

# BeNeLux Bioinformatics Conference

Luxembourg, 12-13 December 2011

Proceedings  
of BBC11



Organised by:

In partnership with:



---

### **The BBC11 Organising Committee**

Francisco Azuaje, CRP-Santé  
Aurélia Derischebourg, CRP-Santé  
Michaël Heymann, Integrated BioBank of Luxembourg (IBBL)  
Danièle Moes, CRP-Santé  
Daniel Struck, CRP-Santé  
Anne-Marie Ternes, CRP-Santé  
Mireille de Waha, CRP-Santé  
Anke Wienecke-Baldacchino, University of Luxembourg

### **The BBC11 Scientific Programme Committee**

Francisco Azuaje, CRP-Santé (Chair)  
Reinhard Schneider, University of Luxembourg (Chair)

Anne-Laure Boulesteix, University of Munich, Germany  
Marie-Dominique Devignes, Loria, France  
Chris Evelo, Maastricht University, The Netherlands  
Ivo Gut, CNAG, Spain  
Michaël Heymann, IBBL  
Yongsheng Huang, University of Michigan, USA  
Andreas Kremer, Erasmus Medical Center, The Netherlands  
Jack Leunissen, Wageningen University, The Netherlands  
Frédérique Lisacek, Swiss Institute of Bioinformatics, Switzerland  
Sara Madeira, Instituto Superior Técnico, Portugal  
Lennart Martens, Ghent University, Belgium  
Yves Moreau, Katholieke Universiteit Leuven, Belgium  
Peter Nazarov, CRP-Santé, Luxembourg  
Isabel Nepomuceno, University of Seville, Spain  
Léon-Charles Tranchevent, Katholieke Universiteit Leuven, Belgium  
Annemie Vandamme, Katholieke Universiteit Leuven, Belgium  
Yves Van de Peer, Ghent University, Belgium  
Gerrit Vriend, Radboud University Nijmegen Medical Centre, The Netherlands  
Louis Wehenkel, University of Liège, Belgium  
Zhongming Zhao, Vanderbilt University Medical Center, USA

<b>WELCOME NOTE.....</b>	<b>1</b>
<b>PROGRAM.....</b>	<b>2</b>
<b>12 DECEMBER 2011.....</b>	<b>2</b>
<b>FLASH ORAL PRESENTATIONS – A.....</b>	<b>3</b>
<b>13 DECEMBER 2011.....</b>	<b>3</b>
<b>FLASH ORAL PRESENTATIONS – B.....</b>	<b>4</b>
<b>POSTER SESSION A.....</b>	<b>5</b>
<b>POSTER SESSION B.....</b>	<b>8</b>
<b>ABSTRACTS ORAL PRESENTATIONS DAY 1 .....</b>	<b>11</b>
<b>ABSTRACTS ORAL PRESENTATIONS DAY 2 .....</b>	<b>20</b>
<b>ABSTRACTS FLASH ORAL PRESENTATIONS A AND POSTER SESSION A.....</b>	<b>26</b>
<b>ABSTRACTS FLASH ORAL PRESENTATIONS B AND POSTER SESSION B.....</b>	<b>30</b>
<b>ABSTRACTS POSTER SESSION A.....</b>	<b>35</b>
<b>ABSTRACTS POSTER SESSION B.....</b>	<b>72</b>
<b>AUTHORS INDEX .....</b>	<b>110</b>

## Welcome Note

### **The BeNeLux Bioinformatics Conference 2011**

Welcome to the BeNeLux Bioinformatics Conference 2011. We are proud to host this conference in Luxembourg for the first time, and offer you a great opportunity to discuss progress in bioinformatics and related fields.

We are delighted to have received more than one hundred abstracts by authors based in fourteen countries and three continents. BBC11 showcases peer-reviewed contributions in a wide range of areas relevant to health and biotechnologies. A key theme of BBC11 is “Bioinformatics: Enabling Translational Biomedical Research”. This will allow us to highlight significant challenges and opportunities to move the lab closer to the clinic.

Apart from a select group of oral presentations and two poster sessions, we are pleased to offer you a programme that fosters cross-disciplinary discussions and cooperation. We are honoured to have four world-class scientists as guest speakers. Their contributions to BBC11 will not only provide novel insights into scientific and technological advancements, but also exciting perspectives and research directions.

Our gratitude to our partners, sponsors and supporters for helping us to put the Lux in BeNeLux.

The BBC11 Organisation

## Program

	<b>12 December 2011</b>
8:00	<b>Collection of delegate bags, new registrations, coffee</b>
9:30	<b>Opening</b> Francisco Azuaje, BBC11 Organising Committee. Jean-Claude Schmit, CEO of CRP-Santé. Peter Sodermans, Luxembourg for Business. Reinhard Schneider, Programme Co-Chair.
	<b>Session 1</b> Chair: Reinhard Schneider, University of Luxembourg
10:00	<b>Guest Talk</b> Evolution teaches protein prediction Burkhard Rost, Technical University Munich, Germany
10:45	Interfering with the interaction network of adenyl cyclase virulence factors Therese Malliavin, Institut Pasteur, France
11:05	Computational annotation and interpretation of single nucleotide variation to identify disease-causing variants by next-generation sequencing Alejandro Sifrim, K.U.Leuven, Belgium
11:30	<b>Session 2 (Flash oral presentations - A)</b> Chair: Michaël Heymann, IBBL, Luxembourg
12:00	<b>Lunch</b>
	<b>Session 3</b> Chair: Andreas Kremer, Erasmus Medical Center, The Netherlands
13:00	Prediction of a phosphorylation network in Arabidopsis thaliana Kris Laukens, University of Antwerp, Belgium
13:20	Spatial Clustering of Protein Binding Sites for Template Based Protein Docking Anisah Ghoorah, INRIA, France
13:40	An integrated workflow for robust alignment and simplified quantitative analysis of NMR spectrometry data Trung Nghia Vu, University of Antwerp, Belgium
14:00	Systems Biology Analysis of Tyrosine Kinase Inhibitor Target Profiles in Leukemia Jacques Colinge, CeMM, Austria
14:20	An Efficient Algorithm to Perform Multiple Testing in Epistasis Screening François Van Lishout, University of Liège, Belgium
14:40	<b>Break</b>
	<b>Session 4</b> Chair: Anna Chioti, CRP-Santé, Luxembourg
15:15	Statistical interpretation of machine learning-based feature rankings for biomarker discovery Vân Anh Huynh-Thu, University of Liège, Belgium
15:35	MSCOMPARE- Data Processing Framework for Quantitative Processing of Label-Free LC-MS data Peter Horvatovich, University of Groningen, The Netherlands

15:55	<b>Guest Talk</b> Applying Bioinformatics to Clinical Practice in Treating HIV Infections: Where Do We Stand? Thomas Lengauer, Max-Planck-Institut für Informatik, Germany
16:45	<b>Poster session A</b> (to 17:45)
17:15	<b>Speed research dating</b> (to 18:00)
19:00	<b>Reception Dinner</b>

### Flash oral presentations – A

eBionics: a Bioinformatics e-Learning Environment for Biologists  
Patrick Koks, Wageningen University, The Netherlands

iSNP: An Integrated, Automatically Updated SNP Database Server Over Web  
Yesim Aydin Son, Middle East Technical University, Turkey

Exposing WikiPathways as Linked Open Data  
Andra Waagmeester, Maastricht University, The Netherlands

Implementation of the Protein Fluorescence And Structural Toolkit (PFAST)  
Cynthia N. Prudence, University of Rhode Island, USA

	<b>13 December 2011</b>
	<b>Session 5</b> Chair: Marie-Dominique Devignes, LORIA, France
9:00	<b>Guest Talk</b> Translational Bioinformatics : From biocuration to model predictions and back Ioannis Xenarios, Swiss Institute of Bioinformatics, Switzerland
9:45	Probic: Simultaneously detecting coexpression modules and their regulatory patterns Yan Wu, K.U.Leuven, Belgium
10:05	Visualizing genotype-phenotype relationships across cell cycle and evolutionary time scales Maria Secrier, EMBL, Germany
10:25	Detection of genes essential for growth of respiratory pathogens Aldert Zomer, Nijmegen Centre for Molecular Life Sciences, The Netherlands
10:45	<b>Break</b>
11:30	<b>Debate</b> Technological and Cooperation Challenges in Bioinformatics
12:30	<b>Lunch</b>
	<b>Session 6</b> Chair: Gert Vriend, Radboud University, The Netherlands
13:30	Bayesian Inference of Protein Complex Modules from Affinity Purification Mass Spectrometry Data Alexey Stukalov, CeMM, Austria

13:50	Data-analytical strategies for enrichment-based genome-wide DNA-methylation profiling by NGS Tim De Meyer, Ghent University, Belgium
14:10	SABIO-RK: A Resource for Biomedical Research Renate Kania, Heidelberg Institute for Theoretical Studies, Germany
14:30	<b>Session 7 (Flash oral presentations - B)</b> Chair: Merja Heinäniemi, University of Luxembourg
15:00	<b>Poster session B (with coffee break)</b>
	<b>Closing Session</b> Chair: Francisco Azuaje, CRP-Santé, Luxembourg
16:00	<b>Guest Talk</b> Genome-Wide Translational Medicine Applications Peter van der Spek, Erasmus Medical Center, The Netherlands
16:45	<b>Closing remarks</b> Yves Moreau, K.U.Leuven, Belgium

