studying the epigenetic basis of human disease. A challenge confronting EWAS though is the assessment of tissue specificity of identified differentially methylated positions (DMPs). To this end, we have developed an analysis approach that determines the tissue-specific regulatory component of a set of EWAS DMPs through the detection of enrichment of overlap with DNase I hypersensitive sites (hotspots) across a wide range of tissues. Our tool, eFORGE (experimentally-derived Functional element Overlap analysis of ReGions from EWAS), is available online (<http://eforge.cs.ucl.ac.uk/>) and provides tabular and graphical summaries of the enrichments. This tool is derived from FORGE (<http://www.1000genomes.org/forg-analysis>), which analyses the tissue-specific regulatory component of SNPs in the context of genome-wide association studies (GWAS). **Description:** For a given set of significant EWAS DMPs (i.e., 450K array probes), eFORGE generates 1000 randomly selected background sets, matched for gene feature

and CpG island relationship, and calculates a binomial P-value of enrichment of overlap for each of the cell types catalogued in NIH Epigenomics Roadmap and ENCODE datasets. eFORGE provided valuable insights into the underlying disease etiology when applied to recently published EWAS. For example, eFORGE enrichments were found in CD56+ cells and thymus tissue in an EWAS for rheumatoid arthritis, and CD4+ cells in an EWAS for multiple sclerosis.

Conclusion: eFORGE provides a user-friendly tool to investigate the tissue-specific component of epigenetic marks identified through EWAS, and has the potential to reveal unforeseen tissue involvements leading to mechanistic insights for disease etiology and progression.

8. Co-regulation of human paralog genes in the three-dimensional chromatin architecture

Jonas Ibn-Salem, Miguel A. Andrade-Navarro

Johannes Gutenberg University Mainz

Introduction: Paralog genes arise from gene duplication events during evolution. The resulting sequence similarity between paralogs often leads to proteins of similar structures and functions in common pathways and protein complexes. Therefore, it can be useful for the cell to have paralogs co-regulated. In eukaryotes, genes are regulated by binding of transcription factors to distal enhancer elements, which perform looping interactions to contact the transcription machinery at gene promoters. These looping interactions can be measured by genome-wide chromatin conformation capture (Hi-C) experiments which revealed conserved megabase-sized self-interacting regions called topological association domains (TADs). We hypothesised that paralogs cluster in the three-dimensional chromatin architecture and share common regulatory mechanisms to enable coordinated expression.

Description: To test this hypothesis, we integrated paralogy annotations with genome-wide data-sets of enhancer-promoter associations, Hi-C experiments, and gene expression in diverse human cell-types. As background control we sampled random gene pairs by taking the linear distances of paralogs and the number of linked enhancers into account. We show that paralog gene pairs share significantly more common enhancer elements than expected. Furthermore, they are located significantly more often in the same TAD and contact each other more frequently than expected. Consequently, paralogs tend to show a positive correlation of gene expression over many cell-types.

Conclusion: Combined, our results indicated that human paralogs share common regulatory mechanisms and cluster not only in the linear genome but also in the three-dimensional chromatin architecture. This enables concerted expression of paralogs over diverse cell-types and indicate evolutionary constraints in functional genome organization.

Session IV

9. A novel feature selection method to extract multiple adjacent solutions for viral genomic sequences classification

Giulia Fiscon, Emanuel Weitschek, Massimo Ciccozzi, Paola Bertolazzi and Giovanni Felici

Department of Computer, Control, and Management Engineering (DIAG), Sapienza University of Rome.

Background: Leveraging improvements of next generation technologies, genome sequencing of several samples and in different conditions led to an exponential growth of biological sequences. However, these collections are not easily treatable by biologists to obtain a thorough data characterization and require a high cost-time investment. Therefore, computing strategies and specifically automatic knowledge extraction methods that optimize the analysis focusing on what data should be sequenced are essential.

Description: We present a new feature-selection algorithm based on mixed integer programming methods able to extract equivalent, multiple, and adjacent solutions for supervised learning problems applied to biological data. In particular, we focus on those problems where the relative position of a feature (i.e., nucleotide locus) is relevant. In this connection, we aim to find sets of discriminating features, which are as close as possible to each other. Our algorithm has been successfully integrated in a rule-based classification framework and applied to three viral datasets (i.e., Rhino-, Influenza-, Polyomaviruses). We succeeded in extracting a wide set of equivalent separation rules focusing on small regions of sequences with high accuracy and low computational time.

Conclusions: Our algorithm enables to extract all the alternative classification solutions of virus specimen to species assignments, by identifying multiple portions of

sequence that are distinctive, compact, and as shorter as possible, in order to provide the biologists with small genome parts to be sequenced. Finally, we obtain advantages in term of sequencing cost and time, as well as a powerful instrument both scientifically and diagnostically (i.e., automatic virus detection).

10. Systematic integration of molecular signatures identifies novel components of the antiviral RIG-I-like receptor pathway

Robin van der Lee, Robin van der Lee, Qian Feng, Martijn A. Langereis, Rob ter Horst, Radek Szklarczyk, Mihai G. Netea, Arno C. Andeweg, Frank J. M. van Kuppeveld, Martijn A. Huynen

Centre for Molecular and Biomolecular Informatics, Radboud University Medical Center, Nijmegen, The Netherlands

Background: The RIG-I-like receptor (RLR) system is critical for the innate defense against viruses. Recognition of viral RNA by the RLRs leads to the production of type I interferons (IFN α/β) that initiate the immune response.

Description: Here, through systematic assessment of a wide variety of available genomics data, we identify 10 molecular signatures of RLR pathway components. Five of these signatures are based on the relationship of RLR signaling with viruses; the other five are based on properties of the pathway itself. We demonstrate that RLR pathway genes, among others, tend to evolve at a high rate, interact with viral proteins, contain a limited set of protein domains, are regulated by a specific set of transcription factors, and form a tightly connected physical interaction network. By weighing these RLR signatures for their ability to predict known RLR pathway genes, and integrating them, we propose 187 novel genes with a likely role in the human RLR system. Using two RNAi screens, we validate an effect on RIG-I-mediated, IFN β promoter-controlled protein production for about half (94) of these genes. For many of the 19 new genes with the strongest effects, knockdown

also affects RLR signaling outcome at the level of IFN β transcription, leading to significantly decreased mRNA expression. Finally, by connecting the results with the known protein interaction network, we suggest for several newly identified RLR genes where in the pathway they could function.

Conclusions: We present our prioritized list as a resource for identifying genes involved in the RLR system.

11. Understanding leishmania development and drug resistance using an integrative omics compendium

Bart Cuypers, Pieter Meysman, Manu Vanaerschot, Maya Berg, Jean-Claude Dujardin, Kris Laukens

University of Antwerp, Antwerp, Belgium

Leishmania donovani causes visceral leishmaniasis (VL), a disease which is lethal without treatment. With only four drugs available and rapidly emerging drug resistance, knowledge about the parasite's resistance mechanisms is essential to boost the development of new drugs. However, only little is known about *Leishmania*'s gene regulation and the few findings indicate major differences to known gene expression systems. Integration of different 'omics could shed light on these gene regulatory mechanisms, but there has been little integration effort so far. Therefore, we developed an easy to use tool, able to collect and connect all the existing *L. donovani* 'omics experiments. Genomics, epigenomics, transcriptomics, proteomics, metabolomics and phenotypic data was collected and added to a MySQL database compendium, further complemented with publicly available data. Relations between the different 'omics levels were explicitly defined and provided with a level of confidence. Python scripts were developed to preprocess, import and access the data. Next to this vast data source a set of integrative data-analysis tools was developed based on data mining strategies. For exam-

ple: One tool uses frequent pattern mining algorithms to look which proteins and metabolites frequently behave in the same way under different conditions. Another tool converts several 'omics data to a network format that can be opened in Cytoscape and can thus be the basis for network analysis. Using the compendium, we characterized the development and drug-resistance in a system biology context (all 'omics). The compendium and its scripts could be used for other organisms with only minor changes.

12. Unravelling signal regulation from large scale phosphorylation kinetic data

Westa Domanova, James Krycer, Rima Chaudhuri, Fatemeh Vafaei, Daniel Fazakerley, David James, Zdenka Kuncic
University of Sydney

Background: A key biological paradigm is that biological processes are tightly-controlled by the temporal behavior of cellular signalling events. Phosphorylation, a prevalent means of signalling, occurs with rapid dynamics but for the majority of events the kinase is unknown. To elucidate the underlying topology of signalling cascades from high-throughput data, we need to be able to predict kinase substrate relationships (KSRs). Existing prediction algorithms do not consider the crucial biological context of KSRs. Description: Here, we predict KSRs in a data-dependent and automatised fashion: given that some kinases are active before others, we use computationally determined kinase specific temporal patterns to predict site-specific KSRs from large-scale in vivo experiments (ssKSR-LIVE).

Conclusions: Applying this to insulin-stimulated phosphoproteomic data we distinguished between AKT and RPS6KB1, two kinases sharing the same consensus motif. We identified several kinases and predicted novel substrates, correlating this with key insulin-regulated biological processes. As a flexible algorithm ssKSR-LIVE can be applied to other high-throughput signaling data and thus can improve our understanding of

complex diseases caused by dysregulated signalling, including cancer and type 2 diabetes.

Benefits of publishing in BioMed Central's open access journals

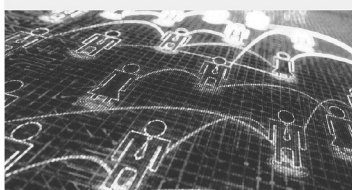


Customer satisfaction

- 89% of authors rated the quality of the overall process as very good
- 90% of authors rated the helpfulness of Editorial staff as very good
- 89% of authors said they would recommend BioMed Central to a colleague

Rapid publication

- Fast and thorough peer review, with an average time to first decision of 6 weeks
- Immediate publication of all articles upon acceptance
- Instant inclusion in all major indexing services including PubMed



High visibility

- Over 25 million page views per month
- Over 7 million user sessions per month
- 1.5 million registered users

"A lot of sincere thanks for your efforts in promoting our paper. The press coverage our article received was unbelievable."
Armen Mulkidjanian

High impact

- Many journals are leaders within their field
- Article-level measures of impact including article accesses and citation tracking
- Over 140 journals with Impact Factors



Innovative publishing solutions

- Innovative publishing tools such as threaded publications and ISRCTN Register facilitate content discovery
- Specialist journals allow integrated publishing solutions of all types of data, e.g. *Trials*, *BMC Research Notes* and *GigaScience*

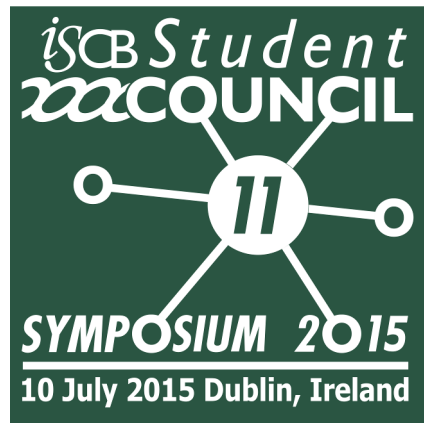
Online publishing

- No space constraints or color figure charges
- Acceptance of multiple file types and formats, including video content
- All journals free to download as apps



Authors retain copyright

- Articles may be published on multiple websites, blogs, publications, etc.
- Redistribute articles to colleagues, opinion leaders and key decision makers without restriction



Poster Presentation Abstracts

bina



Future Shapers.

Meet Bina. **We #code2cure**

A black and white photograph of a woman's face, looking upwards and to the right. The face is framed by a white circular graphic. The background is a dark gray rectangle filled with faint, light gray DNA sequence letters (A, T, C, G) arranged in a pattern that suggests a genome map. At the bottom of the rectangle, the text "NGS informatics for solving genomic data challenges at www.bina.com" is written in white.

NGS informatics for solving genomic data challenges at www.bina.com

For Research Use Only. Not For Use In Diagnostic Procedures.
©2015 Bina Technologies, part of Roche Sequencing. AD-1250-001

Poster Presentation Abstracts

Category : Bioinformatics of Disease and Treatment

1. A Literature Review of Informatics-based approaches to Drug Development for Treating Addiction

Alexander G Christou

Indiana University Bloomington

Background: Chemical dependence or addiction is a chronic disease that affects hundreds of millions of people worldwide. The direct medical and indirect societal costs of addiction are staggering. However, existing medications that specifically target substance dependence are few and research into medication development for substance dependence is limited.

Description: This poster is a comprehensive literature review of previously published papers using computational biology and/or cheminformatics to examine the genetics and chemistry of substance dependence. The scholarly databases considered in this review are ACM, Access Science, EBSCO, Google Scholar, and Science Direct. The review focuses on two main subject areas: the genetics of addiction, and the neurochemistry of substance dependence and substance dependence treatment. The purposes and goals, as well as methodology of the published papers are analyzed. These areas are primarily but not exclusively examined with the goal of medication development in mind.

Conclusion: The amount of published research into addiction science from an informatics perspective is limited and the subject remains largely unexplored. There is great potential in this area for research from both computational biology and cheminformatics approaches.

2. Study to test the inhibitory activity of derivatives of THC- Δ 9-tetrahydrocannabinol on Acetylcholinesterase (AChE) enzyme: A Virtual Screening, Molecular Docking, ADME & Tox Study

Asif Naqvi, Ruhi Jahan, Gajendra Dangi, Mukesh Kumar, Neha Paliwal, Khusboo Jain, Monu Saini, Neetu Sharma, Dr. Kanika Sharma

BioDiscovery Group

Background Tetrahydrocannabinol (THC), also known as Δ 9-THC, Δ 1-THC or dronabinol, is the main psychoactive substance found in the cannabis plant. Previous studies indicate that THC has an anticholinesterase action which may implicate it as a potential treatment for Alzheimer's.

Description The study conducted for new inhibitors for Acetylcholinesterase, also known as AChE-target of Alzheimer's Dementia drugs. Nearly 3500 molecules on structure similarity of Δ 9-THC were virtually screened & molecular docking approach using Lamarckian Genetic Algorithm was carried out to find out the inhibitors for AChE on the basis of calculated ligand-protein pairwise interaction energies. The grid maps representing the protein were calculated using auto grid and grid size was set to 60*60*60 points with grid spacing of 0.375 Å. Docking was carried out with standard docking protocol on the basis of a population size of 150 randomly placed individuals; a maximum number of 2.5×10^7 energy evaluations, a mutation rate of 0.02, a crossover rate of 0.80 and an elitism value of 1. Ten independent docking runs were carried out for each ligand and results were clustered according to the 1.0 Å rmsd criteria.

Conclusions The docking result demon-

strated that the binding energies were in the range of -10.23 kcal/mol to -3.15 kcal/mol, with the minimum binding energy of -10.23 kcal/mol. 8 molecules showing hydrogen bonds with the active site catalytic triad Ser203-Glu334-His447 as well as showed potential ADMET properties. Further in-vitro and in-vivo study is required on these molecules for the future design of new derivatives with higher potency and specificity.

3. Exploration of Gallic acid and their derivatives using in silico and in vitro approaches against amyloid curli of extended beta lactamase (ESBL) producing Escherichia coli

Beema Shafreen

Centre for Nanoscience and Nanotechnology,
Sathyabama University, Chennai, India

Background: E. coli produces curli which is the major proteinaceous component of the complex extra-cellular matrix of the biofilm. The curli produced by E. coli are mainly involved in adhesion to surface, initiates cell aggregation and leads to the formation of biofilm. During infection amyloid curli helps bacterial pathogens to adhere to the host tissues elicits the host inflammatory response and invade deep into the cells and establishes the infection. Biofilm is an escape mechanism adapted by the bacteria to overcome antimicrobial resistance. Hence inhibiting protein essential for curli can significantly inhibit the formation of biofilm by E. coli

Methods: Amyloid inhibitors docked into the binding pocket and were examined for their anti-amyloid and antibiofilm potential using microscopy and biochemical assay.

Results: In absence of crystal structure for CsgA of E.coli, homology model was constructed using modeler from Discovery studio 2.1. The modeled 3D structure was used for docking. Gallic acid and their derivatives were used as the test compounds. Among them, gallic acid (GA) and fluorogallate was revealed with highest docking score (35.377 and 36.231) and binding energy (BE) (373.88 and 242.82 kcal/mol) respectively. Further

the GA with good BE was used for in vitro studies. Gallic acid at 200 µg/ mL showed significant reduction of biofilm. Further, from congo red assay GA showed significant reduction of curli.

Conclusion: Therefore GA known as antioxidant is also capable of inhibiting the biofilm formed by the E. coli. Hence further investigation with GA can help to combat biofilm related infections.

4. Computational genomics approaches for kidney diseases in African patients.

Darlington S Mapiye, Galen Wright, Ikechi Okpechi, Nicki Tiffin

South Africa National Bioinformatics Institute, University of the Western Cape

End-stage renal disease (ESRD) is a complex trait involving multiple processes working together on a background of a significant genetic susceptibility. Black Africans bear an unequal burden of this disease compared to Caucasians. Whole exome sequencing (WES) uses next generation sequencing technology to identify disease associated variation within genes. To date, its application to kidney diseases in African populations has been limited. This study investigates the genetics underlying rare familial ESRD in a South African family and make comparisons to the wider knowledge base of renal diseases. We performed WES on a large South African family of mixed ancestry with an autosomal dominant phenotype of adult-onset nephropathy characterized by early onset abnormal serum creatinine and developmental defects affecting the kidneys. SamTools, Novoalign, Picard, Genome analysis tool kit (GATK), VAAST and Ingenuity variant analysis (IVA) were applied for bioinformatics analysis. 1 unaffected and 5 affected family member were sequenced. We identified 3 novel indels and 16 missense variants in 10 genes (FBXL21, SYCE1L, KCNN3, COL4A1, ICAM1, COL16A1, ZMYM1, STXBP3, ANXA9, and CEBPZ). These were identified in all affected family members and were consistent with autosomal dominant inheritance.

Of these only 3 rare missense variants in 3 genes (COL4A1 [p.R476W], ICAM1 [p.P352L], COL16A1 [p.T116M]) were considered potentially disease causing based on functional analysis. None of these variants were detected in the unaffected family member. This study shows a successful application of WES for the identification of pathogenic mutations, illustrating the power of molecular genetic diagnostics techniques that may explain complex renal phenotypes.

5. Preliminary data mining and statistics on systemic amyloidosis in Rhesus macaques

Eric Leung, Michael Raboin, Anne Lewis, Kamm Prongay, Aaron Cohen, Amanda Vinson

Oregon Health & Science University

Background: Systemic amyloidosis is a disease that is characterized by misfolded proteins becoming insoluble. This insolubility allows these proteins to deposit in body organs and disrupts normal organ function, causing significant morbidity and mortality. We have found a high prevalence of SA of unknown etiology in a colony of rhesus macaques (*Macaca mulatta*) at the Oregon National Primate Research Center in Beaverton, Oregon, USA.

Description: The goal of this study is to use the extensive animal records database maintained on these macaques to describe the etiology of systemic amyloidosis over time within this population. Using the Python programming language, we extracted features and summarized descriptive statistics on the 287 positive cases of systemic amyloidosis in rhesus macaque deceased between January 1st, 2010 and January 1st, 2015.

Conclusions: This preliminary work lays the foundation for feature selection in macaques. Ultimately, these analyses will allow us to develop classification and prediction algorithms, which can be used to identify animals at risk for systemic amyloidosis, and to prevent the development of the disease. This initial retrospective analysis is the first to characterize naturally occurring systemic amyloi-

dosis in a population of rhesus macaques that are important models of human disease.