### **Flash oral presentations – B**

BioGraph: unsupervised biomedical knowledge discovery via automated hypothesis generation  
Anthony ML Liekens, Universiteit Antwerpen, Belgium

Origin and Evolution of the Organellar Release Factor Family  
Isabel Duarte, Radboud University, The Netherlands

Protein regulation dynamics analysis in R  
Florian P Breitwieser, Research Center for Molecular Medicine, Austria

Inheritance analysis and quality control for Next Generation Sequencing data  
Joep de Ligt, UMC St. Radboud, The Netherlands

Reflect: an augmented browsing tool for life scientist  
Janos Binder, EMBL, Germany

## Poster Session A

ID	Poster
A1	<b>eBiomics: a Bioinformatics e-Learning Environment for Biologists;</b> Patrick Koks, Pascale Berthault, Guy Bottu, Jacques van Helden, Jean-Pierre Kraehenbuhl, Frédérique Lisacek, Grégoire Rossier, Jean Sylvestre, Jack Leunissen; Laboratory of Bioinformatics, Wageningen University, The Netherlands.
A3	<b>iSNP: An Integrated, Automatically Updated SNP Database Server Over Web;</b> Ceyhun Gedikoğlu, Levent Çarkacıoğlu, Yeşim Aydın Son; Bioinformatics Graduate Program; Health Informatics Department, Middle East Technical University, Ankara, Turkey.
A4	<b>Exposing WikiPathways as Linked Open Data;</b> Andra Waagmeester, Helena F. Deus, Chris T. Evelo; Department of Bioinformatics - BiGCaT, Maastricht University; Digital Enterprise Research Institute, National University of Ireland, Galway.
A5	<b>Implementation of the Protein Fluorescence And Structural Toolkit (PFAST);</b> Cynthia N. Prudence, Yana K. Reshetnyak; Physics Department, University of Rhode Island, USA.
A6	<b>Gene set analysis in the cloud;</b> Lu Zhang, Shengchang Gu, Yuan Liu, Bingqiang Wang, Francisco Azuaje; Laboratory of Cardiovascular Research, CRP-Santé, Luxembourg; BGI, China.
A7	<b>Breaching the surface with HOPE;</b> Jules Kerssemakers; CMBI, Radboud University Nijmegen Medical Centre, The Netherlands.
A8	<b>Towards a Standard for Cooperative Interactions;</b> Kim Van Roey, Henning Hermjakob, Samuel Kerrien, Toby J. Gibson; Structural and Computational Biology Unit, European Molecular Biology Laboratory, Heidelberg, Germany.
A9	<b>Gene expression evolution on the emergence of pathogenicity in Ascomycetes;</b> Amina Sanchez-Rodriguez, Riet De Smet, Kristof Engelen, Qiang Fu, Yan Wu, Kathleen Marchal; Centre of Microbial and Plant Genetics, Department of Microbial and Molecular Systems, K.U.Leuven, Belgium.
A10	<b>Prediction of bacterial relationships in the human microbiome;</b> Karoline Faust, J. Fah Sathirapongsasuti, Curtis Huttenhower, Jeroen Raes; Bioinformatics and (Eco-) Systems Biology, VIB-Vrije Universiteit Brussel, Belgium.
A11	<b>Union makes strength: Building baseline Tracks from 69 open access full human genomes;</b> Stephane Plaisance, Mark Veugelers; BITS, VIB, Belgium.
A12	<b>Using Hilbert curves to visualize structural variations with Meander;</b> Georgios A. Pavlopoulos, Alejandro Sifrim, Jan Aerts; Faculty of Electrical Engineering - ESAT/SCD, Katholieke Universiteit Leuven, Leuven, Belgium.
A13	<b>Genomic Profiling of HMGN;</b> Arjan van der Velde; National Center for Biotechnology Information (NCBI/NIH), United States.
A14	<b>Proteins in the orthology twilight zone. The Ortho-Profile iterative method and the experimental function confirmation;</b> Radek Szklarczyk, Bas F.J. Wanschers, Thomas D. Cuypers, Leo G. Nijtmans, Martijn A. Huynen; CMBI, Radboud University Nijmegen Medical Centre.
A15	<b>Specificity of allostery networks within the SH3 domain family;</b> Ana Zafra Ruano, José Couceiro, Javier Ruiz Sanz, Joost Schymkowitz, Frederic Rousseau, Irene Luque, Tom Lenaerts; Département d'Informatique, Université Libre de Bruxelles, Belgium.
A16	<b>Identifying common structural DNA properties in transcription factor binding site sets of the LacI-GalR family;</b> Meysman Pieter, Marchal Kathleen, Engelen Kristof; M2S, K.U.Leuven, Belgium.



A17	<b>Modelling the dynamics of chronic myeloid leukemia under therapy;</b> Tom Lenaerts, Fausto Castagnetti, Arne Traulsen, Jorge M. Pacheco, Gianantonio Rosti, David Dingli; Département d'Informatique, Université Libre de Bruxelles, Belgium.
A19	<b>PROCAR-SEQ An analysis and visualization framework for next-generation sequencing based quantification of prokaryotic communities;</b> Joachim De Schrijver, Pieter-Jan Volders, Frederiek-Maarten Kerckhof, Dagmar Obbels, Elie Verleyen, Wim Vyverman, Tim De Meyer, and Wim Van Criekinge; Laboratory of Computational Genomics and Bioinformatics (BioBix), Ghent University, Ghent, Belgium.
A20	<b>The Fc receptor complex from human neutrophils;</b> Florentinus AK, Jankowski A, Petrenko V, Bowden P, Marshall JG; Department of Chemistry and Biology, Ryerson University, Canada.
A21	<b>Proteomic identification of differentially expressed proteins in curcumin-treated prostate cancer cells;</b> Anthoula Gaigneaux, Marie-Hélène Teiten, Sébastien Chateauvieux, Anja Billing, Jenny Renaut, Claude P. Müller, Mario Dicato, Marc Diederich; Laboratoire de Biologie Moléculaire et Cellulaire du Cancer.
A22	<b>Divergence in the length of unstructured protein termini drives the evolution of protein half-life;</b> Robin van der Lee, Kai Kruse, Benjamin Lang, Jörg Gsponer, Natalia Sánchez de Groot, Monika Fuxreiter, M. Madan Babu; MRC Laboratory of Molecular Biology, Cambridge, United Kingdom; Centre for Molecular and Biomolecular Informatics, Radboud University Nijmegen Medical Centre, The Netherlands.
A23	<b>PuMaQC: R-based pipeline for the search, import and QC/QA of public microarray data;</b> Joana P. Corte-Real, Petr V. Nazarov, Arnaud Muller, Tony Kaoma and Laurent Vallar; Microarray Center, CRP-Sante, Luxembourg.
A24	<b>Identification of crosstalk strength in signalling networks by optimization of Probabilistic Boolean Network models;</b> Thomas Sauter, Panuwat Trairatphisan; Life Sciences Research Unit, University of Luxembourg.
A25	<b>Accuracy of information flow predictions within the PTP1E PDZ2 domain;</b> Elisa Cilia, Geerten W. Vuister, Tom Lenaerts; Département d'Informatique, Université Libre de Bruxelles, Belgium.
A26	<b>DALAS: an R-Java desktop application for Affymetrix exon array data analysis;</b> Tony Kaoma, Christelle Ghoneim, Arnaud Muller, Petr Nazarov, Laurent Vallar; Microarray Centre, CRP-Sante, Luxembourg.
A27	<b>Cis-regulation of toxin clusters in Fusarium;</b> Valeria Montis, Francesca Cardinale, Ivan Visentin, Marco Beyer, Hoffmann Lucien, Harold Corby Kistler, Matias Pasquali; Dept of Plant Physiology, University of Torino; CRP- Gabriel Lippmann, EVA Department; USDA ARS Cereal Disease Laboratory, St Paul USA.
A28	<b>Analysis of the X!TANDEM correlation of the HuPO blood consortium results by SQL and SAS;</b> Bowden P, Beavis R, Marshall JG; Department of Chemistry and Biology, Ryerson University, Canada.
A29	<b>Quantitative statistical analysis of blood proteins from liquid chromatography, electrospray ionization and tandem mass spectrometry;</b> Bowden P, Zhu P, McDonnell M, Thiele H, Marshall JG; Department of Chemistry and Biology, Ryerson University, Canada.
A30	<b>Analyzing gene and protein expression variance in cellular pathways using high-throughput experimental data;</b> Enrico Glaab, Reinhard Schneider; Luxembourg Centre for Systems Biomedicine.
A31	<b>COLOMBOS: Access Port for Cross-Platform Bacterial Expression Compendia;</b> Kristof Engelen, Qiang Fu, Pieter Meysman, Amina Sánchez-Rodríguez, Riet De Smet, Karen Lemmens, Ana Carolina Fierro, Kathleen Marchal; Department Of Microbial And Molecular Systems (M2S), KU Leuven, Belgium.
A32	<b>Receptor-Ligand Prediction through Machine Learning;</b> Ernesto Iacucci, D. Popovic, L.-C. Tranchevent, B. De Moor, and Y. Moreau; KU Leuven, ESAT/SCD.

A33	<b>Metabolic Modeling of Human Adipogenesis;</b> Mafalda Galhardo, Merja Heinäniemi, Thomas Sauter; LSRU, University of Luxembourg, Luxembourg.
A34	<b>Merging partially labelled trees;</b> Anthony Labarre, Sicco Verwer; Department of Computer Science, Katholieke Universiteit Leuven, Belgium.
A35	Evaluation of novel SNP prioritization and sub-set selection approaches for GWAS; Yesim Aydin Son; Bioinformatics Graduate Program; Health Informatics Department, Middle East Technical University, Ankara, Turkey.
A36	<b>The influence on protein structure of single nucleotide polymorphisms identified in Mycobacterium tuberculosis;</b> E Vandermarliere, T Muth, J Blackburn, L Martens; Department of Biochemistry, Ghent University, Belgium and Department of Medical Protein Research, VIB, Belgium.
A37	<b>Comparative genomics of GlnR-mediated transcription regulation in Gram-positive bacteria;</b> Tom Groot Kormelink, Eric Koenders, Yanick Hagemeijer, Lex Overmars, Roland J. Siezen, Willem M. de Vos and Christof Francke; Kluyver Centre for Genomics of Industrial Fermentation, GA Delft, The Netherlands; I Food and Nutrition, Wageningen, The Netherlands; Laboratory of Microbiology, Wageningen University & Research Centre, Wageningen, The Netherlands; Netherlands Bioinformatics Centre, Nijmegen, The Netherlands; Centre for Molecular and Biomolecular Informatics, NCMLS, Radboud University Nijmegen Medical Centre, Nijmegen, The Netherlands.
A38	<b>Use of domain knowledge for dimension reduction: Application to drug side effects;</b> Emmanuel Bresso, Sidahmed Benabderrahmane, Malika Smail-Tabbone, Gino Marchetti, Arnaud Sinan Karaboga, Michel Souchet, Amedeo Napoli and Marie-Dominique Devignes; LORIA CNRS, Nancy University, INRIA Nancy Grand-Est and Harmonic Pharma (SAS), France.
A39	<b>Identify me, says the alien peptide. Combining publicly available mS/MS repositories and clustering tools to identify peptides from non-model organisms;</b> Gerben Menschaert, Eisuke Hayakawa, Wim Van Criekinge, Geert Baggerman; Laboratory of Computational Genomics and Bioinformatics (BioBix), Ghent University, Ghent, Belgium.
A40	<b>A method to detect mono- and biallelic DNA-methylation from MBD-seq using single nucleotide polymorphisms;</b> Sandra Steyaert, Ayla De Paepe, Geert Trooskens, Simon De Nil, Wim Van Criekinge, Tim De Meyer; Department of Mathematical Modelling, Statistics and Bioinformatics, Ghent University, Ghent, Belgium.
A41	<b>Multivariate Framework for Biomarker Discovery;</b> Yousef El Aalamat, Dusan Popovic, Etienne Waelkens, Bart de Moor; Katholieke Universiteit Leuven, Dept. of Electrical Engineering, (ESAT), CD-SISTA (BIOI), Leuven, Belgium.
A42	<b>Preprocessing approach for single-cell array Comparative Genomic Hybridization;</b> J. Cheng, P. Konings, E. Vanneste, T. Voet, J. Vermeesch, Y. Moreau; Department of Electrical Engineering, ESAT-SCD, Katholieke Universiteit Leuven, Belgium; IBBT-K.U.Leuven Future Health Department, Belgium; Center for Human Genetics, Katholieke Universiteit Leuven.
A43	<b>Analysis of critical transitions in Parkinson's disease;</b> Christophe Trefois, Paul Antony, Aidon Baumuratov, Sandra Koeglsberger, Olga Boyd, Rudi Balling; Luxembourg Centre for Systems Biomedicine, Luxembourg.

## Poster Session B

ID	Poster
B1	<b>BioGraph: unsupervised biomedical knowledge discovery via automated hypothesis generation;</b> Anthony ML Liekens, Jeroen De Knijf, Walter Daelemans, Bart Goethals, Peter De Rijk, Jurgen Del-Favero; Universiteit Antwerpen.
B2	<b>Origin and Evolution of the Organellar Release Factor Family;</b> Isabel Duarte, Sander Nabuurs, Ramiro Magno, Martijn Huynen; CMBI, Radboud University, The Netherlands.
B3	<b>Protein regulation dynamics analysis in R;</b> Florian P Breitwieser, Alexey Stukalov, Jacques Colinge; Research Center for Molecular Medicine, Austria.
B4	<b>Inheritance analysis and quality control for Next Generation Sequencing data;</b> Joep de Ligt, Lisenka E.L.M. Vissers, Christian Gilissen, Joris A. Veltman, Jayne Y. Hehir-Kwa; Department of Human Genetics, UMC St. Radboud, Nijmegen.
B5	<b>Reflect: an augmented browsing tool for life scientist;</b> Janos Binder, Sean O'Donoghue, Sune Frankild, Lars Juhl Jensen, Reinhard Schneider; Structural and Computational Unit, EMBL, Germany.
B6	<b>Prediction of cardiac perturbations of non-cardiovascular drugs with My-DTome;</b> Lu Zhang, Yvan Devaux, Daniel R. Wagner, Francisco Azuaje; Laboratory of Cardiovascular Research, CRP-Santé, Luxembourg.
B8	<b>Charting The Methylome;</b> Geert Trooskens, Tim de Meyer, Simon Denil, Wim Van Criekinge; Ghent University.
B10	<b>Mining the rest-fraction in RNA-seq experiments;</b> Simon Denil, Tina Kyndt, Annelies Haegeman, Geert Trooskens, Tim De Meyer, Wim Van Criekinge, Godelieve Gheysen; Laboratory of Computational Genomics and Bioinformatics (BioBix), Ghent University, Ghent, Belgium.
B11	<b>Functional classification and co-expression analysis of genetically imprinted Genes;</b> M.Hamed, Siba Isma'el, Martina Polsky, Volkhard Helms; Chair of Computational Biology, Saarland University Germany.
B12	<b>Micro peptides as a new class of bio-active peptides in Eukaryotes;</b> Jeroen Crappé, Gerben Menschaert, Geert Trooskens, Joachim Deschrijver, Geert Baggerman, Wim Vancrickinge; Biobix, Ghent University.
B14	<b>Predicting the Interactions between G-Protein Coupled Receptors:</b> Computational and Experimental Approaches; Mehmet Emre Şahin, Tolga Can, Cagdas D. SON; Biyological Sciences Department, METU, Turkey.
B15	<b>“Last In First Out” gain and loss of the Intra Flagellar Transport components;</b> John van Dam, Martijn Huynen; Centre for Molecular and Biomolecular Informatics, Nijmegen Centre for Molecular Life Sciences, Radboud University Nijmegen Medical Centre, Nijmegen, the Netherlands.
B16	<b>Are REPs genetic insulators that enable differential regulation of gene expression in bacteria?;</b> Lex Overmars, Tom Groot Kormelink, Roland Siezen and Christof Francke; TI Food and Nutrition, The Netherlands; Radboud University Medical Centre; Centre for Molecular and Biomolecular Informatics, The Netherlands, 3) Kluyver Centre for Genomics of Industrial Ferm
B17	<b>Automated comparative genome annotation in prokaryotes: a halt to error propagation?</b> Thomas H. A. Ederveen, Amy de Bruin, Brechtje Hoegen, Bernadet Renckens, Roland J. Siezen, Sacha A. F. T. van Hijum; Centre for Molecular and Biomolecular Informatics, Radboud University Medical Centre, the Netherlands; HAN University of Applied Sciences, the Netherlands; NIZO food research, the Netherlands.
B18	<b>Dynamic co-regulation of miRNAs and mRNAs following cytokine stimulation of melanoma cells;</b> Susanne Reinsbach, Peter V. Nazarov, Martina Schmitt, Demetra Philippidou, Nathalie Nicot, Iris Behrmann, Laurent Vallar, Stephanie Kreis; Universtiy of Luxembourg; Microarray Center, CRP-Santé, Luxembourg.