6. Rationally designed drug blending as a mechanism to overcome drug resistance in cancer

Francisco Martínez-Jiménez, John P. Overington, Bissan Al-Lazikani and Marc A. Marti-Renom

Centro Nacional Análisis Genómico

Drug resistance is one of the major problems in cancer treatment. Rapid mutation and selective pressure can efficiently select drug-resistant mutants, although there are many mechanisms for drug resistance, a classic mechanism is due to coding mutations in the drug-target binding site. Some well-characterized examples are the resistance to BRAF inhibitors in the BRAF (V600E) positive melanomas, resistance to gefitinib in non-small-cell lung cancers due to point mutations in epidermal growth factor receptor, or resistance to topotecan caused by topoisomerase I mutations. To systematically analyse the mutational landscape that can potentially cause resistance to targeted therapies, we have studied the current targets of small-molecule treatments for 30 different classes of cancer. Using the mutational frequencies from Alexandrov et al. (1) we have generated the 3D models for each of most likely mutants for the current drug targets. Next, for all the 3D-generated mutants, we have predicted the resistance potential to the current drug treatments. Finally, for each of the mutants evaluated as likely to confer resistance, we have computationally defined a set of molecules targeting the same protein, which would theoretically overcome drug resistance. This study aims to reduce the difficulties in the choice of the optimal treatment and subsequently, it is a step further in the development of the personalized medicine for the treatment of cancer. 1. Alexandrov, L.B, A.L. et al. (2013) Signatures of mutational processes in human cancer. *Nature*, 500, 415-421

7. Prioritizing a drug's targets using both gene expression and structural similarity

Griet Laenen, Sander Willems, Daniela Börnigen, Amin Ardeshtirdavani, Lieven Thorrez, Yves Moreau
KU Leuven

The pharmaceutical industry is facing unprecedented pressure to increase its productivity. Attrition rates in the later stages of drug development have risen sharply, with toxicity and lack of efficacy being the main bottlenecks. To address both these safety- and efficacy-related issues, a better understanding of the complex biological response to drug treatment is vital. Although many drugs exert their effects through the modulation of multiple targets, these targets are often unknown and identification among the thousands of gene products remains difficult. We propose a computational method to support this identification of putative targets of a drug. The first component of our method prioritizes proteins as potential targets by integrating experimental gene expression data with prior knowledge on protein interactions. More specifically, genes are ranked based on the transcriptional response of functionally related genes by diffusing differential expression signals following treatment over a protein interaction network. The second component of our method prioritizes proteins as drug targets based on the interaction with compounds structurally similar to the drug of interest. To this end compound-compound similarity scores are combined with compound-protein interaction scores. Our method has been evaluated on a test set of small molecule drugs for which the known targets were derived from ChEMBL. AUC values of up to 90% were obtained. These AUC values indicate the predictive power of combining gene expression data and structural information for a drug of interest with known protein-protein and protein-compound interaction information respectively, to identify the targets of that drug.

8. Reconstruction of the temporal signaling network in Salmonella-infected human cells

Gungor Budak, Oyku Eren Ozsoy, Yesim Aydin Son, Tolga Can, Nurcan Tuncbag
Middle East Technical University

Salmonella enterica is a bacterial pathogen that usually infects its host through food sources. Translocation of the pathogen proteins into the host cells changes signaling mechanism either by activating or inhibiting the host proteins. Using high-throughput ‘omic’ technologies, these changes can be quantified at different levels; however, experimental hits are usually incomplete to represent the whole signaling system as some driver proteins stay hidden in the experimental data. Given that the bacterial infection modifies the host response network, more coherent view of the underlying biological processes and the signaling networks can be obtained by using a network modeling approach with reverse engineering principles in which a confident region from the interactome is found by inferring hits from omic experiments. In this work, we used a published temporal phosphoproteomic data of *Salmonella*-infected human cells and reconstructed the temporal signaling network of the host by integrating interactome and phosphoproteomic data. We combined two well-established network modeling frameworks, the Prize-collecting Steiner Forest and the Integer Linear Programming based edge inference approach. The resulting network conserves the temporality and directionality, while revealing hidden entities, such as SNARE binding, mTOR signaling, immune response, cytoskeleton organization, and apoptosis pathways. *Salmonella* effectors' targets in the host cells such as CDC42, RHOA, 14-3-3 δ , Syntaxin family, Oxysterol-binding proteins were included in the reconstructed network although they were not in the phosphoproteomic data. We believe integrated approaches, such as the one here, have a high potential for the identification of clinical targets in infectious diseases, especially in the *Salmonella* infections.

9. Regulatory network analysis identifies miR-146b-5p as a key regulator of BCR-ABL-positive microvesicles transforming mononuclear cells into leukemia-like cells

Hongmei zhang, Hong-mei Zhang, Qing Li, Xiaojian Zhu

Huazhong University of Science and Technology

Background: Malignant transformation of normal hematopoietic transplants induced by residual leukemia cells is considered as a pivotal mechanism of donor cell leukemia. Microvesicles (MVs) play critical roles in this transformation. Incubating with BCR-ABL1-positive MVs, mononuclear cells separated from hematopoietic transplants exhibited a leukemia-like phenotype, which provided a good model for studying the malignant transformation of leukemia. We constructed transcription factors (TFs) and miRNAs co-regulatory networks to study the molecular mechanism of this transformation process. **Results:** RNA sequencing and small RNA sequencing were performed for 5 samples of the key transformation time points. We found that at the start of transformation, many anti-tumor miRNAs were up-regulated and the immune response pathway were initiated, while in the later anti-tumor miRNAs were decreased and oncomiRs were increased, as well as the cell cycle, DNA replication and energy metabolism pathways were activated. TF-miRNA co-regulatory network analysis revealed many important TFs and miRNAs, such as YBX1, MYC, STAT5A, miR-17~92 and miR-146b-5p, are responsible for the differential expression of cancer related genes. MiR-146b-5p as the highest expressed miRNA in the transformation process was predicted to target Notch signaling pathway. Our experiments verified that the up-regulation of miR-146b-5p in K562 could accelerate the transformation process through targeting NUMB and NOTCH2, accompanied by increasing of ROS and genome instability. **Conclusion:** TFs-miRNAs co-regulatory network analysis elucidated the dys-regulatory mechanisms of transformation

process. Experiments verified the up-regulation of miR-146b-5p promote the transformation of mononuclear cells. This study provided important clues for understanding the malignant transformation of leukamia.

10. Data Mining and Pattern Recognition Models for Identifying Inherited Diseases: The Challenges Ahead

Lahiru Manohara Iddamalgoda, Achala Chathuranga Aponso, Naomi Krishnarajah and Prashanth Suravajhala

Department of Computing, Informatics Institute of Technology, Colombo, Sri Lanka

Background: Data mining and pattern recognizing have been in use for genetic studies, specifically applicable for identifying inherited diseases. Several researchers have proposed data mining models and biomedical approaches in finding pathogenesis of inherited disease. The main challenge in prediction of genetic inherited disease is to correctly classify the genetic factor.

Description: In this work, we review state-of-art data mining and pattern recognizing models for identifying inherited diseases and describe the inherent challenges we pose in identifying diseases states. We describe binary classification methods and scoring based prioritization in order to provide the scientific background of the problem. Of different approaches, we consider SNP based binary classification and scoring based genes prioritization as better models for diseasome studies.

Conclusions: While several methods, viz. gene prioritization, protein-protein interactions are preferentially used, we conclude that a combination of the general framework models including SNPs are valuable in easy detection and identification of inherited diseases.

11. A comprehensive survey of ncRNAs in the genome and transcriptome of the tropical parasite Trypanosoma cruzi

Maina Bitar, Martin Smith, John Mattick, Gloria Franco

Universidade Federal de Minas Gerais / Garvan Institute

Trypanosoma cruzi is the etiologic agent of Chagas disease, a neglected tropical disease that mainly affects South America and leads to substantial socio-economic losses. This protozoan parasite was first reported in 1909 and in the following decades a variety of studies have elucidated several unique aspects of its biology. In recent years the genomes of six *T. cruzi* strains have been sequenced, but their annotation remains very poor, specially regarding non-coding RNAs (ncRNAs). ncRNAs are known to have crucial roles in virtually all biological processes and the constantly growing list of functions related to such molecules has influenced even the classical definition of a gene. To better elucidate the possible roles these molecules may play in a parasite with such a complex life-cycle, we have scanned a *T. cruzi* genome using similarity search methods to annotate ncRNA genes based on public databases. As result, over 1500 candidates were identified, more than 40% representing new findings, many of which represent ncRNAs not previously explored in this parasite and thus worthy of further studies. Publicly available RNASeq data have confirmed the expression of 300 of these candidates. We have then sequenced the set of *T. cruzi* RNAs both before and after gamma radiation exposure and we are currently analyzing the differential expression pattern of these previously identified ncRNA candidates and also of newly identified genes. This work will shed light on the function of ncRNAs in the parasite biology and possibly in its resistance to ionizing radiation.

12. Investigating evolutionary models of genome structure in aggressive prostate cancer

Marek Cmero, Natalie Kurganovs, Jessica Chung, Jan Schroeder, Kangbo Mo, Clare Sloggett, Niall M. Corcoran, Christopher M Hovens, Cheng Soon Ong, Geoff Macintyre
The University of Melbourne

Tumour evolution is a complex and multifaceted process. Recently, many approaches have arisen for inferring the evolutionary dynamics of tumour cell populations from point-mutation and copy-number data. Studying the role of structural variations (SVs) in cancer evolution however, particularly balanced rearrangements, has been less thoroughly explored. We present a method of reconstructing cancer phylogeny from multiple single-patient samples using large scale genomic aberrations and apply it to prostate cancer, which is particularly rearrangement-driven. We demonstrate that tumour phylogenies are able to be reconstructed using rearrangement data alone, and we further expand our model to characterise subclonal SVs. We demonstrate our methods by applying them to longitudinal samples from patients undergoing second-line anti-hormone therapy to gain insight into the mechanisms of castration resistance.

13. Computational Analysis of Anopheles gambiae Metabolism to Facilitate Insecticidal Target Discovery

Marion Olubunmi Adebisi
Covenant University

Background: Insecticide resistance is an inherited characteristic involving changes in one or more insect genes. It is also a major public health challenge combating world efforts on malaria control. The malaria vector, *Anopheles gambiae* (*A. gambiae*) has developed resistance to existing classes of insecticides, particularly pyrethroids (the only class approved for Indoor Residual Spray [IRS] and Long-Lasting Insecticide Treated Net

[LLITNs]). Identification of novel insecticidal targets for the development of more effective insecticides is critical, however, determining which gene products are ideal insecticidal targets remains a difficult task.

Description: The dissection and comprehensive study of biochemical metabolic networks has great potential to effectively and specifically identify essential enzymes as potential insecticidal targets. Using the PathoLogic program, we have constructed AnoCyc, a pathway/genome database (PGDB) for *A. gambiae* AgamP3, using its annotated genomic sequence and other annotated information from UNIPROT and KEGG databases. AgamP3, the first PGDB for *A. gambiae* has been deployed to BioCyc Database collection. Next we applied a graph-based model that analyzed the topology of the metabolic network of *A. gambiae* to determine essential enzymatic reactions in the networks.

Conclusion: We obtained a refined list of 61 potential insecticidal candidate targets, which include one clinically validated insecticidal target and host of others with biological evidence in the literature. Seven targets appear novel and have no significant homology with vertebrates.

14. Online Detection of Malaria Parasite in Digital Microscopic Image

Muteba Ayumba Eustache

Independent Researcher

Background: Malaria is caused by the protozoan parasite *Plasmodium* and is transmitted by infected mosquitos. The most species of *Plasmodium* are the following: *P. falciparum*, *P. vivax*, *P. malariae* and *P. knowlesi*. Clinical diagnosis of malaria is based on the patient's history, symptoms and on physical findings at examination. However, for a definitive diagnosis to be made, it is mandatory to obtain the laboratory confirmation of the presence of malaria parasites in the patient's organism. Malaria parasites can be identified by examining under the microscope a drop of the patient's blood, spread out as a "blood smear" on a microscope slide. The morphological features and the estimation of parasi-

taemia represent the essential basis of a correct laboratory diagnosis confirmation of malaria infection and species of *Plasmodium*. The correct and timely diagnosis of malaria infection in a ill patient is in fact of critical importance. Digital image processing presents a wide area of applications including diagnosis.

Description: The proposed solution has two main goals: the estimation of parasitaemia and the morphological features detection. To achieve these goals, we make use of the content based image recognition technique. Our approach demonstrates that an image is a set of subsquares (Classes), each subsquare contains a sequence of pixel and each pixel contains value. If an image is to be process, in the first step, the user must select an element of the image that appears as possessing the information related to the parasite. From the region containing this element, our system can proceed to the analysis and to the detection of species of *Plasmodium* through comparison to images' pattern of species.

Conclusions: In order to contribute to fighting malaria from technological front, we developed an approach for malaria diagnosis based on automated detection of malaria parasite in digital microscope image. Utilizing our system in developing countries where there are disparities between urban and rural areas can benefit to healthcare practitioners to have a remote access expertise and support.

15. Bioinformatics challenges, assessment and implications for identification of fusion genes in an RNA sequencing assay for clinically relevant theranostic targets in solid tumors and leukemias

Numrah Fadra, Jaime Davila, Asha Nair, Eric Klee, Amber McDonald, Xiaoke Wang, Jesse Voss, Benjamin Kipp, Bard Crusan, Joseph Blommel, Lisa Peterson, Xianglin Wu, Veldhuizen Tamra, Jin Jen, Robert Jenkins.

Mayo Clinic

Background: Identification of gene fusions using next generation RNA sequencing has

become widely available owing to the use of complex chemistry leveraging information from mRNA isolation and cDNA synthesis followed by alignment and fusion detection using robust bioinformatics approaches. While RNA sequencing is increasingly being used in research for gene expression quantification and novel spliced variant discovery, the implementation of a clinical assay continues to remain a bioinformatics challenge due to the relatively unstable nature of RNA as well as lack of comparable standards for performance assessment.

Description: We evaluated the bioinformatics performance characteristics for identification of gene fusions using variables such as gene length, tumor percentage, RIN (RNA integrity number) and distance of fusion breakpoint from the poly(A) in mRNA. The impact of RIN on number of sequencing reads and accuracy of fusion detection was calculated using in house degradation of cell lines.

Conclusions: Using well characterized samples with known gene fusions, we have generated robust quality control metrics by establishing accurate limit of detection using Illumina TrueSeq, 3' poly(A) pull down protocol. We aim to generate a profile for the distances to poly(A) tail for clinical gene targets to establish a cut-off distance given the chemistry for library synthesis. These approaches would provide clinicians the ability to make quality control guided decisions for tumor specimens without being confounded by multiple variations and bias given the complex nature of RNA thereby successfully implementing a test for identification of theranostic markers in malignant or benign cancers.

16. Estimation of Fetal Fraction in Maternal Plasma through Read Distribution Relative to Nucleosome Positions

Roy Straver, Marcel Reinders, Erik Sistermans

VU University Medical Center Amsterdam

Background: Recent developments allowed to replace invasive testing procedures with non invasive alternatives for prenatal diag-

nostics. Although this does reduce the danger of miscarriage, its reliability depends on the fraction of fetal DNA in the maternal plasma. No methods exist to determine the fetal fraction without changing wet lab analysis pipelines. Such changes include additional labwork and increasing the amount of sequenced DNA. Both options increase the cost per sample, rendering them too expensive for routine diagnostic purposes.

Description: We present a method which estimates the fetal fraction using only low coverage single end reads from maternal samples, independent of their length. The estimation is achieved through the positional distribution of reads mapped relative to nucleosome positions we determined on this type of data. We obtained a correlation of 0.654 with a P-value of 1.86e-08 over 59 test samples between our results and reference values based on the ratio of reads mapped to chromosome Y. We also obtained a correlation of 0.891 (P-value 6.00e-08, 21 samples) with an existing, more expensive method based on paired-end sequencing. Additionally, our method shows the correlations between DNA fragment size and the amount of reads positioned relative to nucleosome positions.

Conclusions: Our results show it is possible to estimate the fetal fraction in any sample without the need of changing the labwork, providing valuable information for NIPT diagnostics without additional costs, while providing additional insight in the biological causes for known differences in fragment sizes between maternal and fetal DNA.

17. Compiling a minicircle genome for Trypanosoma brucei

Tyler Faits, Tian Yu, Gary Benson, Stefano Monti, Ruslan Afasizhev
Boston University

Background: Trypanosomes, protozoan parasites infecting millions worldwide, are characterized by a prominent kinetoplast, a single large mitochondrion with unique DNA structure. Kinetoplast DNA comprises approximately 10,000 interlinked molecules of circu-

lar DNA known as minicircles (1kb) and maxicircles (25kb). Due to the existence of conserved sequence boxes (CSBs) and DNA bend regions, typical short-read sequencing technology is ill-suited for compiling a minicircle genome. Pacific Biosciences' single molecule real-time sequencing (PacBio-SMRT), while suffering from higher error-rate, offers near-complete minicircle reads and dramatically reduces the difficulty of sequence assembly.

Description: We used PacBio-SMRT to sequence minicircles, and Illumina MiSeq to sequence the gRNA transcriptome of steady-state *Trypanosoma brucei* cells. We filtered raw PacBio reads and clustered them based on sequence-similarity. We generated consensus sequences based on multiple-alignments for each cluster, and validated those sequences by examining gRNA coverage and presence of CSBs. We estimated copy-number for each consensus sequence, and correlated the abundance of each unique minicircle with that of its gRNA transcripts. We used gRNA alignment rates to estimate the completeness of our compiled minicircle genome. We created a searchable online database to allow easy access to our results.

Conclusions: Our analysis yielded 143 minicircles, including 128 unpublished sequences. All identified minicircles contained CSBs and DNA bend regions, which cannot be found in most raw reads. Our result showed that the relatively frequent errors of PacBio-SMRT reads can be corrected through clustering and multiple alignment. The abundant gRNA coverage provided evidence that minicircles are likely to encode multiple gRNAs on both strands.

18. Prognostic long non-coding RNAs involved in prostate cancer progression and treatment resistance

Varune Rohan Ramnarine, Alexander W Wyatt, Fan Mo, Kendric Wang, Mohammed Alshalalfa, Elai Davicioni, Yuzhuo Wang, Colin C Collins

Univeristy of British Columbia

Neuroendocrine prostate cancer (NEPC) is a treatment-resistant lethal disease where most patients die within 1 year and treatment options are strictly palliative. New therapeutic strategies and a greater understanding of how resistance emerges are urgently required to improve patient outcome. To comprehensively characterize the disease we have pioneered a high fidelity xenograft model of NEPC. Adenocarcinoma(AD) patient tumors are grafted into mice, exposed to clinically used treatment, and then develop aggressive NEPC. As the tumor progresses, using deep next-generation sequencing, we profile at various time points generating a time series of events that result in terminal NEPC. To help elucidate novel mechanisms of resistance we apply robust sequence analysis algorithms focusing on underexplored regions of the genome that don't code for protein (98.78%). Within these non-coding RNAs (ncRNAs) are a subclass of long ncRNAs (lncRNAs) – most lacking known function and/or unannotated. A core feature of our sequencing pipeline is the quasi de novo transcriptome assembly, which identifies unannotated transcripts, novel spliced-isoforms, as well as known transcripts. We detect ~90K and ~40K ncRNAs including 13,730 and 20,463 lncRNAs, annotated and novel respectively. Our results show unique patterns of lncRNA expression along our time series that we have classified into five transcript categories, dissecting the pathogenesis from AD to NEPC. Overlaying these patterns with patient data have identified clinically relevant candidates and prognostic associations to patient outcome. Furthermore our research is the first to provide evidence of lncRNAs involved in NEPC and most importantly insights into

novel mechanisms driving the disease.

Category : Computational Aspects

19. An extensive and interactive database of super-enhancers – dbSUPER

Aziz Khan, Xuegong Zhang

MOE Key Laboratory of Bioinformatics, Bioinformatics Division and Center for Synthetic and Systems Biology, TNLIST/ Department of Automation, Tsinghua University, China

Background: Super-enhancers are large clusters of transcriptional enhancers that can drive cell-type-specific gene expression and are crucial in cell identity. Many disease-associated sequence variations are enriched in these super-enhancer regions of disease-relevant cell types. Thus, super-enhancers can be used as potential biomarkers for disease diagnosis and therapeutics. Current studies have identified super-enhancers for more than 100 cell types in human and mouse. However, no centralized resource to integrate all these findings is available yet.

Description: We developed dbSUPER, the first extensive and interactive database of super-enhancers in mouse and human genome. We provide an interactive data transfer platform to easily send the data to Galaxy, GREAT and Cistrome web servers for further downstream analysis. dbSUPER provides a responsive and user-friendly web interface to facilitate efficient and comprehensive searching and browsing. dbSUPER provides downloadable and exportable features in a variety of data formats, and can be visualized in UCSC genome browser while custom tracks will be added automatically. Further, dbSUPER lists genes associated with super-enhancers and links to various databases. Our database also provides an overlap analysis tool, to check the overlap of user defined regions with the current database. Currently, our database contains 66,033 super-enhancers for 96 human and 5 mouse tissue/cell types.

Conclusions: The primary goal is providing a resource for further study of transcriptional

control of cell identity and disease by archiving computationally produced data. We believe, is a valuable resource for the bioinformatics and genetics research community. dbSUPER is freely available at:

<http://bioinfo.au.tsinghua.edu.cn/dbsuper/>

20. phpDAVIDws: a class-based PHP platform for DAVID web services

Aziz Khan, Xuegong Zhang

MOE Key Laboratory of Bioinformatics, Bioinformatics Division and Center for Synthetic and Systems Biology, TNLIST/ Department of Automation, Tsinghua University, China

Background: Functional annotation is one of the fundamental step in downstream analysis in genomics and proteomics to give biological meaning to genes and proteins. DAVID (database for annotation, visualization and integrated discovery), is a web-based bioinformatics platform, that provides tools for functional interpretation of large lists of genes/proteins. DAVID team recently developed a web service interface (DAVID-WS) by giving a full control over all its functionalities.