B19	<b>Critical assessment of candidate gene prioritization methods;</b> Daniela Börnigen, Léon-Charles Tranchevent, Francisco Bonachela-Capdevila, Koenraad Devriendt, Bart de Moor, Patrick De Causmaecker, and Yves Moreau; Department of Electrical Engineering, KULeuven, BE; IBBT-K.U.Leuven Future Health Department, BE; Department of Computer Science, KULeuven, BE; Center for Human Genetics, KULeuven, BE.
B20	<b>A kernel based framework for cross-species candidate gene prioritization;</b> Shi Yu, Léon-Charles Tranchevent, Sonia Leach, Pooya Zakeri, Bart De Moor, and Yves Moreau; Department of Electrical Engineering ESAT-SCD, Katholieke Universiteit; IBBT-K.U.Leuven Future Health Department, Leuven, Belgium.
B21	<b>A novel approach for the identification of miRNAs;</b> Bart Aelterman, Peter De Rijk, Jurgen Del-Favero; Applied Molecular Genomics Unit, VIB Department of Molecular Genetics, VIB, University of Antwerp, Belgium.
B22	<b>Arrayanalysis.org: friendly solutions for standardised microarray analysis;</b> Lars Eijssen, Magali Jaillard, Michiel Adriaens, Anwesha Dutta, Martina Kutmon, Philip de Groot, Chris Evelo; Department of Bioinformatics, BiGCaT, Maastricht University, NL; Nutrition, Metabolism and Genomics Group, Wageningen, NL.
B23	<b>Comparative studies of genome-wide DNA methylation arrays and gene expression data: a breast cancer as an example;</b> Singhal Sandeep K., Desmedt Christine, Ignatiadis Michail, Sotiriou Christos, Michiels Stefan; Breast Cancer Translational Research Laboratory, Institut Jules Bordet, Université Libre de Bruxelles, Bruxelles, Belgium.
B24	<b>Classification of MCAD deficiency using tandem MS neonatal screening data;</b> Tim Van den Bulcke; biomina, Antwerp University Hospital, Belgium.
B25	<b>Capture of an activated receptor complex from the surface of live cells by affinity receptor chromatography;</b> Jankowski A, Zhu P, Marshall JG; Department of Chemistry and Biology, Ryerson University, Canada.
B26	<b>A Custom-Designed Peak Picking Algorithm for Mass Spectral Imaging Data;</b> Nico Verbeeck, Raf Van de Plas, Etienne Waelkens, Bart De Moor; ESAT-SCD, Katholieke Universiteit Leuven, Belgium.
B27	<b>A non-parametric method to assess the presence of significant DNA-methylation in enrichment-based NGS data;</b> Klaas Mensaert, Geert Trooskens, Wim Van Criekinge, Olivier Thas ,Tim De Meyer; Dept. Mathematical Modelling, Statistics and Bioinformatics, Ghent University, Belgium.
B28	<b>Finding the differences: whole genome sequencing and analysis of a monozygotic twin discordant for schizophrenia;</b> Peter De Rijk, Joke Reumers, Anthony Liekens, Maarten Van Den Bossche, Diether Lambrechts, Jurgen Del-Favero; VIB DMG / University of Antwerp.
B29	<b>Evaluating the use of clustering trees for protein subfamily identification;</b> Eduardo de Paula Costa, Celine Vens, Hendrik Blockeel; Dept. of Computer Science, Katholieke Universiteit Leuven, Belgium; Leiden Institute of Advanced Computer Science, Universiteit Leiden, The Netherlands.
B30	<b>Analyzing NF-κB signaling networks in UVB treated Cutaneous T-Cell Lymphoma cell-lines using Gene expression profiling and Ingenuity Pathway Analysis;</b> Amit Kumar, Thomas Sauter , Silvia Racolta , Dagmar Kulms, Petr Nazarov, Laurent Vallar; Life Sciences Research Unit, University of Luxembourg.
B31	<b>Diabetes and Parkinson's - two old friends? Unraveling potential connections between diseases by text-mining techniques;</b> Maria Biryukov, Serge Eifes, Janos Binder, Venkata P. Satagopam, Reinhard Schneider; onScale Solutions, Germany.
B32	<b>DTSpine: DTProbLog for pathway inference from cause-effect experiments;</b> Joris Renkens, Guy Van den Broeck, Siegfried Nijssen, Kathleen Marchal K.U.Leuven.
B33	<b>Top-down, bottom-up and middle out perspectives to model mitochondrial dysfunction and ROS generation in relation to neurodegenerative diseases;</b>

	Alexey Kolodkin, Hans V. Westerhoff, Antonio del Sol Mesa and Rudi Balling; Luxembourg Centre for Systems Biomedicine, University of Luxembourg, Luxembourg; Molecular Cell Physiology, VU University Amsterdam, the Netherlands.
B34	<b>The impact of copy number variation in mental retardation: exploratory analysis;</b> Peter Konings, Joris R. Vermeesch, Yves Moreau; ESAT, KU Leuven, Belgium.
B35	<b>Clinical Data Miner - an Electronic Data Capture software framework that improves interrater agreement;</b> Arnaud Installé, Dirk Timmerman, Thierry Van den Bosch, Bart De Moor; ESAT - SCD, Katholieke Universiteit Leuven, Belgium.
B36	<b>Generate pseudo expression scores from ChIP-seq data;</b> Filip Pattyn, Frank Westermann, Frank Speleman, Jo Vandesompele; Center for Medical Genetics, Ghent University, Belgium.
B37	<b>Introducing conformational variability into X-ray structures based on experimental NMR data;</b> Wim Vranken, Alex Volkov, Tom Lenaerts, Nico van Nuland; Department of Structural Biology, VIB and Structural Biology Brussels, Vrije Universiteit Brussel, Brussel, Belgium.
B38	<b>Transcriptor: a web-tool for mining whole-genome transcriptomes of prokaryotes;</b> Tilman Todt, Roland J. Siezen, Sacha A.F.T. van Hijum; Center for Molecular and Biomolecular Informatics, Nijmegen Center for Molecular Life Sciences, Radboud University Medical Centre, Nijmegen, The Netherlands.
B39	<b>Network analysis of differential expression for drug target prioritization;</b> Griet Laenen, Lieven Thorrez, Yves Moreau; ESAT, K.U.Leuven, Belgium.
B40	<b>Tree-based machine learning methods for Zebrafish Image Analysis;</b> Olivier Stern, Nathalie Jeanray, Raphaël Marée, Jessica Aceto, Benoît Pruvot, Pierre Geurts, Marc Muller, Louis Wehenkel; GIGA-Systems Biology & Chemical Biology, GIGA-R and Dept. EE CS, University of Liege, Belgium.
B41	<b>Inferring gene association network from gene expression data using quantitative association rules;</b> María Martínez-Ballesteros, Isabel Nepomuceno-Chamorro, José C. Riquelme; Lenguajes y Sistemas Informáticos, Universidad de Sevilla, Spain.
B42	<b>Inferring gene co-expression networks with Biclustering based on linear correlations among genes;</b> Juan A. Nepomuceno, Isabel Nepomuceno-Chamorro, Alicia Troncoso, Jesus Aguilar-Ruiz; Lenguaje y Sistemas Informaticos, Universidad de Sevilla, España.
B43	<b>Comparative analyses imply that the enigmatic sigma factor 54 is a central controller of the bacterial exterior;</b> Christof Francke, Tom Groot, Kormelink, Yanick Hagemeijer, Lex Overmars, Vincent Sluijter, Roy Moezelaar and Roland J. Siezen; TI Food and Nutrition, Wageningen; Kluyver Center for Genomics of Industrial Fermentation, Delft; Wageningen University RC; NBIC, Nijmegen; CMBI, RU Nijmegen MC, the Netherlands.
B44	<b>Mutalyzer 2: improved sequence variant descriptions from next generation sequencing data and gene variant databases;</b> J.F.J. Laros, M. Vermaat, G.R. Stouten, J.T. den Dunnen, P.E.M. Taschner; Center for Human and Clinical Genetics, Leiden University Medical Center, The Netherlands.

## Abstracts Oral Presentations Day 1

### **Interfering with the interaction network of adenyl cyclase virulence factors**

E Laine, C Goncalves, L Martinez, E Selwa, A Blondel, S Rault, D Ladant, TE Malliavin\*

Unite de Bioinformatique Structurale Institut Pasteur and URA CNRS 2185

Contact: terez@pasteur.fr

**Background:** Adenyl cyclases play an important role into the virulence of several pathogens, in particular the agents of whooping cough (*Bordetella pertussis*) and of anthrax (*Bacillus anthracis*). The virulence is attained by interaction with the ubiquitous protein calmodulin, which leads to adenyl cyclase activation and cAMP overproduction, disorganizing the signaling network into the host cell. Several X-ray crystallographic structures are available for these adenyl cyclases, giving insights into the interaction with calmodulin at the molecular level. Starting from this structural information, computational approaches permits to propose biochemical and medicinal approaches to block the adenyl cyclase activation.

**Methods:** Molecular dynamics simulations and transition path calculations have been used to analyze the dynamics and energetics of the proteins interaction. A list of possible interaction inhibitors was obtained by virtual screening. Predicted protein mutations and inhibitors were experimentally confirmed.

**Results:** Simplified models of the calmodulin-adenyl cyclase complex energetics were proposed, based on the measure of the energetic influences between the complex domains. These influences describe the system energetics as an oriented graph network, allowing to define important interactions and binding pockets which should be targeted in order to influence the adenyl cyclase activation. Protein sequence mutations and inhibitors proposed by computational methods, were proved by biochemical assays and mutagenesis to have indeed a strong influence on the interaction. In that way, a new family of inhibitors was discovered. Mutations of the B pertussis adenyl cyclase, annealing the protein activation, were put in evidence.

**Conclusions:** The extensive use of structural bioinformatics approaches allowed to extend the information provided by X-ray crystallographic structures in order to obtain a better understanding of the dynamics and the energetics of adenyl cyclases activation. The experimental findings made possible by the bioinformatics prediction, will be further investigated using additional biochemical modifications and high-resolution structural methods.

**Computational annotation and interpretation of single nucleotide variation to identify disease-causing variants by next-generation sequencing**

Alejandro Sifrim\*, Zeynep Kalender, Yuching Lai, Georgios A. Pavlopoulos, Stein Aerts, Joris R. Vermeesch, Yves Moreau, Jan Aerts

Department of Electrical Engineering, ESAT/SCD, K.U.Leuven, Belgium

Contact: [alejandro.sifrim@esat.kuleuven.be](mailto:alejandro.sifrim@esat.kuleuven.be)

**Background:** Interpreting exome and genome sequencing data from patients to discover disease-causing variants requires new computational and statistical analysis methods because of the size and complexity of such data. Bottlenecks towards identifying the variation underlying rare Mendelian diseases currently include poor cross-sample querying, difficulty in setting cutoffs for data filtering, and computationally intensive and constantly changing functional annotation.

**Methods:** We describe a methodology (embodied in our web application Annotate-it) that makes it possible for geneticists to explore and analyze patient sequencing data towards the identification of causal variants under several underlying genetic hypotheses (recessive, dominant, and de novo inheritance). As a novelty, Annotate-it offers interactive visual analytics to effectively determine appropriate filtering criteria. We also propose a novel pathway analysis technique based on regression models to discover significantly overmutated pathways in groups of sequenced exomes, this allows for the discovery of oligogenic etiologies.

**Results:** We demonstrate the effectiveness of our strategy on two case studies: Schinzel-Giedion syndrome and a semi-synthetic Miller syndrome data set. In these datasets we show that simple analysis strategies show significant power when trying to discover the cause of rare Mendelian disease.

**Conclusions:** Here we describe Annotate-it, a versatile framework for the analysis of multisample SNV data generated by NGS. Annotation of samples is performed on the server side, eliminating the need for the installation of complex tools and annotation sources by the end-user and automatically keeping those annotations up to date. Thanks to novel interactive visual analytics techniques, we facilitate the selection of optimal thresholds for filtering through exploration of the underlying characteristics of the data. The query and filtering interface enables the geneticist to quickly test different genetic hypotheses (recessive, dominant, de novo) in multisample setups and aggregates available information at the gene and variant level, facilitating the manual revision of candidate gene lists.

**Prediction of a phosphorylation network in *Arabidopsis thaliana*.**

Thanh Hai Dang, Stefan Naulaerts, Alain Verschoren and Kris Laukens\*

Biomedical informatics research center Antwerpen (biomina), University of Antwerp, Belgium

Contact: kris.laukens@ua.ac.be

**Background:** In contrast to other model organisms our understanding of system-wide phosphorylation networks in higher plants is rather limited. Whereas we do know that the kinome of the model plant *Arabidopsis thaliana* is more complex than the human kinome, we do not have sufficient positive substrate data to generate computational substrate prediction models for individual plant kinases. In the work presented here we employ insights in phosphorylation in well-studied organisms, functional information and the limited plant phosphorylation data available to generate a large-scale candidate phosphorylation site-specific kinase-substrate network in *Arabidopsis thaliana*.

**Methods:** We develop an integrative computational approach, for the reconstruction of a phosphorylation site-specific kinase-substrate interaction network in the target organism *Arabidopsis thaliana*. The core of the method is a Conditional Random Field-based prediction model. Kinase-specific phosphorylation prediction models from well-studied (“reference”) organisms are transferred to the “target” organism through orthology mapping, and merged with global plant phosphorylation prediction models.

**Results:** The resulting predictions are significantly enriched with confirmed phosphorylation sites and confirmed kinase-substrate interactions, respectively. Gene ontology analysis demonstrates significant functional coherence between predicted kinases and their putative substrates. This functional coherence is further used to filter the network to obtain a small high-confidence predicted network that is interpreted with regards to literature evidence. Statistical, functional and literature-based analyses of the predicted network demonstrate that the prediction and orthology mapping method yields a biologically relevant network.

**Conclusions:** The resulting predicted *Arabidopsis* network is a valuable resource for further biological and/or computational analyses and will further expand our current understanding of *Arabidopsis thaliana* complexes and pathways.



## **Spatial Clustering of Protein Binding Sites for Template Based Protein Docking**

Anisah W Ghoorah <sup>\*1</sup>, Marie-Dominique Devignes<sup>2</sup>, Malika Smail-Tabbone<sup>3</sup>, David W Ritchie<sup>1</sup>

<sup>1</sup>INRIA, <sup>2</sup>CNRS, <sup>3</sup>Nancy Université, LORIA, 615 rue du Jardin Botanique, 54600 Villers-lès-Nancy, France

Contact: [anisah.ghoorah@inria.fr](mailto:anisah.ghoorah@inria.fr)

**Background:** In recent years, much structural information on protein domains and their pair-wise interactions has been made available in public databases. However, it is not yet clear how best to use this information to discover general rules or interaction patterns about structural protein-protein interactions. Improving our ability to detect and exploit structural interaction patterns will help to guide docking-based predictions of the 3D structures of unsolved protein complexes in order to provide a better 3D picture of protein-protein interactions and disease mechanisms. This contribution presents KBDOCK, a 3D database system for spatially clustering protein binding sites and for performing template-based (knowledge-based) protein docking.

**Methods:** KBDOCK integrates protein domain-domain interaction (DDI) information from 3DID, Pfam domain family classification, and structural information from the PDB. For each Pfam family, KBDOCK superposes and spatially clusters hetero DDIs in order to identify and store domain family-level binding sites. In order to find a docking template for a given pair of query domains, KBDOCK retrieves all DDIs involving the query Pfam families, and it outputs a proposed docking model for each distinct pair of stored binding sites. If no such pair-wise template exists, KBDOCK can often still propose a small number of candidate binding sites for each individual query domain, in order to focus and constrain docking simulations.

**Results:** KBDOCK provides a straight-forward way to analyse the spatial distributions of protein binding sites at the Pfam family level. We find that most Pfam domain families have up to four hetero binding sites, and nearly 70% of all domain families have just one hetero binding site. This supports the notion that domain binding sites are often re-used in different DDIs. The utility of our approach for template-based docking is demonstrated using 73 complexes from the Protein Docking Benchmark. Overall, 45 out of 73 complexes may be modelled by direct homology to existing domain interfaces, and key binding site information is found for 24 of the 28 remaining complexes. These results show that KBDOCK can often provide useful information for predicting the structures of unknown protein complexes.

**Conclusions:** KBDOCK provides a systematic way to store and analyse the 3D structures of protein domain binding sites. KBDOCK can be used to find automatically homologous hetero DDIs with which to model the unknown 3D structure of given protein complex. KBDOCK is publicly available at <http://kbdock.loria.fr/>.

**An integrated workflow for robust alignment and simplified quantitative analysis of NMR spectrometry data**

Trung Nghia Vu\*, Dirk Valkenburg, Bart Goethals, Kris Laukens

Biomedical Informatics Research Center Antwerp (biomina), University of Antwerp, Belgium

Contact: TrungNghia.Vu@ua.ac.be

**Background:** Nuclear magnetic resonance spectroscopy (NMR) is a powerful technique to reveal and compare quantitative metabolic profiles of biological tissues. However, chemical and physical sample variations make the analysis of the data challenging, and typically require the application of a number of preprocessing steps prior to data interpretation. For example, noise reduction, normalization, baseline correction, peak picking, spectrum alignment and statistical analysis are indispensable components in any NMR analysis pipeline.

**Methods:** We introduce a novel suite of informatics tools for the quantitative analysis of NMR metabolomic profile data. The core of the processing cascade is a novel peak alignment algorithm, called hierarchical Cluster-based Peak Alignment (CluPA). The algorithm aligns a target spectrum to the reference spectrum in a top-down fashion by building a hierarchical cluster tree from peak lists of reference and target spectra and then dividing the spectra into smaller segments based on the most distant clusters of the tree. To reduce the computational time to estimate the spectral misalignment, the method makes use of Fast Fourier Transformation (FFT) cross-correlation. Since the method returns a high-quality alignment, we can propose a simple methodology to study the variability of the NMR spectra. For each aligned NMR data point the ratio of the between-group and within-group sum of squares (BW-ratio) is calculated to quantify the difference in variability between and within predefined groups of NMR spectra. This differential analysis is related to the calculation of the F-statistic or a one-way ANOVA, but without distributional assumptions. Statistical inference based on the BW-ratio is achieved by bootstrapping the null distribution from the experimental data.

**Results:** The performance was evaluated using a Wine dataset and a Huntington dataset. Correlation maps, spectral and grey scale plots show clear improvements in comparison to other methods, and the down-to-earth quantitative analysis works well for the CluPA-aligned spectra.

**Conclusions:** The workflow is unique in the way the subsequent steps are ordered. Because of rearranging the order of processing steps, simple and robust methods could be adopted for the workflow. The whole workflow is embedded into a modular and statistically sound framework that is implemented as an R package called “speaq” (“spectrum alignment and quantitation”), which is freely available from <http://code.google.com/p/speaq/>.