Description: DAVID-WS clients have been already made available for Python, Perl, Matlab, Java and R. PHP is a server-side scripting language and one of the most used web development programming language to develop online bioinformatics tools and a databases. But, currently there is no PHP based client is available to access the DAVID web services. Hence we developed a class-based PHP interface to access all the DAVID-WS functionalities in a PHP environment.

Conclusions: phpDAVIDws is an object oriented, flexible and fast interface to access all the functionalities available by DAVIDWeb-Service. This will help bioinformaticians to easily perform real time functional analysis for their data on their web servers. A beta version of source code and documentation is freely available at:

<http://asntech.github.io/phpDAVIDws/>

21. A parallel implementation of genomic prediction including genetic by environment interaction

Arne De Coninck, Jan Fostier, Steven Maenhout, Drosos Kourounis, Fabio Verbosio, Olaf Schenk, Bernard De Baets
Ghent University

Background: Genomic prediction for plant breeding requires taking into account environmental effects and variations of genetic effects across environments. The latter can be modelled by estimating the effect of each genetic marker in every possible environment, which leads to a huge amount of effects to be estimated. Nonetheless, the information about these effects is only sparsely present, due to the fact that plants are only tested in a limited number of environments.

Description: A parallel implementation for predicting the genetic marker effects and their variations in different environments was conceived, combining sparse and dense matrix formalisms to enable the analysis of large-scale datasets (up to 100,000 observations in 100 environments). A study with simulated data sets showed that the accuracy of predicting the variations of the genetic marker effects in the different environments depends mainly on the number of plants tested in each environment.

Conclusions: To achieve a better prediction accuracy of the agronomic potential of future hybrid crops it is thus essential to possess of training data with a lot of observations in the environments of interest. We implemented a prototype that is optimized for dealing with large numbers of observations by efficiently applying the power of a supercomputing cluster. This clears the path for large-scale data analysis and prediction, which may lead to hybrid crops that are highly optimized for specific environments.

22. Identification of binding sites and favorable ligand binding moieties by virtual screening and Self-Organizing Map analysis

Emna Harigua-Souiai, Isidro Cortes-Ciriano, Nathan Desdouits, Therese E. Malliavin, Ikram Guizani, Michael Nilges, Arnaud Blondel, Guillaume Bouvier
Institut Pasteur of Tunis

Background: Identifying druggable cavities on a protein surface is a crucial step in structure based drug design. The cavities have to present suitable size and shape, as well as appropriate chemical complementarity with ligands. 2. Description: We present a novel cavity prediction method that analyzes results of virtual screening of specific ligands or fragment libraries by means of Self-Organizing Maps. We demonstrate the method with two thoroughly studied proteins where it successfully identified their active sites (AS) and relevant secondary binding sites (BS). Moreover, known active ligands mapped the AS better than inactive ones. Interestingly, docking a naive fragment library brought even more insight. We then systematically applied the method to the 102 targets from the DUD-E database, where it showed a 90% identification rate of the AS among the first three consensual clusters of the SOM, and in 82% of the cases as the first one. Further analysis by chemical decomposition of the fragments improved BS prediction. Chemical substructures that are representative of the active ligands preferentially mapped in the AS. 2. Conclusions: The new approach provides valuable information both on relevant BSs and on chemical features promoting bioactivity.

23. Visualization and statistical analysis of mass-cytometry data using SPADEVizR and CytoAnnot.

Nicolas Tchitchek, David Pejoski, Ludovic Platon, Brice Targat, Roger Le Grand, Anne-Sophie Beignon
CEA, Division of Immuno-Virology

Background: Flow and mass cytometry are experimental techniques used for the characterization of cell properties at a single cell resolution. While flow cytometry is currently able to measure up to 18 cell markers, the recently introduced mass-cytometry technology is able to measure up to 40 markers. The Spanning Tree Progression of Density Normalized Events (SPADE) clustering algorithm has been proposed to analyze mass-cytometry data and to identify clusters of cells sharing similar intensities of selected expression markers. While SPADE offers new unique data mining opportunities, it lacks additional post-clustering methods to better characterize the identified cell populations.

Description: We present here, SPADEVizR and CytoAnnot, two R packages designed for the analysis of mass-cytometry data and SPADE results. SPADEVizR is devoted to better interpret SPADE results using new visualization and statistical approaches. We demonstrate that the proposed methods, such as parallel coordinates, multidimensional scaling or streamgraph representations offer efficient ways to represent features, similarities and kinetics of identified cell clusters. Moreover, the proposed statistical methods allow the identification of cell populations with pertinent biological information and further integration with other biological variables. Additionally, CytoAnnot was developed to annotate cell or cell clusters in cytometry profiles and SPADE results with a new statistical hypothesis test. Using this package, cells or populations can be compared and labeled using previously defined cell sub-set references, regarding their marker expression similarities.

Conclusions: SPADEVizR and CytoAnnot

are extremely valuable for bioinformaticians and biologists aiming to explore mass-cytometry data and SPADE results using novel visualization and statistical approaches.

Category : Comparative Genomics

24. Enabling genotype associations of large insertions and deletions by canonicalizing variants across collections of bacterial strains

Alex Salazar, Christopher Desjardins, Ashlee Earl, Thomas Abeel
Delft University of Technology

The comparison of chromosomal sequence variants across hundreds of bacterial genomes from whole genome sequencing data allow in-depth investigations of the evolution of genetic characteristics in bacterial populations. Unfortunately, the comparison of structural variants, such as large insertions and deletions (indels), across bacterial genomes are difficult to perform because of the difficulties in mapping and assembling large variants with short-read sequencing technology. To address this limitation, new algorithms that can identify large indels using short-reads have been developed. However, for these tools to be effective, they must be robust and accurate when identifying large indels that are identically present across multiple genomes. In this study, we investigated the robustness and accuracy of several variant callers for predicting the same large indel across multiple genomes. By simulating large indels of varying sizes across several hundred genomes of *Mycobacterium tuberculosis*, we benchmarked two recently developed variant callers: Pilon and MindTheGap. We found that both tools can be highly inconsistent and inaccurate when identifying the same large indel across multiple genomes. We extended our investigation to analyze how variables such as indel size, indel type, read-depth, read-type, and nearby variation around the indel site can affect a variant caller's ability to consistently and accurately report a large variant. Our findings allow us to better incor-

porate and analyze information from large structural variants when comparing genomes, enabling us to associate large indels with phenotypes of interest. Furthermore, understanding the error-mode of existing algorithms will assist in the development of algorithms that can more consistently and accurately identify large variants across multiple genomes.

25. Nonribosomal peptide synthetase A-domain – substrate interaction investigation

Candice Ryan, Kevin Lobb, Özlem Tastan Bishop
Rhodes University

Nonribosomal peptides (NRPs), synthesized by bacterial nonribosomal peptide synthetases (NRPSs), have important properties useful in reaction to plant phytopathogens. Resistance has led to NRPS substrate specificity being investigated to generate novel natural compounds. A-domains of NRPS are responsible for the amino acid activation and substrate specificity and therefore were the focus of the study. Initial investigation was conducted on a few representative NRPSs for which there is a crystal structure. This was then expanded to include a larger subset of sequences known to contain NRPS modules. Substrate specificity of sequences was determined followed by a phylogenetic study to determine the evolutionary relation of the sequences. A docking investigation was conducted on 5 known crystal structures with 39 ligands, consisting of the 20 standard amino acids in both L and D configuration, where available, and a few non-standard amino acids. The ability of the synthetases to take up these substrates was analyzed in silico and residues that were found to interact regularly with differing substrates were considered to be critical in binding. Future work will include in silico mutation of identified interacting residues, homology modeling of the altered structure followed by re-docking to give insight into the activity of these synthetases. Molecular dynamics will be used to obtain accurate binding energies and to probe the longer-term stability of these ligands within

the binding sites. These residues could be ultimately modified in vitro to alter substrate uptake by NRPSs and aid in development of novel natural products for use as plant phytopathogens.

26. Using bioinformatics to elucidate the structure, function and evolution of an enigmatic class of multifunctional eukaryotic proteins - the caleosins

Farzana Rahman, Negusse Kitaba, Denis Murphy
University of South Wales

The caleosin group is a major family of proteins that occur in two major eukaryotic clades, namely Viridiplantae and Fungi. This pattern of occurrence is not consistent with the evolution of caleosin genes from a common ancestor because the Fungi, along with animals and many protists, are members of the Opisthokonta, while the Viridiplantae are derived from green algal predecessors. This suggestion that caleosin genes may have been acquired in one of the current clades via horizontal gene transfer from the other. We have studied the variation in caleosin-related gene and protein sequences across several hundred species and have developed de novo structural information for the protein using a combination of computational prediction and experimental work. Protein structure prediction suggests that the calcium-binding domain is widely conserved across species, while there is large variation in the loop region of the structure. While the biological functions of caleosins have yet to be determined in detail it is clear that these proteins have several subcellular locations and participate in a range of physiological processes in both plants and fungi. Our biochemical and modeling analyses demonstrate that two forms of caleosins are present in plants, a lipid-droplet form and a bilayer membrane form. Both caleosin forms have canonical calcium binding and phosphorylation domains and appear to have peroxygenase activity. In this presentation we describe further modelling and bioinformatics studies that are

beginning to shed light on the origin and functions of this intriguing group of proteins.

**27. What sequence information can reveal:
*The functional evolution of arrestins in deuterostomes***

Henrike Indrischek, Sonja Prohaska
University Leipzig

The cytosolic arrestin proteins mediate desensitization of activated G-protein coupled receptors via competitive binding of the receptor or via internalization mediated by clathrin binding motifs. As different arrestin conformations can result in specific signaling outcomes, this protein family is a possible target in drug therapeutics. The aim of the current study was to improve the existing incomplete and error-prone annotations of arrestin genes in order to reveal details about the functional evolution of these proteins. Identity and number of arrestin paralogs were determined searching vertebrate genomes or, alternatively, gene expression data with individual Hidden Markov models for each exon and paralog. Unlike standard gene prediction methods, our pipeline can detect exons situated on different scaffolds and assign them to the same gene, increasing completeness of the annotation. We uncovered the interesting duplication- and deletion history of arrestin paralogs in deuterostomes including tandem duplications, pseudogenization and retrogene formation. At the root of vertebrates, two whole genome duplications have given rise to four arrestin paralogs from a single arrestin as it is found in ciona today. An additional clathrin binding motif was gained after the duplications. The precursor of visual arrestins lost the first clathrin binding motif establishing a functional difference in arresting G-protein coupled signaling in non-visual and visual arrestins. The current work shows how an improved annotation of a multi-exon gene family can result in a detailed understanding of the link between gene architecture and functional evolution.

28. A Comparative Analysis of the Genetic Relationships Between the Pathogens of Ebola Hemorrhagic Fever, Marburg Virus, HIV, Hepatitis A, Hepatitis B, Hepatitis C, Hepatitis D and Hepatitis E

Olaitan Awe, Segun Fatumo, Olugbenga Oluwagbemi, Angela Makolo
University of Ibadan

Ebola is a public health problem and a global monster currently ravaging many nations of the world especially West Africa. Ebola viruses are highly pathogenic, exotic agents that can cause severe hemorrhagic fever disease in human and/or nonhuman primates. Ebola is a member of the negative-stranded RNA virus family Filoviridae. Ebola is a very deadly virus. A comparison of gene content and genome architecture of Ebola Hemorrhagic Fever, Marburg virus, HIV, Hepatitis A, Hepatitis B, Hepatitis C, Hepatitis D, Hepatitis E; major, eight related pathogens with different life cycles and disease pathology, revealed a conserved core protein sequence of genes in large syntenic polycistronic gene clusters. In this paper, we highlight the genetics of the Ebola genome with the genome of seven other viruses to identify points of significant similarities and disparities. The basic structure of Ebola is long and filamentous, essentially bacilliform, but the virus often takes on a U shape, and the particles can be up to 14,000 nm in length and average 80 nm in diameter. Genomics provides an unprecedented opportunity to probe in minute detail into the genomes of West Africa's most recent deadly disease - Ebola Hemorrhagic Fever. Here we report comparative genomics of Ebola strain, Zaire ebolavirus isolate Ebola virus/H.sapiens-wt/SLE/2014/Makona-EM095, complete genome. Knowledge gained from this comparative analysis can help provide innovative methods in solving the Ebola menace. An integrative knowledge of genetics and skills in bioinformatics can form a formidable tool in fighting the Ebola menace.

29. Identification Of Candidate Genes Associated With Disease Resistance Related Genes In Oil Palm Via Comparative Genomics

Rozana Rosli, Rozana Rosli, Mohd Amin Ab Halim, Chan Kuang Lim, Leslie Low Eng Ti, Rajinder Singh, Ravigadevi Sambanthamurthi, Denis J Murphy
University of South Wales

The availability of genome sequences have considerably facilitated the identification of key genes involved in the regulation of important agronomic traits. The completion of the oil palm genome sequence in 2013 has provided an important dataset that is enabling researchers to analyse and mine this genome for functions such as R genes that are involved in aspects of disease resistance. Efforts towards identification of R genes may help in improving disease resistance screening for the most major disease in oil palm - Ganoderma. While several strategies have been developed in order to characterize the R genes in oil palm, this study uses a comparative genomics analysis for their elucidation. In this report, we compare *E. guineensis* gene model sequences with four different plant gene models: *Arabidopsis thaliana*, *Phoenix dactylifera*, *Musa acuminata* and *Oryza sativa*. Orthologous genes between two genomes identified with InParanoid and MultiParanoid were then used to analyze protein relationship between the different species. The findings can provide information about the species evolution as well as the identification of sequences that are unique to a particular species.

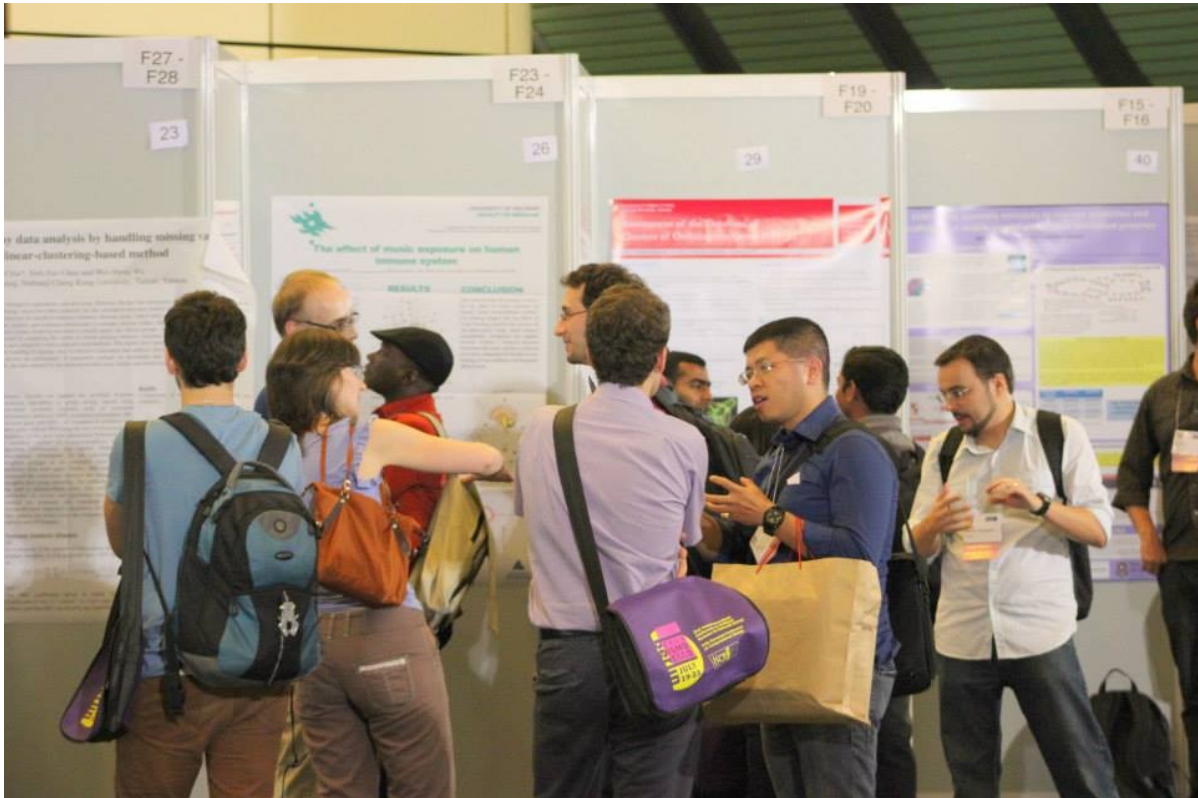
Category : Education

30. Evidence based intervention curriculum for medical students and interns on disease mapping, health and bioinformatics in Nigeria Evidence based intervention curriculum for medical students and interns on disease mapping, health and bioinformatics in Nigeria

Callistus Akinleye, Adewale, Adejimi Adebola Afolake, Olugbenga-bello Adenike I, Olarewaju Sunday Olakunle
University of Ibadan

Introduction: Doctors' knowledge of bioinformatics in Sub Saharan African is still below average. To accelerate biomedical research, reduce healthcare costs, evidence optimal treatment and dosing for clients and patients is required across our health facilities in Africa. This study sort to assess interventions to adapt the curriculum of medical students and interns in Nigeria on the disease mapping, health and bioinformatics. Methods: A cross sectional survey of 400 respondents in both preclinical, clinical years and intern using structural questionnaires at Ladoke Akintola university of technology university teaching hospital Osogbo using the Leslie fisher's formula. Three Focus group discussions among interns was also carried out. Data was analysed with SPSS version 22, and Multivariate regression analysis of the background characters of respondents and primary outcome. The FGDs was transcribed, translated, and analysed using Atlas Ti. Results: The mean age of respondents was 25.7 +/- 2.1 years, with 63% been males and 37% females. Fifty percent were preclinical, 40% clinical students, and 10% interns. Knowledge of health, disease mapping, and bioinformatics was poor (27%-preclinical vs 38%-clinical, 57% intern: $p < 0.005$). Curriculum change to reflect increase in teaching units in disease mapping, and bioinformatics, was suggested as (58% preclinical vs 79% clinical and 85% for interns; $p < 0.05$) . Results from FGDs suggested that National university Commission (NUC) and Medical and Dental Council of Nigeria

Blast from the past



During a poster session at SCS2013, Berlin, Germany

(MDCN) should review the medical schools' curriculum and it should be evidence based. Conclusion: ISCB, NUC and MDCN are needed to accredit evidence based bioinformatics curriculum.

31. Comparative modeling of proteins as a method for engaging students interest in bioinformatics tools

Maina Bitar, Fernanda Badotti, Alan Sales Barbosa, Andre Luiz Martins Reis, Italo Faria do Valle, Lara Ambrosio
Universidade Federal de Minas Gerais

Ever since the invention of the microscope we have seen biology through its lenses. Now, we are looking through the lenses of bioinformatics. And this is a whole new world! In the last few years, universities from all over the world have included bioinformatics in their curricula, encouraging undergraduate students to incorporate computational skills. Nevertheless, despite such efforts to align the current scientific knowledge with undergraduation teaching, there are still extensive gaps to be filled. The aim of this work is to report the experience of awakening students interest in bioinformatics tools during a course entitled "Molecular Modeling of Proteins". The aim of the course was to generate a 3D structure for a provided protein sequence (suggested by professors and other students) through comparative modeling and assess important characteristics of such structure. The class included 33 undergraduate, graduate and post-graduate students from a multitude of University programs and with different backgrounds. To evaluate the course quality, we have randomly selected 15 students to anonymously represent the class. Results indicate that students were able to significantly improve their theoretical knowledge and practical skills on comparative modeling during the course, although no formal statistical analysis was performed. In addition, we observed in our measurements a tendency for the students familiarity to the subject to be much more homogeneous after the course. We expect this work may serve as

a guide for professors who teach subjects for which bioinformatics tools are useful and for universities that plan to incorporate bioinformatics to the curriculum.