## Systems Biology Analysis of Tyrosine Kinase Inhibitor Target Profiles in Leukemia

Uwe Rix, Keiryn L. Bennett, Giulio Superti-Furga, Jacques Colinge\*

CeMM, Austria

Contact: jcolinge@cemm.oeaw.ac.at

**Background:** To be able to understand the mechanism of action of drugs, predict their efficacy, and anticipate their potential side-effects is an important goal. In diseases where the genetic background of patients modulates treatment response, it might allow personalizing the therapy. Chemical proteomic methods, where a compound is immobilized and binding proteins are affinity purified and identified by mass spectrometry, can measure drug-protein interactions and reveal complex tyrosine kinase inhibitor (TKI) target spectra comprising 5-50 kinases that might each be involved in multiple pathways. Hence, TKIs used in patient therapy induce global changes in cells, which require systems approaches. We analyzed four TKI profiles (dasatinib, bosutinib, nilotinib, bafetinib) in Philadelphia positive acute lymphoblastic leukemia (Ph+ ALL) cell.

**Methods:** Protein-drug interaction lists were obtained by chemical proteomic methodologies. After elimination of nonspecific interactors, drug targets were mapped onto a human interactome obtained integrating several public interaction databases. Mass spectrometry data (spectral count) provided a rough estimate of drug affinity. By means of a diffusion method over the interactome (random walk with restart), we determined a subset of the interactome that was significantly influenced by each drug. Using precise information regarding gene deletions in two distinct Ph+ ALL cell lines, we similarly determined subsets of the interactome where deleted genes and the fusion protein BCR-ABL were likely to cause perturbation. The comparison of the two areas of influence, computing a correlation score, yielded a drug efficacy measure. IC50s were determined experimentally for each cell line (Z-119, BV-173, SUP-B15) and drug pair as well as for a LIN knock down (LIN shRNA in SUP-B15).

**Results:** Predicted drug efficacies and IC50 measures were in reasonable agreement and, in particular, validated the surprising results that bosutinib, a potent and promiscuous TKI, was the weakest compound for Ph+ ALL. Dasatinib was predicted to be the strongest, which was confirmed by the experiments. Knocking down in silico the important target LIN, predicted a strong impact on dasatinib efficacy, which was also confirmed by experiments.

**Conclusions:** We have already shown diffusion methods potential to predict drug side-effects and patient genetic background dependent drug response in chronic myeloid leukemia. This new study confirms these initial results with direct measurements. The methods presented are not specific to TKI and leukemia but can be applied broadly and constitute an important step towards the implementation of personalized medicine.

### **An Efficient Algorithm to Perform Multiple Testing in Epistasis Screening**

Francois Van Lishout\*, Tom Cattaert, Jestinah M. Mahachie John, Elena Gusareva, Victor Urrea, Isabelle Cleynen, Emilie Théatre, Benoît Charlotiaux, Alex Kvasz, M. Luz Calle, Louis Wehenkel, Kristel Van Steen

Systems and Modeling Unit, Montefiore Institute, University of Liège, 4000 Liège, Belgium AND Bioinformatics and Modeling, GIGA-R, University of Liège, 4000 Liège, Belgium

Contact: f.vanlishout@ulg.ac.be

**Background:** Research in epistasis or gene-gene interaction detection for human complex traits has grown exponentially over the last few years. It has been marked by promising methodological developments, improved translation efforts of statistical epistasis to biological epistasis and attempts to integrate different omics information sources into the epistasis screening to enhance power. The quest for gene-gene interactions poses severe multiple-testing problems. In this context, the maxT algorithm is one technique to control the false-positive rate. However, the memory needed by this algorithm rises linearly with the amount of hypothesis tests. In main-effects detection, this is not a problem since the memory required is thus proportional to the number of SNPs. In contrast, gene-gene interaction studies will require a memory proportional to the squared amount of SNPs. A genome wide epistasis would therefore require terabytes of memory. Hence, cache problems are likely to occur, increasing the computation time.

**Methods:** In this work we present a new version of maxT, requiring an amount of memory independent from the number of genetic effects to be investigated. This algorithm was implemented in C++ in our epistasis screening software MB-MDR-2.6.2 and compared to MB-MDR's first implementation as an R-package (Calle et al., Bioinformatics 2010). We evaluate the new implementation in terms of memory efficiency and speed using simulated data. The software is illustrated on real-life data for Crohn's disease.

**Results:** The sequential version of MBMDR-2.6.2 is approximately 5,500 times faster than its R counterparts. The parallel version (tested on a cluster composed of 14 blades, containing each 4 quad-cores Intel Xeon CPU E5520@2.27 GHz) is approximately 900,000 times faster than the latter, for results of the same quality on the simulated data. It analyses all gene-gene interactions of a dataset of 100,000 SNPs typed on 1000 individuals within 4 days. Our program found 14 SNP-SNP interactions with a p-value less than 0.05 on the real-life Crohn\_s disease data.

**Conclusions:** Our software is able to solve large-scale SNP-SNP interactions problems within a few days, without using much memory. A new implementation to reach genome wide epistasis screening is under construction. In the context of Crohn's disease, MBMDR-2.6.2 found signal in regions well known in the field and our results could be explained from a biological point of view. This demonstrates the power of our software to find relevant phenotype-genotype associations.

## **Statistical interpretation of machine learning-based feature rankings for biomarker discovery**

Vén Anh Huynh-Thu\*, Yvan Saeys, Louis Wehenkel, Pierre Geurts

Department of EE and CS & GIGA-R, University of Liège, Belgium

Contact: vahuynh@ulg.ac.be

**Background:** Univariate statistical tests are widely used for biomarker discovery in bioinformatics. These procedures are simple and fast, but can only identify variables that provide a significant amount of information about the output variable in isolation from the other inputs. When one seeks for multivariate interacting effects, one can nowadays resort to variable relevance scores provided by machine learning techniques. However, unlike the p-values returned by univariate tests, these relevance scores are usually not statistically interpretable. This lack of interpretability prevents the wide adoption of these methods by practitioners and also makes the identification of the truly relevant variables among the top-ranked ones, i.e. the determination of a relevance threshold, a very difficult task in practice.

**Methods:** In this work, we study several, existing and novel, procedures that extract relevant features from a ranking returned by a multivariate algorithm. These procedures replace the original relevance score with a measure that can be interpreted in a statistical way and hence allow the user to determine a significance threshold in a more informed way. Most of these methods exploit a resampling procedure to estimate the FDR or FWER among the k top-ranked features, for increasing value of k. Just like for standard univariate tests, the user can then choose a threshold on this new measure depending on the risk he/she is ready to take when deeming that all features above this threshold are relevant.

**Results:** Experiments are performed on several artificial and real microarray datasets. Most of the methods provide a statistical score, which greatly improve the interpretability of the original machine learning-based importance score. The different procedures differ greatly in terms of computing times and the tradeoff they achieve in terms of false positives and false negatives. Our experiments also highlight that the common approach to this problem, i.e. selecting the top k features minimizing some cross-validated error, is not a good practice in general, as it typically leads to the selection of several irrelevant features.

**Conclusions:** The proposed measures greatly help in the extraction of truly relevant features from a ranking derived from a multivariate approach and should thus be of great interest for the practitioners. The choice of a particular method among them will depend on computational considerations and on the precise tradeoff one wants to achieve in terms of false positives and false negatives when selecting features.

## **MSCOMPARE- Data Processing Framework for Quantitative Processing of Label-Free LC-MS data**

Peter Horvatovich

University of Groningen, The Netherlands

Contact: [p.l.horvatovich@rug.nl](mailto:p.l.horvatovich@rug.nl)

**Background:** Liquid chromatography coupled to mass spectrometry (LC-MS) is a well-established analysis technique used for comprehensive comparative profiling of pre-classified sets of complex proteomics samples. Quantification accuracy of data processing workflow plays an important role in these type studies, for example to find reliable biomarkers which can distinguish between samples obtained from healthy and diseased individuals.

**Methods:** Many research groups have developed data processing pipelines for label-free LC-MS data such as MZmine, TOPP, and SuperHirn. These pipelines implement different feature detection/quantification and feature alignment/matching modules. Until now, it was only possible to use these programs by installing them individually on local computers, and it was not possible to assess, which pipeline or which combination of feature detection/quantification and feature alignment/matching modules provide the most accurate quantification of particular type of LC-MS data. We have coupled msCompare with a toolbox containing validated statistical methods such as Nearest Shrunken Centroids, PLS-DA, PCA-DA to select set of discriminating compounds. We have integrated msCompare and the statistical toolbox in Galaxy infrastructure program and coupled them with help of Annotated Putative peptide Markup Language format.

**Results:** The developed framework is able to explore all possible combinations between the main modules of the different programs. We implemented a statistical method to assess and find the performance of the different modules combinations using data obtained from sample sets spiked at different levels. The statistical tool box with double cross-validation enables the selection of discriminating compound set.

**Conclusion:** Galaxy framework can be used to develop complex data intensive processing workflow for proteomics LC-MS data and apply them using simple web interface to analyse LC-MS. We are further developing this workflow to annotate quantitative information with peptide identification and to transfer the developed workflow in other infrastructure programs such as Data Analysis Framework, which provides access to large parallel computational facility such as grid and clusters and allow enhanced data management.

## Abstracts Oral Presentations Day 2

### **Probic: Simultaneously detecting coexpression modules and their regulatory patterns**

Yan Wu\*, Lieven Verbeke, Carolina Fierro, Jan Fostier, Kathleen Marchal

CMPG/Bioinformatics, Dep Microbial and Molecular Systems, K.U.Leuven, Belgium

Contact: yan.wu@biw.kuleuven.be

**Background:** Approaches to explore transcriptional regulation are customarily based on using clustering/biclustering to search for condition dependent coexpression, followed by motif detection (using de novo search methods or by screening for known motifs). However, a module is ill defined (that is its size in number of genes is very much parameter dependent). By simultaneously searching for coexpressed genes that also contain an overrepresented motif (motif set) we can focus the search on those modules that are truly coregulated. Therefore, we developed ProBic-II, a tool to simultaneously search for coexpression modules and their regulatory motifs.

**Methods:** In this work, we present ProBic-II, an iterative and efficient framework for predicting the regulatory modules in large gene expression compendia together with their regulatory motifs. The key of our framework is a data integration strategy relying on the Probabilistic Relational Models (PRMs), which takes the advantage of Bayesian integration of multiple data types. ProBic-II optimizes the combined task of learning motifs and their motifs in an iterative way (EM based strategy). First, ProBic-II discovers tightly co-expressed modules (biclusters). Then, genes in these biclusters are used as query to retrieve regulation patterns. Updated gene list are again used as seeds for coexpression modules. These alternate steps end when reaching a stable number of genes both in biclusters and regulation patterns.

**Results:** We applied ProBic-II on a large scale Escherichia coli (E.coli) compendium together with regulatory motif data obtained by screening the whole genome with known motifs extracted from RegulonDB. To demonstrate the effectiveness of our approach, we validated our method by predicting interactions for FNR.

**Conclusions:** In this study we developed ProBic-II as the data integration framework for a combination of publicly available microarray data and regulatory motif data from E.coli.

## Visualizing genotype-phenotype relationships across cell cycle and evolutionary time scales

Maria Secrier\*, Reinhard Schneider

European Molecular Biology Laboratory, Heidelberg, Germany

Contact: secrier@embl.de

**Background:** Highly time-dependent biological processes like the cell cycle pose interesting questions in terms of visual representation of temporal events and the overall understanding of dynamic aspects of regulation. We look at the projection of the phenotypic space of cell division defects unto the temporal landscape of cell cycle regulation and then place it in the larger context of evolution. A visualization tool to help capture these aspects is presented.

**Methods:** Phenotypic transitions between cell populations upon knockdown are represented as color-coded arcs that can be traced throughout the time course. In parallel, the GO term network highlights terms corresponding to those genes whose suppression is causing a phenotypic transition at a particular time point. The GO representation can change between biological process, molecular function and cellular component. For the evolutionary analysis, peaks of protein activity throughout the cell cycle are plotted as bar charts, with arcs representing shared orthologs between any other two species visualized in parallel. The visualization tool was developed in Processing (<http://processing.org/>) and it supports bar plot, connecting arcs and heatmap view.

**Results:** The temporal succession of phenotypes in cell populations upon knockdown of genes essential to cell division is described. We analyze phenotypic transitions within cell populations, as observed by time-lapse microscopy of HeLa cells that show mitotic effects upon selective gene silencing. By visualizing patterns of these time-driven events in relation to overrepresented GO terms, we can get substantial insight into the key triggers of phenotypic changes and generate hypotheses about gene co-activation/co-regulation or participation in the same pathway.

To get an evolutionary perspective of timing in the cell cycle, we analyze how transcriptional activation levels compare in different organisms. This allows us to understand how the human homolog landscape maps to the cell cycle events. By comparing the degree of conservation of genes that are being actively transcribed during the cell cycle, one can identify temporal hotspots of novel activation events in human.

**Conclusions:** The visualization tool presented helps create a time-encoded mapping of genotype-phenotype interactions in the context of the cell cycle. This method is nevertheless applicable to any datasets where time-resolved connections between phenotypes, pathways or other variables of interest are questioned. Such representations allow for a better interpretation of how a system's dynamics can shape the morphology and interaction topology of the cell. Additionally, they may enable potential linking of diseases that have common regulators.



**Detection of genes essential for growth of respiratory pathogens.**

Aldert Zomer\*, Peter Burghout, Stefan de Vries, Hester Bootsma, Jeroen Langereis,  
Hendrik Stunnenberg, Peter Hermans and Sacha van Hijum

Centre for Molecular and Biomolecular Informatics, Nijmegen Centre for Molecular  
Life Sciences, Nijmegen, The Netherlands

Contact: a.zomer@cukz.umcn.nl

**Background:** Respiratory tract infections are a leading cause of global mortality and morbidity. For instance, It has been estimated by the WHO that annually 4-5 million people die of pneumonia. Infections that rarely lead to death include sinusitis and otitis media, and are the second most common disease of childhood after upper respiratory infection in developed countries. Infection by and growth of a respiratory pathogen is a complex process dependent on a number of essential pathways, of which members could form ideal candidate targets for drug design and/or vaccine development.

**Methods:** To identify microbial genes essential for growth of respiratory pathogens, we have used an insertion knockout strategy and developed a bioinformatics tool that allows rapid identification of disrupted genes. This method employs the next generation sequencing method Tn-Seq to generate footprints of bacterial transposon mutant libraries. Genes that are not detected in the knockout library are likely essential for growth if a sufficiently large knockout library is used. To identify shared essential pathways in bacterial species we have used statistical analysis, pathway analysis and functional category enrichment methods to determine these ideal candidates for drug design and vaccine development.

**Results:** In *S. pneumoniae* and two other respiratory pathogens we found that roughly 10 % of all genes is essential, similar to what has been found in classical knockout studies in literature. Most of these genes encode functions involved in transcription, translation or replication, however for several genes encoding for hypothetical proteins we also have not been able to generate knockouts in the used conditions. The products of these genes possibly play a role in essential processes in the cell and could form novel candidate targets for drug design and/or vaccine development

**Conclusions:**

High throughput screening of essential genes is feasible using GSF. A large knockout library is preferred to decrease the number of false positives. Genes encoding for orthologous proteins in all three species have been found to be essential, including hypothetical proteins, these could potentially play a role in critical processes in the cell.

## **Bayesian Inference of Protein Complex Modules from Affinity Purification Mass Spectrometry Data**

Alexey Stukalov\*, Florian Breitwieser, Jacques Colinge

Research Center for Molecular Medicine, Austria

Contact: astukalov@cemm.oeaw.ac.at

**Background:** Biological pathways and protein complexes are essential components of the organization of cells and hence they provide a very useful “reference system” to study changes and deregulation in biomedical research. Affinity Purification Mass Spectrometry (AP-MS) allows researchers to analyze chosen pathways or an ensemble of related protein complexes by measuring physical interactions between bait and prey proteins. The bioinformatic analysis of such datasets, typically involving 20-300 bait proteins, is challenging because many more binding proteins (preys) are detected than there are baits, and conventional methods based on dense subgraph detection have limited applicability.

**Methods:** We introduce BI-MAP (Bayesian Inference of protein Modules A Posteriori), a novel method for protein complex modules prediction based on a Bayesian approach. The method works by inferring chessboard biclustering of AP-MS dataset that directly corresponds to the modular organization of protein complexes. To achieve high sensitivity and accuracy BI-MAP takes into account both quantitative and topological components of AP-MS dataset and thus compensates for the limited number of baits used. We also introduce a new concept of module stability index that makes large and noisy datasets amenable to analysis by BI-MAP, which then can extract only stable modular structures.

**Results:** We have applied BI-MAP method to publicly available AP-MS datasets of different complexity: TIP49a/b dataset (27 baits and 125 prey proteins, Sardiù et al.) and autophagy dataset (67 baits and 2550 preys, Behrends et al.). While showing good agreement with the previous reports, BI-MAP was able to identify potential new components of protein complexes among proteins regarded as “background” interactors in previous studies. To further evaluate BI-MAP performance we have introduced a comprehensive generative model of protein complex affinity purification that can produce realistic synthetic test datasets. We used it to demonstrate that BI-MAP compares very favorably with the competing tools and study the influence of different factors (such as replicate experiments availability or protein complex size) on the quality of predictions.

**Conclusions:** We showed that BI-MAP has achieved a performance level permitting reliable application to essentially any AP-MS dataset to decompose complex protein interaction maps into smaller structures that are associated with a biological function. These decompositions can be further exploited to describe the dynamics of interaction maps under different conditions such as disease stages or drug treatments. BI-MAP implementation and test data are made freely available.