Category : Epigenetics

32. eFORGE: a tool for identifying tissue-specific signal in epigenetic data

Charles Edmund Breeze, Dirk S. Paul, Lee M. Butcher, Javier Herrero, Ewan Birney, Ian Dunham, Stephan Beck
University College London

Background: Epigenome-wide association studies (EWAS) provide a novel means of studying the epigenetic basis of human disease. A challenge confronting EWAS though is the assessment of tissue specificity of identified differentially methylated positions (DMPs). To this end, we have developed an analysis approach that determines the tissue-specific regulatory component of a set of EWAS DMPs through the detection of enrichment of overlap with DNase I hypersensitive sites (hotspots) across a wide range of tissues. Our tool, eFORGE (experimentally-derived Functional element Overlap analysis of ReGions from EWAS), is available online (<http://eforge.cs.ucl.ac.uk/>) and provides tabular and graphical summaries of the enrichments. This tool is derived from FORGE (<http://www.1000genomes.org/forg-analysis>), which analyses the tissue-specific regulatory component of SNPs in the context of genome-wide association studies (GWAS). Description: For a given set of significant EWAS DMPs (i.e., 450K array probes), eFORGE generates 1000 randomly selected background sets, matched for gene feature and CpG island relationship, and calculates a binomial P-value of enrichment of overlap for each of the cell types catalogued in NIH Epigenomics Roadmap and ENCODE datasets. eFORGE provided valuable insights into the underlying disease etiology when applied to recently published EWAS. For example, eFORGE enrichments were found in CD56+ cells and thymus tissue in an EWAS for rheu-

matoid arthritis, and CD4+ cells in an EWAS for multiple sclerosis.

Conclusion: eFORGE provides a user-friendly tool to investigate the tissue-specific component of epigenetic marks identified through EWAS, and has the potential to reveal unforeseen tissue involvements leading to mechanistic insights for disease etiology and progression.

33. Performance evaluation and benchmarking of differential DNA methylation analysis methods

Helen McCormick, Eleni Giannoulitou, Peter Hickey, Jennifer Cropley, Catherine Suter
Victor Chang Cardiac Research Institute

DNA methylation is one of the most widely used markers for the study of epigenetic contributions to phenotypic variation and disease. There are several methods for analyzing genome-wide DNA methylation data in common use, but there has been no rigorous evaluation of their performance. We have performed a systematic assessment and comparison of four packages: MethySig, methylKit, eDMR and DSS, using an empirical dataset of 12 reduced representation bisulphite sequencing libraries (6 test, 6 control). Surprisingly, we observed very low concordance among these commonly used model-based and binomial test-based approaches: using equivalent pre-processing and filtering parameters for each method, we found that the four methods identified significant differentially methylated cytosines at a concordance rate of less than 1%. Similarly low levels of concordance were observed with identification of differentially methylated regions using tiled data. Our study highlights the need for systematic approaches to reliable differential methylation analysis via data simulation. This concept of simulation will be discussed in the context of the growing implementation of epigenomic data in human medicine.

Category : Functional Genomics

34. ASpli: an integrative R package for the analysis of alternative splicing using RNA-Seq

Estefanía Mancini, Marcelo Yanovsky, Ariel Chernomoretz
Fundacion Instituto Leloir

Alternative splicing (AS) is a prevalent mechanism of post transcriptional gene regulation in multicellular eukaryotes. It allows a single gene to increase functional and regulatory diversity, through the synthesis of multiple mRNA isoforms encoding structurally and functionally distinct proteins. AS occurs via 4 main events: intron retention (IR), exon skipping (ES) and alternative use of donor and acceptor sites (alt 5' and alt 3'). The development of novel high-throughput sequencing methods for RNA (RNA-Seq) provided a very powerful mean to study alternative splicing under multiple conditions at unprecedented depth. As long as new studies on post-transcriptional regulation arises, there are an increasing evidence than AS frequency is higher than expected. Despite It has become the new standard for studying gene and transcription expression, the use of RNA-seq for the study of transcripts repertoire in a given condition is not trivial. Here we introduce a very flexible and easy to use R package named ASpli. We propose a count based integrative method taking into account gene expression, exon and intron differential usage and their relationship with junctions spanning those features. Using an annotated transcriptome we are able to classify subgenic features into alternative or not alternative regions. ASpli is intended to facilitate the analysis of RNAseq data for the quantification and discovery of AS events and it has been used in many recent publications from our lab. Results of the analysis are presented in a user friendly manner, including plots of the most relevant AS events discovered.

35. Using transcriptograms for gene expression data analysis in whole genome from mammalian cells

Isabela Berger

Universidade Federal do Rio Grande do Sul

One of the great challenges of biological sciences is to determine the exact number and type of genes being expressed in a given tissue or organ, and the expression level and regulation of these genes under a variety of conditions both normal and pathological. The level of gene expression of an organism can reveal information about its condition at a particular time. In this work we performed gene expression analysis of different human cells during the cell cycle process, using a new method for analyzing gene expression in genome whole scale, such as those produced by microarrays or RNA-Seq assays, the Transcriptogramer. Transcriptograms allow to analyze cell profiles in different states, comparing control versus treatment or healthy and diseased cells, identifying relevant points for cell diagnosis. This method yields global assessments of cellular metabolism, providing a genome-wide view of altered expression of pathways not necessarily anticipated in experiments design. It is specially designed to find differentially expressed gene sets with the necessary significance assessments. We show the differences between classes of samples, such as control versus treatment using gene expression analysis by transcriptograms for both microarray and RNA-Seq data. Thus, supporting the fact that the metabolic state of the cells vary greatly during the cell cycle, we could make a characterization of these findings, detecting significant changes in the transcription of genes of interest groups such as metabolic pathways or Gene Ontology terms.

36. Co-regulation of human paralog genes in the three-dimensional chromatin architecture

Jonas Ibn-Salem, Miguel A. Andrade-Navarro

Johannes Gutenberg University Mainz

Introduction: Paralog genes arise from gene duplication events during evolution. The resulting sequence similarity between paralogs often leads to proteins of similar structures and functions in common pathways and protein complexes. Therefore, it can be useful for the cell to have paralogs co-regulated. In eukaryotes, genes are regulated by binding of transcription factors to distal enhancer elements, which perform looping interactions to contact the transcription machinery at gene promoters. These looping interactions can be measured by genome-wide chromatin conformation capture (Hi-C) experiments which revealed conserved megabase-sized self-interacting regions called topological association domains (TADs). We hypothesised that paralogs cluster in the three-dimensional chromatin architecture and share common regulatory mechanisms to enable coordinated expression.

Description: To test this hypothesis, we integrated paralogy annotations with genome-wide data-sets of enhancer-promoter associations, Hi-C experiments, and gene expression in diverse human cell-types. As background control we sampled random gene pairs by taking the linear distances of paralogs and the number of linked enhancers into account. We show that paralog gene pairs share significantly more common enhancer elements than expected. Furthermore, they are located significantly more often in the same TAD and contact each other more frequently than expected. Consequently, paralogs tend to show a positive correlation of gene expression over many cell-types.

Conclusion: Combined, our results indicated that human paralogs share common regulatory mechanisms and cluster not only in the linear genome but also in the three-dimensional chromatin architecture. This en-

ables concerted expression of paralogs over diverse cell-types and indicate evolutionary constraints in functional genome organization.

37. Discerning Systematic Bias In *S. cerevisiae* Pathways Using A Novel Bayesian Statistics Problem Structuring Method

Jacob M. Luber, A.X. Jiang, Matthew A. Hibbs

Trinity University

Background: One of the major problems in computational biology is developing algorithms that cope with multi-functionality of proteins due to the reuse of biological pathways and components in different contexts. Because of this inherent complexity, new work coping with genome scale problems is potentially pressured to conform to existing biases or limitations present in the literature.

Description: We propose an algorithm that utilizes a novel form of Bayesian problem structuring to correlate overfitting with these aforementioned problems in an ensemble of classifiers attempting to predict gene function based on incomplete or inherently biased expert knowledge. We tested this novel form of Bayesian problem structuring through iteratively removing random genes from the well-studied MAPK pheromone response pathway in *S. cerevisiae*. We then subsequently used this modified pathway as a training set in our classifier ensemble that is applied to features from freely available heterogeneous data sources. We then performed cross validation, creating features from the normalized results and the amount of overfitting observed in the first round of classification. Classification outputs from this initial ensemble are converted into features and fed into a second "Black Box" classifier that outputs the probability of expert knowledge being inherently biased.

Conclusion: Early results indicate that this problem structuring method has potential, based on a small subset of very well studied *S. cerevisiae* pathways, to predict pathways that may significantly suffer from study bi-

ases or lack of complete knowledge.

38. Transcriptional networks driven by human endogenous retroviruses mark the naive cells in the midst of conventional human embryonic stem cells

Manvendra Singh

Max-Delbruck Centre for Molecular Medicine

Endogenous Retroviruses (ERVs) occupy more than 8% of the human genome. ERVs are repressed by de novo DNA methylation during gametogenesis but during embryonic development some are demethylated and expressed. We sought to understand both the underlying mechanisms of expression and the downstream consequences, what might be called cross talk. To understand cross-talk we developed pipelines and methods for the analysis of high throughput data to survey the direct and indirect association of several families of repetitive elements with known regulators. Moreover, we also applied in silico approaches to perform cross-species data analysis in order to resolve the evolution of cross-talk. The analysis revealed high and restricted expression of different classes of ERV transcripts marking the different stages of embryonic development. Finally, we estimated human specific ERVs (HERVs) transcripts and its human embryonic stem cell (hesc) specific cross-talk with host-factors which maintains the identity of hesc. In particular we show strategic analysis of high throughput data for ERVs which reveals the differential cross regulation of subsets of HERVs by host factors and vice-versa in vivo and in vitro stem cells. We identified a number of host factors having an evolutionary interplay with ERVs during the primate stem-cell evolution that have re-wired the hesc specific transcriptional networks for natural stem cells such that they resemble cells of the inner cell mass. Our results suggest that only genetically marked human naive cells found, are defined by the expression of selective loci of HERVH via various interactions with host factors.

39. Systematic integration of molecular signatures identifies novel components of the antiviral RIG-I-like receptor pathway

Robin van der Lee, Qian Feng, Martijn A. Langereis, Rob ter Horst, Radek Szklarczyk, Mihai G. Netea, Arno C. Andeweg, Frank J. M. van Kuppeveld, Martijn A. Huynen
Centre for Molecular and Biomolecular Informatics, Radboud University Medical Center, Nijmegen, The Netherlands

Background: The RIG-I-like receptor (RLR) system is critical for the innate defense against viruses. Recognition of viral RNA by the RLRs leads to the production of type I interferons (IFN α/β) that initiate the immune response.

Description: Here, through systematic assessment of a wide variety of available genomics data, we identify 10 molecular signatures of RLR pathway components. Five of these signatures are based on the relationship of RLR signaling with viruses; the other five are based on properties of the pathway itself. We demonstrate that RLR pathway genes, among others, tend to evolve at a high rate, interact with viral proteins, contain a limited set of protein domains, are regulated by a specific set of transcription factors, and form a tightly connected physical interaction network. By weighing these RLR signatures for their ability to predict known RLR pathway genes, and integrating them, we propose 187 novel genes with a likely role in the human RLR system. Using two RNAi screens, we validate an effect on RIG-I-mediated, IFN β promoter-controlled protein production for about half (94) of these genes. For many of the 19 new genes with the strongest effects, knockdown also affects RLR signaling outcome at the level of IFN β transcription, leading to significantly decreased mRNA expression. Finally, by connecting the results with the known protein interaction network, we suggest for several newly identified RLR genes where in the pathway they could function.

Conclusions: We present our prioritized list as a resource for identifying genes involved in the RLR system.

Category : Genome Organisation and Annotation

40. 3D-NOME: 3D NucleOme Multiscale Engine for data-driven modeling of three-dimensional genome architecture

Przemysław Szalaj, Zhonghui Tang, Oskar Luo, Paul Michalski, Yijun Ruan, Dariusz Plewczynski
Medical University of Bialystok, Poland

Human genome is folded into three-dimensional structures. The 3D organization of the genome is thought to facilitate compartmentalization, chromatin organization and spatial interaction of genes and their regulatory elements. Recently developed high-throughput ChIA-PET method allows us to capture the genome-wide map of physical contacts between distal genomic loci. We present a 3D-NOME, a multiscale computational engine we developed to model the 3D organization of the genome. First, we use a bottom-up approach to create a hierarchical model representing the nucleus based on the underlying data features. Then we use Monte Carlo simulations to sequentially reconstruct all the levels in a top-down manner, i.e. we reconstruct more general levels first and we use them to guide the simulation on the following levels. This approach allows us to efficiently model the chromatin folding on a level of whole chromosomes as well as single topological domains. In our modeling we consider CTCF (which is long known to be responsible for chromatin weaving) and RNAPII (which activates genes transcription) interactions. Taken together these two protein factors provides a comprehensive map of human genome interactions. The specificity of the ChIA-PET data allows us to model the shape of individual chromatin loops and their mutual interactions. In this work we describe both the model construction and simulation steps of our algorithm. We do also highlight main advantages of our approach compared to existing methods for genome architecture modeling.

41. Immunogenomics of the Egyptian fruit bat, an important viral reservoir

Stephanie D'Souza, Chandri Yandava, Sean Lovett, Galina Koroleva, Elyse Nagle, Albert Lee, Raul Rabadan, Mariano Sanchez-Lockhart, Jonathan Towner, Gustavo Palacios, Thomas Kepler
Boston University School of Medicine

The Egyptian fruit bat (*Rousettus aegyptiacus*) is the suspected reservoir host for Marburg virus, and there is mounting evidence for the long-term circulation and evolution of the virus in these bats. Currently, there is no available reference genome for the *Rousettus* bat, limiting the ability to study this virus in its natural host. The lack of genomic data for this bat also prevents detailed study of the molecular mechanisms and genetic changes that allow bats of this species to coexist with Marburg and other highly pathogenic viruses. To address this need, we are constructing a high quality, annotated genome of *R. aegyptiacus* with a hybrid assembly approach. Using a combination of short and long read data, we have produced a draft genome of 99,254 scaffolds with 11,314 scaffolds representing 50% of the assembly. We are testing a few hybrid assembly pipelines to improve our current assembly. For automated annotation of the whole genome, we are using *ab initio* software trained with paired-end RNA-Seq data from ten bat tissues. We are simultaneously annotating key immune loci in both innate and adaptive immune systems. Thus far, we have found evidence for seven families of immunoglobulin heavy chain variable genes, and an expanded family of Type I Interferons, an important component of innate antiviral immunity. These reference annotations will open a new suite of tools in bat immunology and will be valuable for assessing how the *Rousettus* bat hosts a virus that is deadly to humans without experiencing any major pathology.

Category : Genetic Variation Analysis

42. Effects of structural variation on fetal hemoglobin expression levels in different populations

Anastasia Gurinovich, Elmutaz Shaikho, Katherine Norwood, Gary Benson, Paola Sebastiani, Martin Steinberg, John Farrell
Boston University

Background: Sick cell anemia and beta-thalassemia are diseases which affect millions of people worldwide. The symptoms of both diseases are modulated by fetal hemoglobin (HbF) levels. HbF prevents the polymerization of the mutant form of adult hemoglobin. HbF expression varies widely among different populations and the different haplotypes associated with these diseases. There exists a need to understand the causes of genetic variation for HbF expression within patients.

Description: Our primary goal is to identify novel genetic variation to account for the different HbF expression levels across 24 whole-genome sequencing samples representative of different populations. We focused our search to the region surrounding the HBB gene cluster located on chromosome 11, since previous studies have shown that regulation for HbF expression is linked to that locus. Initial structural variant analysis was performed using CNVnator, Lumpy and Pindel. CNVnator and Pindel identified deletions within the HBG1 and HBG2 region. Further investigations led us to believe that these deletions are false positive calls due to BWA alignment errors caused by the sequence similarity between HBG1 and HBG2. These gene paralogs encode proteins that differ by only one amino acid.

Conclusions: In order to address this issue, we tested multiple alignment software to improve the alignment mapping quality for the reads in these ambiguous regions. We found that the GEM mapper was the most successful at correcting the misalignment, performing better than both the widely used BWA and BWA-MEM software. This will prove highly beneficial in further downstream

analysis of our data.

43. Inferring clonal evolution from single-cell sequencing data

Edith Ross, Florian Markowitz

Cancer Research UK Cambridge Institute,
University of Cambridge

Tumour evolution leads to genetic intra-tumour heterogeneity, which poses major challenges to cancer therapy. To improve our understanding of tumour evolution reliable methods for inferring tumour phylogenies are needed. To date, most methods use bulk sequencing data, but they struggle to deconvolute the mixed signal into distinct subpopulations. Single-cell sequencing can overcome this challenge, but methods that account for high error rates of single-cell sequencing are needed. We present oncoNEM, a probabilistic method for inferring intra-tumour evolutionary lineage trees from noisy exome- or genome-wide single-cell sequencing data. OncoNEM jointly infers the tree structure, the number of subpopulations, their composition and their genotypes. The core of oncoNEM is a scoring function that is based on the nested structure of mutation patterns of related cells and defines the likelihood of a tree given the observed genotypes. We evaluate the accuracy of oncoNEM in the controlled setting of a simulation study and demonstrate that it (i) can accurately infer trees of tumour evolution despite the high error rates of current single-cell sequencing technologies, (ii) is robust to inaccuracies in the estimation of model parameters and (iii) substantially outperforms competing methods. We also present results on three published single-cell data sets. In summary, oncoNEM is an accurate probabilistic method for inferring intra-tumour phylogenies from single-cell sequencing data. It identifies subpopulations within a sample of single-cells and estimates their evolutionary relationships. In simulations, oncoNEM performs well for realistic error rates, is robust to inaccuracies in the estimation of model parameters and outperforms competing methods.

44. Molecular characterization of *Chlamydomphila pneumoniae* detected in Moroccan patients with cardiovascular diseases

El Yazouli Loubna, Hicham Hejjaji, Aziz Alami, Naima Elmdaghri, Nadia Dakka, Fouzia Radouani
Institut Pasteur du Maroc

Chlamydomphila pneumoniae (*C.pneumoniae*) is an obligate intracellular gram negative bacterium, which cause respiratory diseases and it's strongly involved in cardiovascular diseases. Both whole-genome sequencing and specific gene typing suggest that there is relatively little genetic variation in human isolates of *C. pneumoniae* worldwide. The main objective of this study is to identify and characterize *C. pneumoniae* detected strains in peripheral blood mononuclear cells (PBMCs) of Moroccan patients with cardiovascular diseases by sequencing. A total of 54 *C.pneumoniae* strains DNA detected in PBMCs were amplified by nested PCR using the primers cpn5P/Cpn3P and Cpn5N/Cpn3N respectively in the first and second round. These primers are targeting 366 bp fragment of the variable domain 4 (VD4) region of the outer-membrane protein-A (ompA) gene. The PCR products were sequenced in both directions using BigDye Terminator Chemistry in an automatic capillary DNA sequencer, chromatograms were analyzed with Bio Edit and the sequences were analyzed using nucleotide Blast. The sequences analysis showed a high homology with human *C.pneumoniae* strains, with more than 99% similarity. For more accuracy of strains analysis, the sequences were compared with *C.pneumoniae* reference strains, this comparison revealed different SNPs. Deep analysis of the sequences is in process for a good epidemiologic characterization. Key words: *Chlamydomphila pneumoniae*, Molecular characterization.

45. Etiology of common diseases: a dialog between gwas and electronic health records

Olubukola Smile, Segun Fatunmo, Oyekanmi Nashiru

University of Ibadan Nigeria

Background: Genome-wide association study provides an insight to the role of genetic variants in the etiology of common diseases . Hundreds of associations of common genetic variants predisposing individuals to many complex diseases have been identified. GWAS usually identifies genetic variants associated with the phenotype in question (disease) by comparing the DNA of two groups of people. The first set of people are those with the disease (cases) and the second group are those without the disease (control). Variants that are more common in the cases are said to be likely associated with the disease. The associated variants are then considered to mark a region of the human genome which influences the risk of disease. Genetic variants however are just a part of other factors that leads to acquiring diseases. Genetic variants are susceptible to certain environmental factors that increase the risk of acquiring such diseases.

Description: Most diseases are not caused by genetic variants alone but are a product of genetic variants, environmental factors and/or the interaction between the two.

Conclusions: Findings from GWAS will revolutionize the field of human medicine, only if other variables in the pathogenesis of diseases are well understood. Environmental factors, environmental-genetic factors and genetic factors must all be studied not in isolation but in combination to bring to relevance the result of GWAS.

46. A Novel Approach to Gene Expression Analysis of Ethnicities

Prathik Naidu

Thomas Jefferson High School for Science and Technology

Microarrays are important tools to measure the gene expression levels in cells. However, these technologies have major limitations, such as their inability to detect genetic aberrations. A computational method was developed and implemented to analyze gene expression patterns more efficiently than current technologies. This approach was applied to biologically understand the gene expression patterns in the American European, Great British, and Yoruba African ethnic groups. The computational method aligned the RNA reads to the reference genome, assembled the transcripts, and performed the gene expression analysis. These steps were streamlined into a software package named DGEST, which expedites the discovery of differentially expressed genes using RNA-sequencing. Statistical tests, implemented in the R programming language, filtered the large gene set from the DGEST analysis into a smaller group of differentially expressed genes. The aligned reads and protein isoforms were then visualized in a genome browser and the Protein Atlas database. Initial results showed that 74 genes with the lowest q-value and the largest log(FPKM) expression value were the most differentially expressed. Protein modeling for the isoforms of these genes indicated differences primarily in cellular processes. The CCL4L1 gene was chosen for further analysis, and indicated higher expression in the Yoruba population, suggesting greater ability for HIV inhibition. In addition, the computational method yielded 95% accuracy in the RNA read mapping, illustrating the efficacy of this approach for gene expression analysis. The results gleaned from this method can assist in the development of targeted cancer therapeutics and optimized drug designs for ethnicities.

Category : Metagenomics

47. Characterization of microbial diversity in both mediterranean and atlantic Moroccan coastal lagoons by a metagenomic approach.

Bouchra Chaouni, El Houcine Zaid, Hassan Ghazal

Faculty of Science of Rabat, University Mohammed V-Agdal and University Mohammed Premier, Morocco

Microbial metagenomics is a new approach allowing the study of the variation of species in a complex microbial sample. Using NGS it is possible to investigate microbial communities by sequencing the marker gene 16S or sampled fragments of the whole genome or transcriptome without the need for culturing. Marine ecosystems are among the most attractive fields in metagenomics. In our present study we chose to target the lagoons as ideal and specific marine locations to study microbial biodiversity. Marchica lagoon displays a major socioeconomic interest for that region. Its biogeochemical features intensely studied have contributed to classify it as the most interesting lagoons among the Mediterranean coast. Unfortunately, it's subject to anthropogenic activities and eutrophication, which has led to massive phytoplankton blooms and thus a reduction of the microbial diversity. In order to protect this ecosystem, several studies have been launched to define its microbial community. However, there is a lack of information regarding many microorganisms, which need to be identified. The aim of our study is to explore the microbial diversity of Marchica using the metagenomics approach. The outcome will be compared to results from other aquatic ecosystems, with regard to suggest which environmental factors influence the selection of microorganisms that might have an impact on the environmental and human health. This will be the first time where such approach is applied in Morocco to study aquatic ecosystems. This work is also part of the Ocean Sampling Day consortium where our group is the coordina-

tor for the Moroccan ODS network.

Category : Pathogen Informatics

48. PathoAssem: A Tool for Genome and Transcriptome Assembly of Novel Pathogens

Demarcus Briers, Sean Corbett, Junmin Wang

Boston University

Background: Pathoscope is a pipeline for rapidly removing host contamination, isolating microbial reads, and identifying disease-causing pathogens in next-generation sequencing data from metagenomic samples. Despite performing pathogen identification, Pathoscope lacks capabilities to generate genome/transcriptome assemblies. Assembly approaches such as de novo and reference-based transcriptome assembly could be used to characterize novel genomic features of pathogens identified by Pathoscope.

Description: We present PathoAssem, an application to incorporate assembly capabilities into the Pathoscope pipeline utilizing pre-existing assemblers. After identifying the most prevalent pathogen in each, we used PathoAssem to assemble reads collected from various environmental sources: a lake trout from a fishery facing population decline, a Titimoney with a presumed adenovirus infection, and nasal brushings from a child with asthma. After assembling pathogens isolated from environmental samples, we used various quality control procedures to probe the accuracy of these assemblies. Based on this analysis, we assembled a described Titimoney adenovirus with good quality, as well as *Moraxella catarrhalis*, a well known respiratory microbe from the nasal brushing samples, with lower coverage. We were unable to definitively assemble a complete genome from the lake trout, due to a majority of reads aligning in tandem repeat regions.

Conclusion: Our efforts to perform assemblies via metagenomic samples were met with varying success. The functionality provided by our software module offered few

insights into the novel pathogens, but allowed us to uncover pitfalls in dealing with metagenomic samples which will inform future assembly efforts, as well as potential issues with PathoScopes species attribution methods.

49. *In silico* global analysis of the interaction of violacein with *Staphylococcus aureus*'s possible protein targets.

Fernanda Luz Paulino da Costa, Bruna de Araujo Lima, Ana Carolina Guimarães Cauz, Marcelo Brocchi
UNICAMP

Staphylococcus aureus is one of the most important human pathogens that cause life-threatening infection and continues to be a serious health problem due to its ability to develop resistance to a great number of antibiotics. Violacein is a small molecule, probably a residue from a secondary metabolism of some known environmental bacteria. Studies have shown that violacein have antiviral, antiparasitic, antitumor and antimicrobial activities, so it's a great antibiotic candidate to be tested against *S. aureus*. We want to find if violacein have a significant interaction to *S. aureus*'s proteins, disrupting the bacteria's main metabolism. We have used *in silico* experiments to analyze the interaction of violacein with *S. aureus*'s proteins. All proteins from *S. aureus* deposited in PDB was used, clustered in groups of 90% of similarity and then used as an input data at Pipeline Pilot. This software was used to read the data and to do the protein-ligand docking. The top five poses (according to RMSD docking score) were selected to further analysis. We found two proteins worth mention, Aminoacyltransferase and Glycyl-glycine endopeptidase, involved at peptideglycan biosynthetic process and cell wall organization, respectively. It suggest that violacein must interact with proteins that maintain the bacteria's integrity and are important for its survival. This initial test imply that violacein is a good antimicrobial candidate. Further studies are needed to better understand the nanoenviron-

ment of the molecular interactions between violacein and *S. aureus*'s proteins and its role as a possible antibiotic, specially against multi-resistant strains of *S. aureus*.

50. *Comparison of TAL effectors and their predicted host targets across diverse strains of the rice bacterial leaf streak pathogen *Xanthomonas oryzae* pv. *Oryzicola**

Katherine Wilkins, Nicholas J. Booher, Li Wang, Adam J. Bogdanove
Cornell University

Bacterial leaf streak of rice, caused by *Xanthomonas oryzae* pv. *oryzicola* (Xoc), can cause yield losses of 30% in this staple crop. Disease progression is mediated in part by pathogen-secreted transcription activator-like (TAL) effectors that bind host gene promoters to upregulate corresponding genes. Across *Xanthomonas* species, TAL effectors upregulate susceptibility genes critical for disease and or resistance genes that trigger a host defense response. Knowledge of these important targets informs breeding of resistant rice varieties. Each TAL effector possesses a central repeat region that determines binding specificity, with repeat variation specifying binding site sequence in a one-to-one fashion based on a degenerate "code". We sequenced the complete genomes of ten diverse strains of Xoc to identify their TAL effector content and used RNA-Seq to compare the transcriptional response of rice inoculated with each strain. Xoc TAL effectors are highly conserved compared to those of the closely related *X. oryzae* pv. *oryzae*. Phylogenetic analysis suggests Xoc TAL effectors are primarily vertically transmitted, although the only Xoc TAL effector known to upregulate a susceptibility gene appears to have been horizontally transferred. Prediction of TAL effector targets using the TAL effector-DNA recognition code yields many false positives, with predicted binding sites in every rice gene for each Xoc strain. We filtered these predictions using a machine learning classifier we previously developed and requiring correlation between TAL effector

presence and gene upregulation across strains. This filtering results in testable numbers of candidate TAL effector targets of potential importance for rice breeding.

51. Antigenic variation and recent evolution of the dengue virus envelope protein

Sarah L. Keasey

United States Army Research Institute of Infectious Diseases

Dengue is an infectious disease that is now endemic to most tropical or semi-tropical regions of the world. The dengue virus (DENV) circulates in four serotypes (DENV-1, 2, 3, and 4) within the human population, and antibody responses provide incomplete protection across the distinct serotypes. The DENV envelope protein is a primary target of neutralizing antibodies, suggesting that persistence of autochthonous infection cycles is partially driven by selective evolution of this receptor molecule. We examined antigenic surfaces of the DENV-2 envelope protein from isolates collected over 7 decades. Bayesian analysis of envelope protein sequences estimated the emergence of the DENV-2 serotype in the late 1800s and revealed significant changes in amino acid residues that were linked to progressive time of isolation. Protein phylogenies recapitulated known aspects of DENV genome evolution, for example divergence from sylvatic strains and clustering of sequences by geographical location of virus isolation from human infections. Using entropy as a measure of sequence variability, we identified highly substituted residues that mapped to surface-exposed regions of the envelope protein monomer as displayed on the mature virus. Notably, a 50% increase in the number of variable surface residues is evident in the envelope protein of DENV-2 isolates in comparison to the other serotypes, and these highly-variable surfaces are dominated by residues that are targets of neutralizing antibodies. The discordant evolution of the DENV-2 envelope protein compared to other viral serotypes should be considered in ef-

forts to control infection by vaccination.

Category : Population Genetics Variation and Evolution

52. MAAD: Moroccan Ancestry Associated Diseases

Yassine Souilmi, Muhammad Ilyas, Hassan Ghazal, Saaid Amzazi, Peter J. Tonellato
Mohamed Vth University

Background: Genome-wide association studies (GWAS) are a powerful tool to identify associated risk alleles in a given population, and the increased genetic variation in admixed populations may increase disease risk variants detection. However, in Morocco we have a highly admixed population, and the lack of resources to conduct GWAS and the absence of systematic clinical genetic testing makes any genetic risk factors estimation impossible.

Description: To overcome this weakness, we propose an alternative approach using the population genetic admixture, where we use the ancestry information for the genomic region of interest to identify the risk alleles for the gene(s) and variants of interest. We use publicly available SNP array data of randomly sampled individuals with no associated phenotype to avoid any disease bias in the sample.

Conclusions: We developed a methodology to study these complex genetic diseases in Moroccans. We believe that the admixture mapping may be a useful approach to classify the genetic determinants of diseases.

Category : Protein Structure and Function Prediction and Analysis

53. Protein disorder promotes protein conformational diversity

Alexander Monzon, Diego Zea, María Silvina Fornasari, Silvio Tosatto, Gustavo Parisi
National University of Quilmes

Background: Large-scale analysis of protein conformational diversity using RMSD as a measure of conformer similarity showed that this distribution has a peak at low RMSD (0.4-0.8Å) but with a skew towards higher RMSD values.

Description: In this work we studied the relationship between conformational diversity and the occurrence of protein disorder. Using 2383 proteins with their corresponding conformers taken from CoDNAs database (36200), we measured RMSD between all the conformers for each protein (almost 800,000 comparisons), along with disorder percentage and number of disordered regions in each conformer. Disordered regions were taken as those with at least five consecutive missing residues (terminal ends were excluded). A protein was classified as “ordered” if all their conformers showed no disordered regions, and as “disordered” protein when shows at least one conformer with a disordered region.

Conclusions We found that ordered and disordered maximum RMSD distributions are different (0.81 and 1.24Å for average RMSD respectively) and above 0.9Å of RMSD there is a 2.5 times enrichment in disordered proteins. Interestingly, ordered and disordered sets differ at the molecular function and cellular compartment levels when an enrichment test with GO terms were performed. Also, in the disordered set, most proteins (587) have their maximum RMSD in a disordered pair of conformers, while the rest (251) in an ordered pair. These subsets differ in the percentage of maximum disorder per protein and in the number of disordered regions. We think that these distributions could reflex functional adaptations depending on protein flexibility and disorder content.

54. Determining the winning SH3 coalition: how cooperative game theory reveals the importance of domain residues in peptide binding

Ashley Mae Conard, Dr. Elisa Cilia, Dr. Tom Lenaerts
Université Libre de Bruxelles

Cell signaling relies on protein-protein and protein-peptide interactions involving signaling domains, which typically recognize specific peptide motifs. For instance, SH3 domains bind preferably to proline-rich amino acid motifs. Phage-display experiments allow one to determine those motifs and whether surface or core domain mutants gain or loose preference for peptide motifs. Here, we present an approach utilizing the Shapley Value from Cooperative Game Theory to determine the importance of seven residues in the Fyn SH3's hydrophobic core. The core positions and the residues in those positions represent the players of a cooperative game in which the worth of each coalition is measured through its capacity to discriminate the binding and non-binding mutants for certain classes of peptides. The players (positions or residues) can be seen as the features of SH3 mutants in a binary classification task. Essentially, we use a feature selection method based on the Shapley Value to assign a payoff to each core position and residue. We quantify their importance to promote peptide binding as well as their joint effects, and their interactions, represented through networks. Our results provide novel insights suggesting that the Fyn SH3 domain must contain different signatures of amino acids to promote binding to various peptide classes. This analysis highlights residue importance for proper domain function, which helps scale conservation profiles (e.g. WebLogo) by adding functionally relevant properties. These detailed pieces of information contribute an effective and novel approach to understanding the role core residues play, next to normally investigated binding-site residues, in binding specific peptides.

55. Unraveling the First Pakhtun Genome for Clinical and Evolutionary Inference

Muhammad Ilyas

Harvard Medical School

Background: Pakistan covers a key geographic area in Asia and played an important role in human history, being both part of the Indus River region that acted as one of the cradles of civilization and as a link between Western Eurasia and Eastern Asia. This region is inhabited by a number of distinct ethnic groups, the largest being the Punjabi, Pathan (Pakhtun), Sindhi, and Baloch.

Description: A total of 3.8 million single nucleotide variations (SNVs) and 0.5 million small indels were identified (Novel: 129,441 and non-synonymous SNVs: 10,315 in 5,344 genes). SNVs were annotated for health consequences and high risk diseases, as well as possible influences on drug efficacy. We confirmed that the Pakistani genome presented here is representative of this ethnic Pakhtun group. Finally, we reconstruct the demographic history by PSMC, which highlights a recent increase in effective population size compatible with admixture between European and Asian lineages expected in this geographic region.

Conclusion: We present a whole-genome sequence and analyses of an ethnic Pakhtun from the north-west province of Pakistan. It is a useful resource to understand genetic variation and human migration across the whole Asian continent. The genotype data and annotated variations for Pakistani genome will represent a valuable public resource enabling clinical genetics research and diagnostics.

56. Organism specific protein-RNA recognition: A computational analysis of protein-RNA complex structures from different organisms

Nagarajan Raju, Sonia Pankaj Chothani, Ramakrishnan C, Sekijima M, Gromiha MM
Indian Institute of Technology Madras

Background: Understanding the recognition mechanism of protein-RNA complexes has been a challenging task in molecular and computational biology. In this work, we have constructed 18 sets of same protein-RNA complexes belonging to different organisms. The similarities and differences in each set of complexes have been revealed in terms of various sequence and structure based features such as root mean square deviation, sequence homology, propensity of binding site residues, conservation at binding sites, binding segments and motifs, preferred amino acid-nucleotide pairs and influence of neighboring residues for binding.

Description and Results: We found that the proteins of mesophilic organisms have more number of binding sites than thermophiles and the binding propensities of amino acid residues are distinct in *E. coli*, *H. sapiens*, *S. cerevisiae*, thermophiles and archaea. Proteins prefer to bind with RNA using a single residue segment in all the organisms whereas RNA prefers to use a stretch of up to six nucleotides for binding with proteins. We developed amino acid residue-nucleotide pair potentials for different organisms, which could be used for predicting the binding specificity. Further, molecular dynamics simulation studies on aspartyl tRNA synthetase complexed with aspartyl tRNA showed specific modes of recognition in *E. coli*, *T. thermophilus* and *S. cerevisiae*.

Conclusion: Based on the structural analysis and molecular dynamics simulations we suggest that the mode of recognition depends on the type of the organism in a protein-RNA complex.

57. Gene Prioritization through Geometric Kernel Data

Pooya Zakeri, Sarah ElShal, Yves Moreau

1-Department of Electrical Engineering (ESAT), STADIUS Center for Dynamical Systems, Signal Processing and Data Analytics, KU Leuven, and iMinds Medical IT

Background: In biology, there is often the need to discover the most promising genes among large list of candidate genes to further investigate. While a single data source might not be effective enough, integrating several complementary genomic data sources leads to more accurate prediction. Finding an efficient technique for fusing these complementary genomic data has received increasing attention.

Description: We propose a kernel-based gene prioritization framework using geometric kernel fusion (GKF) which we have recently developed as a powerful tool for protein fold classification and protein sub-nuclear localization. It has been shown that GKF is less sensitive in dealing with complementary and noisy kernel matrices compared to standard multiple kernel learning methods. Since genomic kernels often encodes the complementary characteristics of biological data, this leads us to research the application of GKF in the gene prioritization task. We use several genomic data sources including annotation-based data sources, protein-protein interaction networks, microarray data, literature, and sequence-based protein features.

Conclusions: We develop a prospective benchmark based on the OMIM associations. For this we use an old version from OMIM (back in 2010), and a newer one (recent in 2013). We automatically adjust both versions such that we have similar disease entries for which we have at least one delta gene (reported after 2010). Afterwards, we manually verify the delta genes which we use in our prospective analysis. Experimental results on both our prospective benchmark and the ENDEAVOUR benchmark show that our model can improve the accuracy of the state-of-the-art gene prioritization model.

58. Local Frustration and the Energy Landscapes of Ankyrin Repeat Proteins

R. Gonzalo Parra, Rocio Espada, Nina Verstraete, Diego U. Ferreira

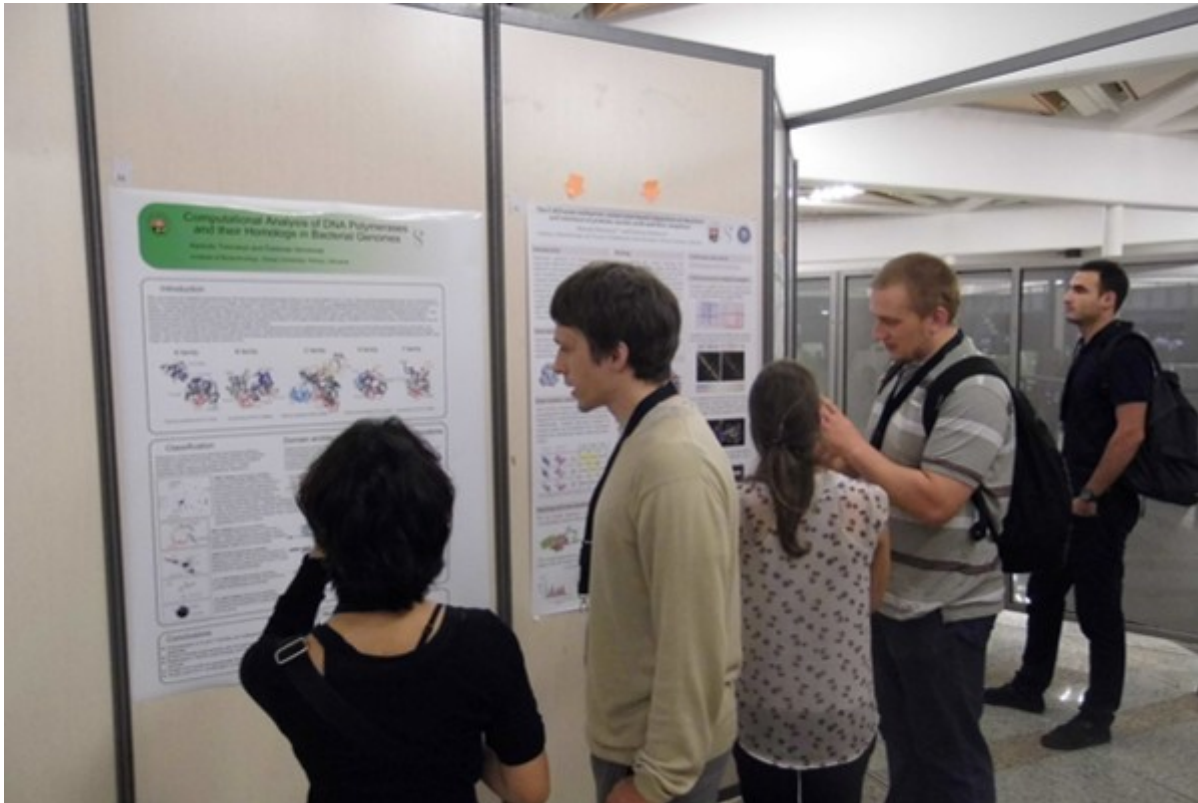
Protein Physiology Lab, Buenos Aires University

Introduction: Protein folding is today understood within the “Energy Landscapes Theory”. In order to fold robustly, proteins must minimize their internal conflicts, following the Principle of Minimal Frustration. However, residual frustration can provide evolutionary advantages on protein native states ensembles, sculpting protein dynamics and modulating protein function. Ankyrin Repeat Proteins (ANKs) are composed of multiple copies of a ~33 residues motif that typically fold into elongated structures. Since residue-residue interactions remain local within and between neighboring repeats, these systems are useful to quantify how local energetics impacts on folding and function.

Description: We have applied a tessellation approach in order to detect and compare the structural variations on all the known structures of ANKs. We calculated the local frustration patterns over these structures and describe which parts of the repeat arrays tend to be energetically (un)favorable for protein folding. We have performed folding simulations of several ANKs, in order to describe their folding mechanisms and compare the results with the previously described parameters.

Results: Energetic conflicts are not randomly distributed over the canonical framework of ANK repeats. Enrichment of highly frustrated interactions in the residues that surround insertions, binding sites and deletions is observed. Highly conserved residues at the sequence level are connected by a network of minimally frustrated interactions. Despite their high structural similarity, ANKs with the same number of repeats can display different folding dynamics with variable levels of complexity. The effects of local frustration and structural variations in the folding ensembles will be discussed.

Blast from the past



Interactive Poster Session, ESCS2014, Strasbourg, France

59. Computational Analysis of Conformational Changes stimulate by Mutants of Macrophage Infectivity Potentiator from Chlamydia trachomatis Catalytic Residues and its Interactions with Rapamycin

Ramachandran Vijayan, Naidu Subbarao, Natesan manoharan
Bharathidasan University, India

The Legionella pneumophila, causes Legionnaires' disease. This human pathogen produces a major virulence factor, called 'macrophage infectivity potentiator protein' (Mip), that is essential for multiplication of the bacteria in human alveolar macrophages. Mip exhibits peptidyl prolyl cis-trans isomerase (PPIase) activity, which can be inhibited by Rapamycin and FK506. Homologous proteins are Chlamydia trachomatis and Chlamydia pneumoniae. Chlamydia causes Sexual transmitted disease, Trachoma and Pneumonia. Mutation of Legionella Mip protein on catalytic residues at Aspartate-142 position replaced to Leucine-142 and Tyrosine-185 position replaced to Alanine-185 that strongly reduces the PPIase activity as reported earlier. In order to design a drug for treating Chlamydia infections, we aim to design an in-silico mutagenesis model of Chlamydia trachomatis Mip for both important catalytic residues, validated the stability of the mutated model. Further, we have docked to the known inhibitor rapamycin with Chlamydia trachomatis Mip (native) and mutants (D170L and Y213A) to examine the details of conformational changes occurred in the binding site. For electrostatic contributions and VanderWaals interactions are important for rapamycin binding and responsible for the binding differences between the Chlamydia trachomatis Mip (native and mutated) proteins. Thus, the observations provide new insights into the structure and function relationship of Mip would be help for designing new drugs against Chlamydia pathogens.

60. FunFHMMer: protein annotations using functional family assignments

Sayoni Das, Ian Sillitoe, David Lee, Jon Lees, Natalie Dawson, Christine Orengo
University College London

Background: Due to the rapid increase in international genome-sequencing initiatives and structural genomics projects, a large amount of protein sequence and structural data are accumulating. Since experimental characterisation of such huge amounts of data is not feasible, computational approaches that can predict protein functions are essential.

Description: We present the FunFHMMer web server (<http://www.cathdb.info/funfhmmer>) which provides automated, domain-based protein function predictions for query sequences based on the functional classification of the CATH-Gene3D resource. CATH-Gene3D provides a comprehensive classification of structure and sequence domains into 2735 structure-based superfamilies, which have been further classified into functional families (FunFams) using a new method, FunFHMMer. This functional classification helps to improve the functional annotation of uncharacterised protein domain sequences assigned to an annotated FunFam within the superfamily and also understand the mechanisms of functional divergence in a superfamily during evolution.

Conclusion: The functional purity of the FunFams was assessed using a set of manually-curated mechanistically diverse enzyme superfamilies, consistency of EC annotations within the FunFams, and an in-house residue enrichment analysis. The preliminary results of the recent CAFA international function prediction experiment have ranked FunFHMMer among one of the top methods for "molecular function" and "biological process" function prediction. Further validation of our function prediction protocol comes from a CAFA-like benchmarking on a test set of 1174 proteins generated by a rollback of the UniProtKB/SwissProt database for which FunFHMMer (Fmax: 0.65) shows better performance compared to BLAST

(Fmax: 0.58) and protein family resources like Pfam (Fmax: 0.58) and CDD (Fmax: 0.6).

61. Validation of an automated procedure for the prediction of potential evolutionary pathways of imp metallo-beta-lactamases

Yifang Liu

Tsinghua University

Background: The indiscriminate use of antibiotics has led to the development of widespread antimicrobial resistance. IMP metallo-Beta-lactamases (MBL) is a new class of β -lactamases (sub class B1) that has a broad substrate spectra. It is capable of hydrolyzing amide bond, thereby inactivating not only various broad-spectrum Beta-lactams but also carbapenems. One important notion in antibiotic resistance is that mutations can be induced by the selective pressure of antibiotics that they will likely evolve into even more efficient enzymes and survive at higher levels of these drugs. In this study, we investigated aspects of MBL activity that cannot be easily observed experimentally by gathering information of MBL sequences, reporting date of their evolution and their ability to hydrolyze various antibiotics. This work will provide the background for the development of newer types of Beta-lactamases resistant antibiotics.

Description: With confidence that our modeling approach provides an adequate model of reality, we went on to the next step and employed the same approach to predict novel MBL point mutants. We have employed MD simulations in MMFF94 force field to calculate the enzyme substrate free binding energy and shows a pretty significant correlation relationship between the catalytic efficiency and MM/GBVI.

Conclusions: In summary, a new method for the prediction of IMP metallo-Beta-lactamases evolution pathway has been established. Taken together, the amalgamation of various disciplines allows for a greater understanding into the development of enhanced MBL-conferred antibiotic resistance. This computational-based approach may have the

potential of guiding the clinical use of available antibiotics.

Category : Proteomics

62. Interactive and integrated visual analytics of Mass Spectrometry Imaging data using MSIdeas

Xian MAO, Nico Verbeeck, Yousef El Aalamat, Etienne Waelkens, Bart De Moor

Department of Electrical Engineering (ESAT), STADIUS Center for Dynamical Systems, Signal Processing and Data Analytics, KU Leuven, Belgium

Background: Mass spectrometry imaging (MSI) is a powerful bio-analytical technique that plays an increasingly important role in a growing number of fields including fundamental biology, medical research and forensic science. In a single experiment, MSI provides the spatial distribution and abundance of a large range of biochemical content in a tissue slice without the need for molecular labelling. The data obtained from this technique is of very high dimensionality, complexity and heterogeneity, which makes extraction of the useful information from these data challenging.

Description: In order to facilitate the interpretation of MSI data, we have developed MSIdeas as an integrated multifunctional software. MSIdeas takes advantage of selected effective algorithms, interactive visualization and fast computing while remedying the shortcomings of them. This free, standalone application enables users from broader fields to gain instant insights on any scale of their MSI datasets. Moreover, thanks to its modular design, MSIdeas aims to integrate the promising data mining algorithms in the MSI community into a single platform, allowing for easy comparison of the results delivered by these different algorithms. Additionally, its intuitive and novel visualizations provide quickly access to the different aspects of the MSI dataset for users. Briefly, MSIdeas intends to address the community's need for vendor-neutral software to identify key char-

acteristics of their MSI data in the simplest, clearest and most informative way.

Conclusions: We present MSIdeas as a software tool that can facilitate the exploration, analysis and interpretation of MSI data.

Category : Sequence Analysis

63. Ensemble Multiple Sequence Alignment via Advising

Dan DeBlasio, John Kececioğlu
University of Arizona

The multiple sequence alignments computed by an aligner for different settings of its parameters, as well as the alignments computed by different aligners using their default settings, can differ markedly in accuracy. Parameter advising is the task of choosing a parameter setting for an aligner so as to maximize the accuracy of the resulting alignment. We extend parameter advising to aligner advising, which chooses among a set of aligners to maximize accuracy. In the context of aligner advising, default advising selects from a set of aligners that are using their default settings, while general advising chooses both the aligner and its parameter setting. We apply aligner advising for the first time, to obtain a true “ensemble aligner.” Through experiments on benchmark protein sequence alignments, we show that parameter advising for a fixed aligner gives a significant boost in accuracy over simply using its default setting, for all of the most accurate aligners currently used in practice. Furthermore, for ensemble alignment, default aligner advising gives a further boost in accuracy over parameter advising for any single aligner, and general aligner advising improves beyond default advising. Our new ensemble aligner that results from general aligner advising, when evaluated on standard suites of protein alignment benchmarks, and selecting from a set of four or more choices, is significantly more accurate than the best single default aligner.

64. A novel feature selection method to extract multiple adjacent solutions for viral genomic sequences classification

Giulia Fiscon, Emanuel Weitschek, Massimo Ciccozzi, Paola Bertolazzi and Giovanni Felici

Department of Computer, Control, and Management Engineering (DIAG), Sapienza University of Rome.

Background: Leveraging improvements of next generation technologies, genome sequencing of several samples and in different conditions led to an exponential growth of biological sequences. However, these collections are not easily treatable by biologists to obtain a thorough data characterization and require a high cost-time investment. Therefore, computing strategies and specifically automatic knowledge extraction methods that optimize the analysis focusing on what data should be sequenced are essential.

Description: We present a new feature-selection algorithm based on mixed integer programming methods able to extract equivalent, multiple, and adjacent solutions for supervised learning problems applied to biological data. In particular, we focus on those problems where the relative position of a feature (i.e., nucleotide locus) is relevant. In this connection, we aim to find sets of discriminating features, which are as close as possible to each other. Our algorithm has been successfully integrated in a rule-based classification framework and applied to three viral datasets (i.e., Rhino-, Influenza-, Polyomaviruses). We succeeded in extracting a wide set of equivalent separation rules focusing on small regions of sequences with high accuracy and low computational time.

Conclusions: Our algorithm enables to extract all the alternative classification solutions of virus specimen to species assignments, by identifying multiple portions of sequence that are distinctive, compact, and as shorter as possible, in order to provide the biologists with small genome parts to be sequenced. Finally, we obtain advantages in term of sequencing cost and time, as well as a powerful

instrument both scientifically and diagnostically (i.e., automatic virus detection).

65. *Jabba: Hybrid Error Correction of Long Sequencing Reads*

Giles Miclotte, Mahdi Heydari, Pieter Aude-naert, Piet Demeester, Jan Fostier

Ghent University, Department of Information Technology - iMinds, Internet Based Communication Networks and Services (IBCN)

Background: Third generation sequencing techniques produce longer reads with higher error rates than second generation methods. While the improved read lengths can provide useful information for downstream analysis, the higher error rates can complicate the required mapping or alignment. Hybrid strategies have been proposed to correct the long reads using accurate short reads. Mapping short reads on long reads may eliminate up to 99% of all errors in bacterial datasets, however this requires significant amounts of computing resources. Mapping the long reads on a k-mer frequencies based de Bruijn graph is significantly more efficient, but loses some accuracy on larger genomes.

Description: We present Jabba, a hybrid method to correct long reads by mapping them on a corrected de Bruijn graph. First, accurate second generation reads are used to build a de Bruijn graph, which is then corrected based on standard topological graph correction methods. Finally a path in the graph is then found by using maximal exact matches between a long erroneous read and the nodes of the de Bruijn graph. This path then dictates the corrected sequence.

Conclusions: Jabba achieves comparable gain to other available tools for bacteria and other small genomes. For larger genomes Jabba keeps performing well, while others either can not practically handle these at all, or only at significantly reduced gain.

66. *A comprehensive analysis of human transcription factor targets*

Liu Wei

Huazhong University of Science & Technology

Background: Transcription factors (TFs) are key regulators for gene expression. Diverse computational and experimental efforts were made to elucidate the control circuitry regulating the transcription of human genes, and generated massive datasets without systematic mining and integration. Moreover, there are still missing integrated tools for fetching TF targets easily and accurately.

Description: We collected all publicly available ChIP-Seq datasets of human TFs, including 627 datasets from ENCODE project and another 688 datasets from NCBI GEO. Peaks were identified using our developed pipeline built upon bowtie and MACS. We introduced a novel algorithm to reduce the false positives by removing peaks which don't contain enriched motifs. Furthermore, targets identified by individual dataset are integrated for further reduction of the false positives. Simultaneously, conserved TF binding sites among Human, Mouse, and Rat were predicted as putative targets. In total, we identified the targets of 236 TFs using ChIP-seq data and the targets of another 637 TFs using TFBS prediction method. All these data are integrated and visualized in hTFtarget, which will be the most comprehensive database of human TF targets. The data and rules of human TF targets were also summarized.

Conclusions: We have deployed a comprehensive database for human TF targets by integrating publicly available experimental resources and computational prediction results. It will be very useful resource for transcriptional regulation.

67. Brownie: correcting second generation sequencing errors using de Bruijn graphs

Mahdi Heydari, Giles Miclotte, Yves Van De Peer, Jan Fostier

Department of Information Technology, Internet Based Communication Networks and Services(IBCN) and iMinds, University of Ghent, Belgium

Background: Next-generation sequencing (NGS) methods enable the production of huge amounts of sequencing data at a low financial cost. However, the presence of sequencing errors in these data challenges applications like de novo assembly methods, potentially causing a suboptimal assembly quality. Therefore, several standalone applications have been proposed that are specialized in the correction of these errors in order to provide cleaner data for downstream analysis tools.

Description: We introduce Brownie for the correction of errors in sequencing data generated from the Illumina platform. Brownie builds a De Bruijn graph from all reads with a user defined k-mer size and applies graph correction algorithms by taking into consideration both the graph topology and statistical evidence in order to detect erroneous nodes. Subsequently, the input reads are individually aligned to the corrected De Bruijn graph in order to correct them by finding mismatches between the reads and the sequence represented by the graph.

Conclusions: We applied Brownie to both simulated and real data set and we showed that Brownie outperforms all previous methods in terms of accuracy, memory efficiency.

68. Elucidating effects of HAART on molecular evolutionary patterns of HIV-1 in countries with different socio-economic status.

Reeba Paul, Madara Hetti Arachchilage, Helen Piontkivska
Kent State University

Background: Significant difference in molecular evolutionary patterns of HIV-1 exists between developed and developing countries. While the umbrella of socio-economic differences between countries likely encompasses a multitude of individual factors - such as availability of baby formula for HIV-positive women or cultural habits, among others - whose contribution may be impossible to decipher at this stage, one of the major factors is HAART.

Description: Isolating the influence of HAART on evolution of the virus in individual countries is challenging, in part because of wide variation in both coverage and compliance between developed and developing countries. On one hand, we can expect a larger share of drug resistant mutations being present in developed countries due to longer exposure to drugs and/or better coverage. On the other hand, in developing countries the treatment may not always be affordable; hence, intermittent HAART use may also lead to emergence of drug resistance. In this study we have attempted to elucidate the effect of HAART on substitution rates in HIV. Conclusion: Our results showed that indeed a significant difference in substitution rates exists for HIV-1 Pol gene sequences sampled in pre- and post-HAART periods in different groups of countries, based on income and stability, although trends appear to differ between different groups of countries. In general, it appeared that sites harboring drug resistance mutations have higher non-synonymous substitution rates than the remainder of the gene in high income countries, as could be expected, while that difference is less pronounced in lower-income countries.

69. Automatic correction of the nucleotide sequences

Zakaria Elyazghi, Ahmed Moussa, Fouzia Radouani

Institut Pasteur du Maroc, Casablanca, Morocco
LabTIC Laboratory - ENSAT, Tangier, Morocco

The computer output for a sequencing run consists of chromatogram in ABI Format. When viewing chromatograms, there is some ambiguities at various sites along the DNA sequence, because the sequencing machine used to call the bases cannot always precisely determine what nucleotide is, when it is represented by either a broad peak or a set of overlaying peaks. In such cases, a letter other than A, C, G, or T is recorded, most commonly "N". The purpose of this work is to develop an application allowing the automatic correction of these ambiguities. This program run under R platform and consist in the development of friendly user interface for an easy exploitation of results. For tests, we used sequences of bacterial strains detected in urogenital samples of patients with urogenital infections. As results, we note that our program, ABI Base Recaller, can give a good correction, very close to the manual one, it increases the rate of identity and coverage and minimizes the number of mis-matches and gaps. Thus, it provides solution to this problem and save biologist's time and labor.

Category : Systems Biology and Networks

70. Chromatin interactions predict co-expression in the mouse cortex

Ahmed Mahfouz, Sepideh Babaei, Marc Hulsmann, Boudewijn P.F. Lelieveldt, Jeroen de Ridder, Marcel Reinders

Leiden University Medical Center

The three-dimensional conformation of the genome in the cell nucleus influences important biological processes such as gene expression regulation. Previous studies have shown a strong correlation between co-expression and chromatin interaction in model organisms and human cell lines. However, predicting gene co-expression from direct chromatin interactions remains challenging. We present the first attempt to predict co-expression based on chromatin interactions. We address this by representing chromatin interactions, measured by Hi-C, as a network which allows a more comprehensive characterization of long-range interactions using network topology. We demonstrate the power of using scale-aware topological measures to describe chromatin interactions at different scales ranging from direct interaction between genes (small-scale) to chromatin compartment interactions (large-scale). Our results show that the spatial co-expression of genes in the mouse cortex, based on data from Allen Mouse Brain Atlas, is predictable from chromatin contact information as measured by Hi-C data. However, the accuracy of prediction can be substantially improved by using scale-aware topological measures to describe chromatin interactions as well as considering Hi-C interactions at different genomic resolutions. These results point to the importance of not only direct chromatin interactions between genes (small-scale) but also chromatin compartment interactions (large-scale) in explaining co-expression.

71. Simulation of early mouse ovarian development using a cellular potts model

Annika Eriksson, Hannah Wear, Karen Watanabe

Oregon Health and Science University

Computational models offer a humane, efficient alternative to reduce the number of animals used in traditional toxicity testing by simulating biological processes to predict whole organism responses. Many endocrine-disrupting chemicals negatively affect reproductive function. To predict adverse effects we first need to understand and simulate normal reproductive system development and function. Our research focuses on in silico simulation of multi-scale, multi-cell phenomena during normal early ovarian development in mouse, one of the most common animals used in toxicity testing. With CompuCell3D software, we constructed a two-part model using published data on normal cell-cell behaviors such as adhesion, growth, apoptosis, mitosis, secretion, and migration. CompuCell3D employs a lattice-based Cellular Potts Model, a probabilistic Monte Carlo step approach that modifies cells in the lattice based on whether an overall effective energy function is minimized. We used an iterative process to parameterize this function by tuning cellular constraints such as volume, surface area, chemical fields, and cell-cell adhesion energies. Part One of our model simulates migration of primordial germ cells into the gonadal ridge, and Part Two simulates proliferation of germ cells in the gonadal ridge and development of primordial germ nests and follicles. When visually compared with experimental observations and histological images, both parts of our model reproduce the data well. This model can be expanded to extrapolate in vitro data as a predictive toxicology tool. It can be used for educational purposes, the generation of new hypotheses, and to demonstrate the potential for modeling whole organ systems using CompuCell3D.

72. Modeling the regulation of β -catenin by Wnt stimulation and GSK3 inhibition

Annika Jacobsen, Nika Heijmans, Renée van Amerongen, Folkert Verkaar, Martine J. Smit, Jaap Heringa, K. Anton Feenstra
Centre for Integrative Bioinformatics (IBIVU), VU University Amsterdam, The Netherlands

Background: Wnt/ β -catenin signaling is crucial for cell renewal and differentiation. Aberrant signaling caused by specific mutations plays an important role in oncogenesis. A better understanding of these signaling mechanisms is therefore crucial. We have constructed a Petri net model of Wnt/ β -catenin signaling that captures the regulation mechanisms of the transcriptional co-activator β -catenin. Main components included are the Wnt receptors, the destruction complex and three of its crucial components AXIN, AXIN2 and GSK3. We included an important feedback loop of upregulated AXIN2 expression by β -catenin.

Description: We simulated the model with Wnt stimulation and GSK3 inhibition i.e. aberrant signaling. Wnt stimulation lead to an initial peak in β -catenin levels followed by a decrease to initial levels. On the other hand, inhibitions by GSK3 lead to high increases of β -catenin levels. We experimentally validated these observations by western blot. In addition we measured the transcriptional activity of AXIN2 by TCF/LEF luciferase reporter assay, which showed both dosage and time dependent transcriptional activity for both Wnt stimulation and GSK3 inhibition.

Conclusions: The feedback from AXIN2 only has a negative effect on β -catenin during Wnt stimulation, where AXIN is the limiting factor, and not during aberrant signaling. Further, our model, together with validating experiments, suggest that the initial β -catenin peak seen during Wnt stimulation is sufficient for transcriptional activation of AXIN2. Using this model we predicted the β -catenin behavior from other important mutations found in breast and colorectal cancer. In summary, our model can be used to explain plau-

sible underlying mechanisms for oncogenic signaling.

73. Targeted destruction of receptors of cancer cells by porphyrins

Aram Gyulkhandanyan, Grigor Gyulkhandanyan

Institute of Biochemistry of NAS of Armenia

Background: The epidermal growth factor receptor (EGFR) is transmembrane protein and its overexpression affects on the state of a cell and leads to tumor growth. Upon binding of the natural peptide ligands to domains I and III of the extracellular domain of EGFR occurs a conformational rearrangements leading to dimerization of intracellular domains of receptors, which converts cells in oncological state. Together with scientists from the University of Nantes we have shown that some small compounds (non-peptide compound nitro-benzoxadiazolyl (NBD)) may purposefully bind to dimerization domain EGFR, which promotes the formation of stable dimers and launching of oncological process.

Description: By molecular docking method, we showed that NBD have high affinity to different sites of EGFR, including to domains I and III. It has been shown that the highest affinity NBD showing to a site between two macromolecules of the extracellular domain of EGFR. On the other hand by the method of molecular docking we also showed high affinity of EGF and NBD with cationic porphyrins and formation of complex systems [NBD + porphyrin] and [EGF + porphyrin]. Porphyrins accumulate selectively in tumors and upon illumination promote generating of reactive oxygen species that result to the destruction of cancer cells.

Conclusions: It allows assuming that the complex type [NBD + porphyrin] or [EGF + porphyrin] at affinity binding with the extracellular dimerization domains of EGFR and by photodynamic illumination, the reactive oxygen species can cause destruction of EGFR domains, prevent the dimerization process and cancer launch.

74. Multiscale mathematical modelling recapitulates breast cancer invasion phenotypes

Arnau Montagud, Andrei Zinovyev, Emmanuel Barillot

Institut Curie

Understanding tumour invasion mechanisms is crucial to improve prognosis and develop new cancer treatment strategies, but this is hindered by the lack of understanding of detailed molecular determinants of this process and their interactions leading to different ways cancer cells invade the surrounding tissues. Tumour invasion varies from individual to collective cell movement or if migrating cells are mesenchymal-or amoeboid-like or also if they use proteases to facilitate their migration. In the past years several efforts have been done in systematising different mechanisms of cell migration and understanding their underlying causes. We devised a multi-scale mathematical model that incorporates information of a series of traits, cellular and environmental, that output in a set of invasion modes. For this, the model incorporates different intracellular and signalling pathways and the resulting influence network has been translated into a mathematical model using discrete logical modelling. We have taken advantage of continuous time Boolean modelling based on Markovian stochastic process defined on the model state transition graph to simulate intracellular molecular processes determining individual cellular properties. We have embedded this Boolean model in a lattice-free individual cell population model to cope with interaction between cells and microenvironment affecting cell properties, leading to various patterns of collective cell behaviour. The model has been tuned by observed phenotypes on existing data from experimental results on tumours, cell lines and cells spheroids. Present work is part of a collaborative effort to model tumour invasion in order to identify treatment strategies and to understand underlying properties of metastasis.

75. A Probabilistic Graphical Model for Interleukin-1 Signaling in Cancer

Aurora Blucher, Anupriya Agarwal, Jeffrey Tyner, Shannon McWeeney, Guanming Wu
Oregon Health & Science University

The interleukin-1 (IL-1) family of cytokines regulates both innate and adaptive immunity by controlling proinflammatory reactions. Upregulation of IL-1 has been found in several types of tumors and is often associated with a poor prognosis for the patient. Probabilistic graphical models provide a concise way of representing a complex distribution of conditional probabilities across many cross-dependent variables, such as those found in biological pathways and networks. These models have successfully been applied in the study of biological systems, such as PARADIGM, which determines patient-specific pathway activity in cancer genomic samples. Using an IL-1 signaling network compiled from the literature in work by Ryll, et al., we have developed a probabilistic graphical model for the IL-1 pathway. By using this model, we can integrate several different types of genomic data, including gene expression and copy number variations (CNVs), from tumor samples along with drug sensitivity information and explore how IL-1 pathway activation differs among samples. This work will help to further elucidate the role of IL-1 signaling in tumor development and sensitivity to chemotherapeutic drugs.

76. The Discordant Method: A Novel Application for Detecting Differential Correlation

Charlotte Siska, Katerina Kechris
University of Colorado Anschutz Medical Campus

A common approach for identifying molecular features (such as transcripts or proteins) associated with disease is testing for differential expression or abundance in -omics data. However, this approach is limited for studying interactions between molecular features, which would give a deeper knowledge of the

relevant molecular systems and pathways. We have developed a method for this purpose that we call the Discordant method. The Discordant method estimates the posterior probability that a pair of features has discordant correlation between phenotypic groups using mixture models and the EM algorithm. We compare our method to existing approaches; one that uses Fisher's transformation in a classical frequentist framework and another that uses an Empirical Bayes joint probability model. We prove with simulations and miRNA-mRNA glioblastoma multiforme (GBM) data from the Cancer Genome Atlas that the Discordant method performs better in predicting related feature pairs. In simulations we demonstrate that while all of the methods have similar specificity, the Discordant method has better sensitivity and is better at identifying pairs that have a correlation coefficient close to 0 in one group and a largely positive or negative correlation coefficient in the other group. Using the GBM data, which has matched samples between miRNA and mRNA, we find that the Discordant method finds relatively more GBM-related miRNAs compared to other methods. We conclude from the results in both simulations and GBM data that the Discordant method is more appropriate for identifying molecular feature interactions unique to a phenotype.

77. Harnessing a large collection of sparse and noisy gene perturbation data to discover mammalian causal gene regulatory networks

Djordje Djordjevic, Andrian Yang, Shu Lun Shannon Kwan, Joshua W. K. Ho
Victor Chang Cardiac Research Institute

Background: Gene regulatory networks (GRNs) play a central role in systems biology. Recent findings, including ours on mammalian causal GRNs, showed that it is virtually impossible to infer causal GRNs in eukaryotes without using gene perturbation data. A huge amount of perturbation data is publicly available but computationally inaccessible, buried in the figures and tables

of published papers, or as data hosted on gene expression study databases.

Description: Manually curating over 6,000 experimental results from genetic or molecular perturbation data enabled us to infer >3,000 causal gene regulatory interactions among >1,000 genes across multiple tissues during embryonic development. This approach has enabled us to uncover biologically useful causal GRNs for multiple organs. We present web-based resources for early tooth (<http://compbio.med.harvard.edu/ToothCODE/>), ocular lens (<http://isyte.victorchang.edu.au/>), and heart development (<http://CardiacCode.victorchang.edu.au/>), and associated diseases. In order to harness the vast amount of data in NCBI's GEO, we are developing an automated pipeline to (i) identify genetic perturbation studies, (ii) assign phenotypic annotation to each sample, and (iii) extract differentially expressed genes from each perturbation dataset. For analyzing processed perturbation data, we present a probabilistic model that can combine diverse and noisy data to infer cell-type or other context specific causal GRNs. Finally, we present a network-based algorithm to identify the minimum set of upstream regulators spanning a candidate gene set at multiple levels, with applications for de-convoluting phenotypic drivers.

Conclusions: Our research is facilitating the automated construction, analysis and interpretation of eukaryotic GRNs, with broad applicability for rapid insight generation in systems biology

78. A Synthetic Data Generator for Kinome Microarray Data

Farhad Maleki, Anthony Kusalik
University of Saskatchewan

In kinome microarray data analysis developing techniques to infer signaling pathways based on the data is a key research area. Synthetic data allows us to construct artificial datasets with a priori knowledge about the signaling pathways represented. Therefore,

synthetic kinome data generators would be an indispensable tool to evaluate various algorithms and methodologies for signaling pathway activity inference. To the best of our knowledge, there is no synthetic data generator designed for kinome microarrays. Further, synthetic data generators for DNA microarray technology may not be appropriate for synthesizing kinome array data. In this paper we propose a method to generate synthetic kinome array data. The proposed method relies on actual measurements from kinome microarray experiments to preserve subtle characteristics of the original information and to prevent oversimplification of the generated data, which is a main concern with synthetic data generators. The proposed method includes two algorithms, one for generating inter-array technical replicates and the other for differentially phosphorylating an arbitrary set of peptides. We use fold change thresholds and one-sample t-test with a significant level to determine differential phosphorylation. We show that the measurements for within-array technical replicates in the synthesized data have the same distribution as actual data from kinome array experiments. In addition, we validate the synthetic kinome array datasets using PIIKA 2, which is a well-established tool for analysis of kinome microarray data. The analysis by PIIKA 2 reveals the same set of differentially phosphorylated peptides as those used for synthesizing kinome array data.

79. A network based approach to understanding altered virulence in closely related Mycobacterium tuberculosis isolates.

Jon Mitchell Ambler, Margaretha de Vos,
Suereta Fortuin, Melanie Grobbelaar, Rob Warren, Samantha Sampson, Nicola Mulder
University of Cape Town

Mycobacterium tuberculosis (MTB) is the causative agent of the severe respiratory disease tuberculosis, affecting millions of people globally and resulting in the deaths of ~1.5 million per year. As both multi-drug resistant and extensively drug resistant strains con-

tinue to emerge, this disease is likely to become an increasing threat. In the Western Cape, South Africa, two strains of MTB showing varying levels of virulence are being used to investigate the biological underpinnings of virulence in this pathogen. Understanding complex phenotypic traits such as virulence requires the integration and interpretation of multiple layers of data from various sources. To this end, biological networks provide an intuitive framework in which the factors influencing these traits may be investigated. We have assembled the genomes of the two isolates of interest, and identified single nucleotide polymorphism (SNP) containing genes. At the gene regulation level, the non-coding RNA sequences predicted in previous studies were screened for SNPs and their potential targets predicted. These data layers are integrated into a larger biological network for MTB containing both experimentally generated data and data obtained from public repositories, including transcription factor over-expression experiments, gene set enrichment analysis, non-coding RNA regulation, and gene expression data. By integrating these various datasets, we are able to generate insights that will guide us in eliciting the cause of the dissimilar virulence of the two isolates, as well as developing additional frameworks for the integration of data from multiple sources that better represent their biological contexts.

80. Computational approaches to identify and dissect the transcriptional influence on metabolism

Kevin Schwahn, Zoran Nikoloski

Max Planck Institute of Molecular Plant Physiology

Background: The availability of high-throughput data from transcriptomics and metabolomics technologies necessitate novel statistic approaches to elucidate the transcriptional influences on metabolism.

Description: Here we introduce two new approaches to identify transcriptional effects on metabolite levels: The first combines partial

correlations with principal component analysis, while the second partials out the covariance of transcript expression levels from the covariance of metabolite levels. Based on these approaches, we also defined and investigated three new concepts-stable correlations, noise-sensitive correlations, and total partialing.

Conclusion: Our findings demonstrate that the proposed approaches are effective in pinpointing the metabolite pathways under strong transcriptional influence. The proposed approaches can also be readily employed to extract network-based descriptions of the data sets, which we use in our subsequent enrichment analysis. Using transcriptomic and metabolomic profiles from *Escherichia coli* under five different environments, we show that the so-extracted networks contain a smaller number of three-cycles, in comparison to correlation-based networks; however, the findings from the enrichment analysis remain unaltered. Therefore, the proposed approaches provide a promising extension to widely used techniques in computational systems biology to dissect relationships between components on different levels of cellular organization.

81. The TOPCONS webserver for consensus prediction of membrane protein topology and signal peptides

Konstantinos Tsirigos, Christoph Peters, Nanjiang Shu, Lukas Käll, Arne Elofsson
Stockholm University

The TOPCONS web server was launched in 2009 and it combined predictions from five individual topology prediction methods using simple grammar architecture in order to create a consensus prediction. We hereby present a major update to the server, with some substantial improvements, including: (i) TOPCONS can now efficiently separate signal peptides from transmembrane regions. (ii) The server can now differentiate more efficiently between globular and membrane proteins. (iii) The server now is even slightly faster, although a much larger database is

used to generate the multiple sequence alignments. For most proteins the final prediction is produced in a matter of seconds. (iv) The user-friendly interface is retained, with the additional feature of submitting batch files and accessing the server programmatically using standard interfaces, making it thus ideal for genome-wide analyses. Indicatively, the user can now scan the entire human proteome roughly in a day's time. (v) For proteins with homology to a known 3D structure, the homology-inferred topology is also displayed. (vi) Finally, the combination of methods currently implemented achieves an overall increase in performance by 4% as compared to the currently available best-scoring methods and TOPCONS is the only method that can identify signal peptides and still maintain a state-of-the-art performance in topology predictions. The web server is freely available at <http://topcons.net/>.

82. Genome-wide ceRNA networks

Mario Antonio Flores
University of Texas

Postranscriptional regulation of gene expression can be modeled as a competitive endogenous RNA (ceRNA) network in which mRNAs compete for miRs binding. Previous research shows that this competition maintains and fine-tunes levels of protein coding genes and the disruption of the network contributes to phenotypic conditions like cancer. Based on our previous studies we provided a tool (TraceRNA) for reconstruction of ceRNA networks around a gene of interest (GoI). The approach used in TraceRNA although practical and useful for gene-based studies provides only a partial landscape of the ceRNA mechanisms and phenotypes. Besides TraceRNA offers an ad-hoc approach for the study of the ceRNA phenomenon. In this work we present a formal genome-wide approach for ceRNA networks study. This novel and formal treatment of the ceRNA phenomenon provides new perspectives in the study of ceRNA networks and its specific phenotype. We divide the study of genome-

wide ceRNA networks in three main sections: network construction, analysis of network components by network perturbation and network stability. In the case of network construction we formalize the definition of ceRNA phenomenon. The construction of a genome-wide network of a specific phenotype (e.g. breast cancer) is modeled having as input datasets of miR binding predictions as well as associated mRNA and miRNA expression datasets. In the case of the analysis of the main components (mRNAs, miRNAs) of the predicted ceRNA network we present an algorithm that is based on perturbation. For ceRNA network stability we present an approach using data subsampling and network stability indicators. This research was supported in part by the Intramural Research Program of the National Library of Medicine, NIH.

83. Inferring Parameters of Cis-Regulation Using Parallelized Optimization

Michal R. Grzadkowski
Massachusetts Institute of Technology

Background: The properties of the interactions between cis-regulation and trans-regulation remain poorly understood. It is well understood that the nuances of these interactions drive many fundamental biological processes, including tissue differentiation, evolution, and disease development. However, attempts to describe the machinery of gene regulation have so far sacrificed depth for breadth, or vice versa, and have not fully exploited the computational resources available to many researchers today.

Description: We propose an ensemble of thermodynamic functions that model how the placement of cis-regulatory elements (CREs) and putative trans-factor (TF) binding sites in the proximity of target genes affects their transcription. To optimize the parameters of this model, we implement a parallelized algorithm that alternates between simulated annealing and genetic selection steps to find the best fit of the thermodynamic functions to a set of co-expression values between TFs

and their targets. This model is implemented on an available compute cluster with up to two hundred nodes running in parallel.

Conclusions: Our model was able to uncover several interesting patterns of gene regulation in a myelogenous leukemia cell line (K562). In particular, different families of transcription factors exhibited unique effects on gene regulation mediated by enhancers and promoters depending on motif placement relative to gene TSS position and strand.

84. RNA secondary structure prediction with pseudoknots using Newtonian dynamics simulations

Nils Petersen, Andrew Torda
University Hamburg

Background: One of the major problems of secondary structure prediction from RNA sequences is that most algorithms are limited to nested structures without pseudoknots. This is due to both computational complexity and limitations of the existing energy models.

Description: We have developed a new method to predict RNA pseudoknots using Newtonian dynamics in an artificial base pair space. This required converting parts of the popular, discrete nearest neighbor energy model into continuous force field terms and simulating non-physical particles (base-pair probabilities) in a one-dimensional space. This model imposes no restrictions on the topology of the structure and thus allows all kinds of pseudoknots. Furthermore, we demonstrate how it can be coupled to three-dimensional structure models and how secondary and tertiary structure representations can be simulated at the same time.

Conclusions: On its own, the model is very simple and brings no improvement compared to other methods. However, it has the major benefit that it can be coupled to a 3D structure model. We will exploit this in the future. Adding a very simple coarse grained model will allow us to predict secondary structures which are physically plausible given the steric constraints. Conversely, we want to use the secondary structure model to bias more

detailed simulations in order to enhance the sampling of 3D structures.

85. Application of a graph-based technique to plasmodium falciparum interactome

Odia Trust Osee, Dr. Ola Oyelade
Covenant University

Plasmodium falciparum is the parasite that causes malaria disease, which according to (W.H.O, 2014) there were 198 million malaria cases and an estimated 584,000 malaria deaths. Many researchers have worked on this disease and have looked at its proteome and interactome, in a bit to find solution to the disease. Proteins are essential parts of organisms and participate in virtually every process within cells. Each protein in the cell of a living organism has a single or multiple functions. Proteins can also work together to achieve a particular function and they often associate to form stable protein complexes. Understanding protein interactions is helpful in deducing its function which plays a great role in drug design, sheds light in diseases, protein structure prediction, understanding molecular biology, molecular mechanism of cellular processes, cell biochemistry and many others. The study seeks to provide functional annotation for hypothetical proteins in the parasites interactome. It employs a graph-based clustering algorithm called Molecular Complex Detection (MCODE) which identifies protein complexes in the parasites interactome. Functional annotation for the identified proteins in the complexes where done by bioinformatics approaches which includes sequence similarity, protein structure alignment and assessment, phylogenetic analysis and protein family classification. Functional annotation was assigned to two hypothetical proteins (PFL0350c and PFL1395c) based on the functional annotation approaches used. PFL1395c seem to be a histone acetyltransferase enzyme that regulates gene expression. Keywords: protein-protein interaction, clustering, protein function, molecular complex.

86. Detection of Heterogeneity in Single Particle Tracking Trajectories

Paddy Slator, Nigel Burroughs
University of Warwick

Background: Single particle tracking trajectories are fundamentally stochastic, which makes the extraction of robust biological conclusions difficult. This is especially the case when trying to detect heterogeneous movement of molecules in the plasma membrane. This heterogeneity could be due to a number of biophysical processes such as: receptor clustering, traversing lipid rafts, binding to the cytoskeleton, or changes in membrane diffusivity.

Description: Working in a Bayesian framework, we developed multiple models for heterogeneity, such as confinement in a harmonic potential well, and switching between diffusion coefficients. We analyse these models using Markov chain Monte Carlo algorithms, which infer model parameters and hidden states from single trajectories. We also calculate model selection statistics, to determine the most likely model given the trajectory. Our methodology also accounts for measurement noise. For LFA-1 diffusing on T cells we found 10-26% of trajectories display clear switching between diffusive states, depending on treatment. Analysis of the motion of GM1 lipids bound to the cholera toxin B subunit in model membranes showed transient trapping in harmonic potential wells. We have also demonstrated that allowing for measurement noise is essential, as otherwise false detection of heterogeneity may be observed.

Conclusions: We have used Bayesian methodology to analyse single particle tracking trajectories. Rather than existing methods, which rely on generic properties of Brownian motions, our approach allows us to test which biophysical model best fits a trajectory. With the continuing improvement in spatial and temporal resolution of trajectories, these methods will be important for biological interpretation of experiments.

87. Unravelling signal regulation from large scale phosphorylation kinetic data

Westa Domanova, James Krycer, Rima Chaudhuri, Fatemeh Vafaei, Daniel Fazakerley, David James, Zdenka Kuncic
University of Sydney

Background: A key biological paradigm is that biological processes are tightly-controlled by the temporal behavior of cellular signalling events. Phosphorylation, a prevalent means of signalling, occurs with rapid dynamics but for the majority of events the kinase is unknown. To elucidate the underlying topology of signalling cascades from high-throughput data, we need to be able to predict kinase substrate relationships (KSRs). Existing prediction algorithms do not consider the crucial biological context of KSRs.

Description: Here, we predict KSRs in a data-dependent and automatised fashion: given that some kinases are active before others, we use computationally determined kinase specific temporal patterns to predict site-specific KSRs from large-scale in vivo experiments (ssKSR-LIVE).

Conclusions: Applying this to insulin-stimulated phosphoproteomic data we distinguished between AKT and RPS6KB1, two kinases sharing the same consensus motif. We identified several kinases and predicted novel substrates, correlating this with key insulin-regulated biological processes. As a flexible algorithm ssKSR-LIVE can be applied to other high-throughput signaling data and thus can improve our understanding of complex diseases caused by dysregulated signalling, including cancer and type 2 diabetes.

88. An integrative approach to unravel the human-Schistosoma mansoni interactome: Who, when and where

Yesid Cuesta-Astroz, Alberto Santos, Lars Juhl Jensen, Guilherme Oliveira
Fiocruz/MG

Background: The study of molecular host-parasite interactions is essential to understand parasite infection and local adaptation within the host. Efforts use several strategies to identify inter-species protein-protein interactions (PPIs) between the host and parasites, viruses and bacteria. Here, we investigate the inferred PPI network between human and *S. mansoni*, one of the parasites causing Schistosomiasis, a neglected tropical disease. **Description:** To this end, we propose an integrative approach that gives context to the interactions according to the parasite's life cycle and subcellular localization of the proteins. We use a homology-based method to predict interactions by looking at intra-species interactions among all organisms within the closest ancestral group common to both, human and *S. mansoni* and uses conservation of interactions as a measure of confidence. Besides, we used publicly available datasets of domain-domain interactions to identify possible PPIs based on common domains. To contextualize the interactions, we limit the interactions to human membrane expressed in tissues that support the parasite's tropism (skin, blood, lung, liver and intestine). Our approach predicted 34,586 PPIs, which show crosstalk between parasite and host proteins enriched in metabolic and tissue-specific secretory pathways essential in the life cycle of the parasite. An initial manual curation of some of the interactions revealed tissue-specific interactions that are also stage-specific according to expression data available for *S. mansoni*.

Conclusions: We believe that applying this systems biology approach will certainly help uncover targetable mechanisms for the therapy of Schistosomiasis, and also opens the possibility for the analyses of any host-parasite pair.

Category : Other - Not Elsewhere Classified

89. Understanding leishmania development and drug resistance using an integrative omics compendium

Bart Cuypers, Pieter Meysman, Manu Vanaerschot, Maya Berg, Jean-Claude Du Jardin, Kris Laukens
University of Antwerp, Antwerp, Belgium

Leishmania donovani causes visceral leishmaniasis (VL), a disease which is lethal without treatment. With only four drugs available and rapidly emerging drug resistance, knowledge about the parasite's resistance mechanisms is essential to boost the development of new drugs. However, only little is known about *Leishmania*'s gene regulation and the few findings indicate major differences to known gene expression systems. Integration of different 'omics could shed light on these gene regulatory mechanisms, but there has been little integration effort so far. Therefore, we developed an easy to use tool, able to collect and connect all the existing *L. donovani* 'omics experiments. Genomics, epigenomics, transcriptomics, proteomics, metabolomics and phenotypic data was collected and added to a MySQL database compendium, further complemented with publicly available data. Relations between the different 'omics levels were explicitly defined and provided with a level of confidence. Python scripts were developed to preprocess, import and access the data. Next to this vast data source a set of integrative data-analysis tools was developed based on data mining strategies. For example: One tool uses frequent pattern mining algorithms to look which proteins and metabolites frequently behave in the same way under different conditions. Another tool converts several 'omics data to a network format that can be opened in Cytoscape and can thus be the basis for network analysis. Using the compendium, we characterized the development and drug-resistance in a system biology context (all 'omics). The compendium and its scripts

could be used for other organisms with only minor changes.

90. Development of a chromosomal visualization tool for genomic data

Karen Yasmine Oróstica Tapia, Ricardo A. Verdugo

Universidad de Chile

In the last years, the new technologies for high-throughput sequencing allowed a considerable reduction of time and monetary cost associated with the sequencing of complete genomes. This development has generated a large volume of data that typically goes through a review process where visualization plays an essential role. Currently there are several programs and browsers that can represent genomic features in chromosomal context such as Cvit, PhenoGram, GViewer, GBrowse, among others. However, most of these programs are external to R, a statistical software that constituted the most used working environment for the analysis of this type of information. Based in the previous and in order to facilitate the workflow in genome projects, the goal of this project was to create a package in R able to graph elements such as genes, variants and regulators along the chromosome elements. The package was developed from previously used codes in the Laboratory of Genetics Systems and Biomedical Genomics (GENOMED Lab). These codes were cleared, complemented and finally packaged in a program called chromPlot. It was also developed a package with cytogenetic data and gaps of human and mouse assembly, for future users that wants to easily test the package. Finally, it was created a tutorial that shows different usage examples, including graphs of differential expression, gene density, genetic polymorphisms and synteny between human and mouse. Notably chromPlot may interact with other R packages for data from public repositories and monitoring of small regions with nucleotide resolution.



Swiss Institute of
Bioinformatics

Extraordinary Science Swiss Quality A Unique Education

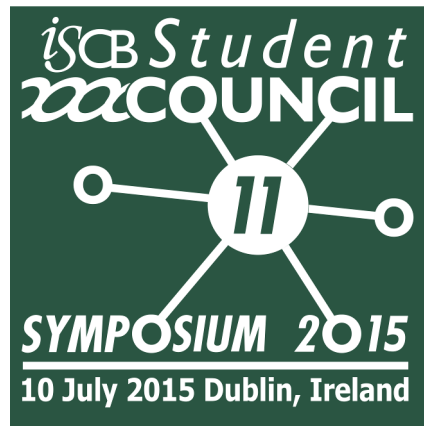


Think Switzerland when you think education in bioinformatics.
Swiss Universities, ETHZ, EPFL and SIB working together to offer Bachelors,
Masters, PhD degrees and training workshops.

Where else?

www.isb-sib.ch/training





Travel Fellowships & Awards

Blast from the past



SCS2013's Travel Fellowship Awardees with SCS2013 Organisers and Prof. Bukhard Rost, ISCB President (2013)

Travel Fellowships

This year, the ISCB Student Council is awarding six travel fellowships to attendees of the 11th ISCB Student Council Symposium taking place in Dublin, Ireland. These include, for the first time in the Symposium's history, an *"Inspiring Youth"* travel fellowship awarded to one **pre-university** student attending.

Winners of SCS2015 travel fellowships are:

SIB Travel Fellowship sponsored by Swiss Institute of Bioinformatics

- Alexander Monzon (Argentina),
- Marek Cmero (Australia),

F1000 Travel Fellowship sponsored by Faculty of 1000 Ltd

- Giulia Fiscon (Italy),

BMC Travel Fellowship sponsored by BioMed Central Limited

- Westa Domanova (Australia),
- GÜngör Budak (Turkey).

ISCB SC *"Inspiring Youth"* Travel Fellowship

- Prathik Naidu (United States of America)

Travel Fellowship Sponsors



Swiss Institute of
Bioinformatics



BioMed Central
The Open Access Publisher

F1000Research

OPEN SCIENCE • OPEN DATA • OPEN PEER REVIEW

ISCB Student
COUNCIL
INTERNATIONAL SOCIETY FOR COMPUTATIONAL BIOLOGY

Awards

The outstanding poster and oral presentations of the 11th ISCB Student Council Symposium will be recognized and awarded. These awards have been made possible by the sponsorship of Oxford University Press. We thank Oxford University Press for their continued support.



Best Presentation Award

This award will acknowledge the best oral presentation by an author 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 and peers. The Best Presentation Award this year is sponsored by Oxford University Press journal Bioinformatics.

Value of the award: 500 USD

Previous winners:

2014 – Harriet Dashnow

2013 – Han Cheng Lie and Severine Affeldt (Runner-Up)

Best Poster Award

The best poster of the symposium will be chosen through voting by the delegates present on the day and by the opinions of a jury of ISCB Student Council Leaders and peers. 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

Previous winners:

2014 – Alex Salazar and Sarah Keasey (Runner-Up)

2013 – Maribel Hernandez-Rosales and Nadezda Kruchkova (Runner-Up)

Awards Ceremony

All awards will be presented at the awards ceremony, to be held together with the ISCB Open Business Meeting on Monday, July 13, 12:45 p.m. - 1:50 p.m.- Room *The Forum*, at the Convention Center Dublin, Ireland.



The Student Council and its Activities

Blast from the past



**A photo from Open Business Meeting
at ISMB2012, at Long Beach, California, USA**



The Organization

The Student Council (SC) is a vibrant international student group made of members from around the world who share a passion for bioinformatics and computational biology. The SC is organized into teams and committees, each with their own tasks and responsibilities, designed to provide students with support and resources. The SC is part of the ISCB and as such it works closely with the ISCB leadership to achieve its goals.

Mission Statement

The mission of the ISCB Student Council is to promote the development of the next generation of computational biologists. We achieve our goal through provision of scientific events, networking opportunities, soft-skills training, educational resources and career advice, while attempting to influence policy processes affecting science and education.

Student Council Executive Team

The positions in the Student Council Executive Team (ET) are filled by the elected leaders. The key role of the ET is the sustainable development of the Student Council and its Regional Student Groups (RSGs) and the coordination of all activities. For this purpose several committees are established that are chaired and occupied by SC members. The SCET is responsible for managing these committees.

Events

The events teams mainly focus on events coordinated with ISCB-related events (ISMB, ECCB, ASBCB, ISCB-LA), such as the Student Council Symposium (SCS), European Student Council Symposium (ESCS), Student Council Symposium - Latin America (LA-SCS) and Student Council Symposium - Africa (SCS Africa) and we now hope to take up the challenge and increase the SC presence further.

Committees

The standing committees focus on sustaining the SC and its RSGs. These include the Education & Internship, Fundraising, Outreach, Publication, Web and RSG committees.

Our activities over the last year

This was the first year that we hosted four independent student-run symposia, each in conjunction with one of ISCBs affiliated conferences.

- The **10th Student Council Symposium** was held during the first workshop day of ISMB 2014 in Boston, USA. The chairs Farzana Rahman and Tomás Di Domenico organized a meeting that hosted 12 student talks and two wonderful keynotes by Dr. David Bartel and Dr. Ashlee Earl.
- The **3rd European Student Council Symposium** was held before ECCB 2014 in Strasbourg, France. The meeting hosted 8 full talks and keynotes by Dr. Lennart Martens, Dr. Jeroen de Ridder and Dr. Lars Juhl Jensen. ESCS 2014 was organized by chairs Pieter Meysman and Margherita Francescatto.
- The **1st Latin America Student Council Symposium** is the first of two new symposia hosted by the student council. It was held in the day before ISCB-LA in Belo Horizonte, Brazil. The chairs R. Gonzalo Parra and Avinash Shanmugam along with their committees put together a meeting that saw keynotes from Dr. Vitor Leite, Dr. Francisco Melo and Dr. Peter F. Stadler as well as 6 stellar student presentations.
- The second new symposium this year was the **1st Student Council Symposium Africa** held in the days preceding ISCB-Africa in Dar es Salaam, Tanzania. Dr. Manuel Corpas gave an excellent keynote talk along with 7 student presentations. The chairs for this meeting were Yassine Souilmi and Chinmay Kumar Dwibedi.

In addition to our flagship symposium series, the student council's executive team and various committees are hard at work to advocate for and provide resources to the student researchers in the ISCB. We would like to thank all of the ISCB Student Council committee members for their help all year long, without your assistance the SC would not be able to function as well as it does. In particular we would like to thank the committee chairs for their continued contribution.

- **Education/Internship** — Emre Guney
- **Fundraising** — Jakob Jespersen
- **Outreach/Volunteer** — Alexander Junge
- **Regional Student Group** — Chinmay Dwibedi
- **Web** — Dan DeBlasio and Mehedi Hassan

The student council recently facilitated the recruitment of it's ninth intern to the SC Internship Program. The **education and internship committee** worked very hard to collect and review applications to a position in the Schneider lab at the Luxembourg Centre for Systems Biomedicine (LCSB). Be on the lookout for additional internships in the coming year!

The **outreach committee** has been hard at work this year increasing the Student Council's presence on social media. We encourage everyone to follow **@ISCBSC** on twitter, facebook and linkedin to find out new and exciting information about goings on around the student council as well as new and interesting research.

This year also saw a slight reorganization of the **regional student groups** with the addition of regional vice-chairs: Yassine Soulimi (Africa), Gonzalo Parra (Latin America) and Marek Cmero (Asia). The RSG program experienced a rapid expansion with the addition of 8 new RSGs this year, taking the total number of active RSGs to 29. The new RSGs that joined the RSG network in the year 2014-15 include: RSG Chile, RSG Ireland, RSG Brazil, RSG Norway, RSG Denmark, RSG Sri Lanka, RSG South Africa and RSG Washington DC. The RSG Committee is also working closely with the web committee to create a unique online presence for each group.

Several regional student groups have organized their own events this year as well and we are excited that they turned out so well.

The **web committee** has recently released a new version of the student council website that is easier to navigate and more up to date. In doing so they are working on new features that will help the SC function more smoothly as well as create more harmonious communication and collaboration for all of our members. One major project that was released before SCS 2015 is the new **community message board** (<http://community.iscbosc.org>) which is open not only to student council members, but the student community as a whole. Not only is this a space for discussing SC and RSG announcements, but also new and exciting research. Keep an eye out for future feature improvements that will keep everyone in touch online.

We have lots of plans for the coming year and encourage all of our members to get involved and stay in touch. Follow us on Facebook (facebook.com/iscbosc), Twitter (@ISCBOSC), LinkedIn (linkedin.iscbosc.org), and our website (iscbosc.org) for the latest updates on Student Council events.

Regional Student Groups

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 fulfil 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 a 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://rsg.iscb.org/content/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.

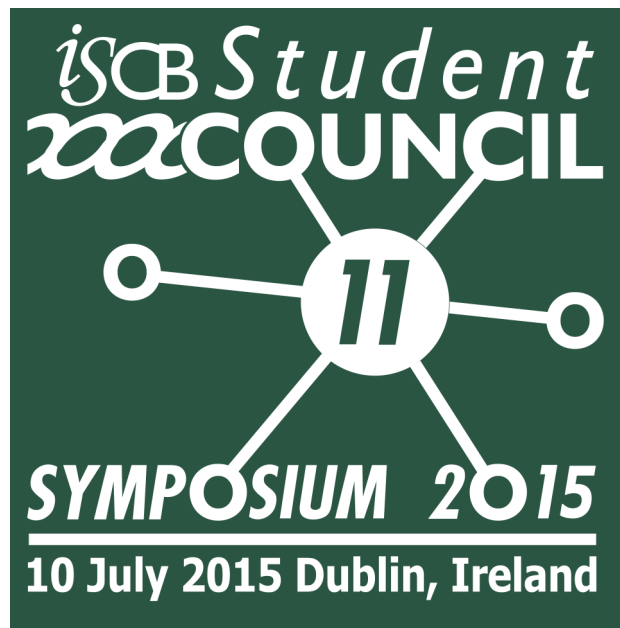
If you would like to find out more about the RSGs initiative or find out how you too can get involved in this, please visit iscbasc.org/rsg or send an email to rsg@iscbasc.org

Blast from the past



**A snap from the first Student Council Symposium, Africa
SCS-Africa 2015, Dar es Salaam, Tanzania**

,



Organisers' Bio

Student Council Executive Team

Pieter Meysman, Chair, ISCB Student Council



Pieter Meysman is a post-doctoral researcher at the Advanced Database Research and Modelling (ADReM) group of the University of Antwerp in Belgium. Trained as a PhD in bioscience engineering and now integrated into a data mining group, he serves as a bridge between the computer science and life science world. He has been serving as the Chair of the Student Council since 2015.

Alexander Junge, Vice Chair, ISCB Student Council



Alexander Junge is a PhD student at the Center for non-coding RNA in Technology and Health, University of Copenhagen, Denmark. His research interests include the application of machine learning techniques in RNA bioinformatics. He obtained a Bachelor of Science (2011) and Master of Science (2014) degree in Bioinformatics (Computational Molecular Biology) from Saarland University, Saarbrücken, Germany. Alexander Junge serves as Vice Chair of the ISCB Student Council since January 2015.

Jakob Berg Jespersen, Treasurer, ISCB Student Council



Jakob Berg Jespersen did both his Bachelor and Master Degree at the Technical University of Denmark, where he focused on Human Life Science and Applied chemistry, before getting interested in Computational Biology, he is now working with structural genomics and protein-protein interaction networks in the Lage Lab at Massachussets General Hospital.

Farzana Rahman, Secretary, ISCB Student Council and Co-chair, SCS2015



Farzana is a PhD student at the Genomics and Computational Biology Research Group at University of South Wales, UK. Her research focuses on developing methods to elucidate the role of gene families and network related to cell development and pathogenesis. Farzana got connected with Student Council since 2013. She has chaired SCS 2014, 1st RSG UK symposium and co-chairing SCS 2015. She is serving as the Secretary of the ISCB Student Council since October 2014.

Chinmay Kumar Dwibedi, RSG committee Chair, ISCB Student Council



Chinmay Dwibedi is a PhD candidate pursuing his research at the Swedish Defence Research Agency [FOI] and is affiliated to the Department of Clinical Microbiology, Umeå University, Sweden. His research includes Microbial Forensics and Population Genetics studies of bacterial pathogens. Chinmay first joined the ISCB-SC in 2007, at the beginning of his undergraduate studies. He chaired RSG-India in 2009 and continued to be associated with it until 2012. Since January 2013 he is chairing the RSG committee of the SC.

Student Council Executive Team

Anupama Jigisha, ISCB Board of Directors Representative



Anupama is a graduate student at University College Dublin working on the prediction of essential genes in pathogenic fungi. Her association with the ISCB Student Council (SC) started 6 years ago and ever since she served the student body in various capacities including its Chair, Finance chair and currently as the SC representative to the ISCB board of directors. She also co-organizes internships for students from developing nations as part of the SC internship program.

Student Council Symposium 2015 Organisers

Katherine Wilkins, Chair



Katie Wilkins is a Ph.D. candidate in the Field of Computational Biology at Cornell University in Ithaca, NY. Her research focuses on using machine learning and other bioinformatics approaches to determine the role of transcription activator-like proteins in plant diseases caused by bacteria of the genus *Xanthomonas*. Last year she served as program chair for SCS2014.

R Gonzalo Parra, Programme Committee Chair



R. Gonzalo Parra is a last year PhD candidate at the Protein Physiology Laboratory at Buenos Aires University in Argentina. His main interests are stem cells transcriptional regulation and protein folding. He holds a MSc. degree in Bioinformatics (2011) from the National University of Entre Rios in Argentina. Beyond science, Gonzalo is deeply interested in expanding the Bioinformatics field in Latin America being RSG Argentina's first president and now acting as the RSG Committee Vice Chair for Latin America.

Mehedi Hassan, Web Committee Chair



Mehedi is a third year Ph.D. student at the University of South Wales, UK. His research interest is in bio-fuel, evolutionary and functional genomics. He is currently studying networks related to lipid storage in oil producing monocots. Mehedi was introduced to the ISCB and the SC through the SCS2011. He was involved in organising the SCS2014. Mehedi is the founder and currently serving President of the ISCB RSG UK. He is also the co-chair to the student council's web committee.

Student Council Symposium 2015 Organisers

Dan DeBlasio, Web Committee Co-Chair



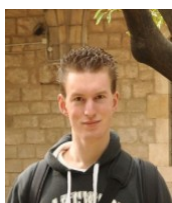
Dan DeBlasio is currently a Ph.D. candidate in the Department of Computer Science at the University of Arizona, USA where his work focuses on algorithms for computational biology, specifically, improving the quality of multiple sequence alignment through accuracy estimation and advising, he plans to defend in Fall 2015. Before coming to Arizona he received his B.S. and M.S. in Computer Science at the University of Central Florida. Dan has been an ISCB-SC Web Committee co-chair since 2012.

Margherita Francescato, Travel Fellowship Committee Chair



Margherita recently graduated and is currently a post-doctoral researcher in the Tübingen site of the German Center for Neurodegenerative Diseases. By analysing NGS expression data she studies the expression patterns that characterize distinct regions of the aged human central nervous system and how these patterns are altered in neurodegenerative diseases. She first came to know the ISCB Student Council in 2011 via the Regional Student Group of the Netherlands, and ever since has been involved in several of its committees and initiatives.

Bart Cuypers, Communications Chair



Bart is a PhD student in the Advanced Database Research and Modelling group at the University of Antwerp and in the Molecular Parasitology Unit at the Institute of Tropical Medicine, Antwerp. His research focusses on the application of systems biology, bio-informatics and data mining techniques to unravel developmental and drug-resistance mechanisms in Trypanosomes. Currently, he is the president of RSG Belgium.

Carla Luciana Padilla Franzotti, Designer



Carla is advanced undergraduate studying racing Biochemistry and Biotechnology at the National University of Tucuman (Argentina). She is part of the commission RSG-Argentina since 2014. Her main areas of interest: Bioinformatics, Computational Biology, Cell Biology and Human Clinic.

Thank you for attending