### **Data-analytical strategies for enrichment-based genome-wide DNA-methylation profiling by NGS**

Tim De Meyer\*, Geert Trooskens, Klaas Mensaert, Evi Mampaey, Simon Denil,  
Peter Pipelers, Wim Van Criekinge

Dept. Mathematical Modelling, Statistics and Bioinformatics, Ghent University,  
Belgium

Contact: Tim.DeMeyer@UGent.be

**Background:** Over the last decade, DNA-methylation research has shifted from a gene-based approach to genome-wide analyses. DNA-methylation, featured by the enzymatic methylation of cytosines in a predominantly CpG-dinucleotide context, is an epigenetic process that is tightly associated with gene expression regulation. A novel generation of methodologies has enabled researchers to profile DNA-methylation in a genome-wide manner, e.g. by Methyl-Binding Domain (MBD)-based affinity purification followed by NGS (MBD-seq). While MBD-seq provides an excellent combination of sensitivity and cost-efficiency, it is featured by a set of bioinformatics and statistical challenges that complicates the subsequent data-analysis. The data-analysis pipeline for quantitative NGS applications typically consists of quality control, sequence mapping, data summary, data normalization and statistical analysis. Several of these steps require specific solutions for MBD-seq data:

- Quality control is difficult yet necessary as sensitivity and specificity of enrichment procedures may vary.
- Data summary is complicated by the lack of a functional unit for DNA-methylation, cf. the exon as unit for RNA-seq data summary.
- Most data normalization procedures assume that the overall profiles are similar between samples, an assumption that is invalid for DNA-methylation.
- The identification of significant enrichment is usually based on a Poisson background model. This model has several restrictions, resulting in suboptimal power.

Therefore, we aimed at developing tailor-made solutions for each of these challenges.

**Methods:** Captured fragments were paired-end sequenced (Illumina GAIIX). Sequenced reads were mapped on the human genome with BOWTIE. R, Perl, Java and MySQL were used to implement the different solutions.

**Results:** We could demonstrate that the CpG-density profile of the sequenced fragments provides a solid basis for quality control, including the evaluation of sensitivity and specificity. A Map of the Human Methylome was constructed based on a large collection of MBD-seq profiles. This map consists of putatively independently methylated genomic regions, i.e. Methylation Cores (MCs), that can be used for data summary. For data normalization, a procedure called "Massively Enriched Loci Normalization" (MELON) was developed, based on the assumption that there exists a set of massively enriched loci of which the degree of DNA-methylation is similar between samples. A novel statistical framework, that provides higher power and sensitivity than the standard Poisson model has (partially) been developed, and can also be used for other enrichment based NGS applications.

**Conclusions:** An optimized pipeline for high-quality MBD-seq data-analysis has largely been developed and implemented.

### **SABIO-RK: A Resource for Biomedical Research**

Renate Kania\*, Ulrike Wittig, Martin Golebiewski, Maja Rey, Lei Shi, Lenneke Jong, Enkhjargal Algaa, Andreas Weidemann, Meik Bittkowski, Elina Wetsch, Isabel Rojas, Wolfgang Müller

Scientific Databases and Visualization, Heidelberg Institute for Theoretical Studies, Germany

Contact: [renate.kania@h-its.org](mailto:renate.kania@h-its.org)

**Background:** Biochemical data in the scientific literature is available sparsely in structured and standardized format. To better understand the processes taking place in biological systems, detailed qualitative and quantitative information about single biochemical reactions is essential. For the quantitative analysis of biochemical reactions by modeling their enzyme kinetics, reliable kinetic data for the individual reaction steps are essential. SABIO-RK was developed to store these kinetic parameters in a structured and standardized form and to facilitate the exchange of kinetic data between experimentalists and modelers, and thereby to support the setup of quantitative computer models.

**Methods:** SABIO-RK is a web application based on a relational database. This makes sense in order to maintain data integrity. However, the relation between data items within SABIO-RK is quite complex, leading to complex queries with poor performance. We remedy this by using inverted file indexing and aggressive caching, dramatically reducing the number of queries that hit the database. This increases speed of query processing, up to a factor of 50 in some cases. Our data entry is performed via an input interface, which enables us to have a distributed group of students perform data entry. The students' entries are then checked and (potentially) modified to meet SABIO-RK data quality standards, before they are made visible in the SABIO-RK database. SABIO-RK has been growing over the years. The majority of the code uses a JSP/Spring/iBatis stack. New parts of the code are written in Grails, a Groovy language based framework that allows seamless integration with JAVA applications.

**Results:** As of October 2011, data from over 3400 publications have been curated and are stored in the database. And what it makes interesting for Biomedical Research is that almost 20% of the entries are related to Human enzymes. The data in SABIO-RK can be accessed either manually via a web-based search interface or programmatically via SOAP and REST-style web services. We are constantly working on facilitating access for biologists as well as developers and further improving efficiency

**Conclusions:** SABIO-RK is a curated database storing kinetic data of biochemical reactions. The database is being populated with experimentally measured kinetic parameters manually extracted from scientific literature or directly submitted from laboratories. Additionally the data is connected to the assay conditions and if available the corresponding kinetic law type and rate equation of the reaction.

<http://sabiork.h-its.org/>

## Abstracts Flash Oral Presentations A and Poster Session A

### A1: eBiomics: a Bioinformatics e-Learning Environment for Biologists

Patrick Koks\*, Pascale Berthault, Guy Bottu, Jacques van Helden, Jean-Pierre Kraehenbuhl, Frédérique Lisacek, Grégoire Rossier, Jean Sylvestre, Jack Leunissen

Laboratory of Bioinformatics, Wageningen University, The Netherlands

Contact: patrick.koks@wur.nl

**Background:** The number and volume of biological data sets grows rapidly. Well-trained bioinformaticians are needed to cope with the challenge of analysing these data. In order to educate more people and promote good practices in bioinformatics, we present eBiomics, an e-learning platform in bioinformatics for biologists. The system is targeted at post-graduate students and scientists interested in -omics (genomics, proteomics, metabolomics) or systems biology. eBiomics is a didactic guide for an extensive range of on-line databases and tools commonly referred to in -omics applications.

**Methods:** eBiomics offers "the biologist" an e-learning environment where he or she can study some of the most relevant bioinformatics topics in common research domains like proteomics, metabolomics and genomics. Central in eBiomics are "Conceptual Flowcharts" that provide you with a comprehensive overview of both the general methods and many specific tools in a research field.

Protocols provide a linear learning sequence through a common bioinformatics task, while Case Studies show a specific (article-based) solutions to complex tasks. Furthermore, in the Resources section we provide reviews of many different tools and hands-on Exercises to learn how to work with them. Taken together, this enables users to learn about many tools and their applications.

**Results:** The eBiomics platform serves both as a knowledge base, a collection of tool descriptions and applications, and an e-learning environment where people can study bioinformatics. We offer three complementary methods for e-learning: Case-Based-Learning, Learning-By-Example and Article-Based-Learning. Together these provide complete self-study sessions.

**Conclusions:** The eBiomics website is an advanced e-learning system to educate the next generation of bioinformaticians. The system is freely available to the public at [www.ebiomics.org](http://www.ebiomics.org). The website is well suited for self study, but could as easily be embedded by a teacher in existing courses. The developers of the eBiomics platform are committed to further development and maintenance in the future.

### **A3: iSNP: An Integrated, Automatically Updated SNP Database Server Over Web**

Ceyhun Gedikoğlu<sup>1</sup>, Levent Çarkacıoğlu<sup>1</sup>, Yeşim Aydın Son<sup>\*1,2</sup>

<sup>1</sup>Bioinformatics Graduate Program, Middle East Technical University, 06800, Ankara, Turkey; <sup>2</sup>Health Informatics Department, Middle East Technical University, 06800, Ankara, Turkey

Contact: yesim@metu.edu.tr

**Background:** Single Nucleotide Polymorphisms (SNPs) are the most frequently observed genomic variations, and are subtle changes in human genome where only one DNA base (nucleotide) differs in the genomic sequence. As genotyping millions of SNPs in a short time and much lower cost is now possible with the microarray and advanced sequencing technologies, Genome Wide Association Studies (GWAS) of SNPs and SNPs as genomic biomarkers are becoming more popular with the potential to aid for population genetic studies and for identifying genetic variations underlying complex diseases. National Center for Biotechnology Information (NCBI)'s current SNP Variation database (dbSNP build 132) holds about 20 million validated SNPs out of estimated 30 million potential SNPs within the human genome. The vast majority of SNPs are shared between populations. The International HapMap Project (HapMap3) have genotyped around 1.4 million consensus SNPs between 11 different populations consistent with previously reported statistics from the earlier phases of the study. One of the current research areas that draw attention in bioinformatics field is the challenge of identifying genetic variations that are the molecular basis of common diseases, such as neurodegenerative, immunological, and cardiovascular disease, diabetes and cancers. Our understanding of the genetic etiology of human disease is still limited because of the enormous number of genetic variations on the human genome, as well as the complex interplay of multiple genes and environmental factors underlying disease.

**Methods:** Information related to SNPs and their associated meta-data can be found individually in many public databases. These databases serve data in a non-uniform format. In order to provide uniformity of SNP data and its associated meta-data, we have build an integrated database structure called iSNP where data from NCBI's dbSNP and Entrez Gene, HapMap, UCSC, Polyphen, Pathway Commons, GAD and GeneRIF-DO is collected. iSNP has the capability of updating the information by automatically synchronizing with the online accessible SNP databases listed above.

**Results and Conclusions:** The java-based METU-SNP (<http://metu.edu.tr/~yesim/metu-snp.htm>) desktop application developed by our group provides all-in-one genome wide association analysis of SNP-disease relations. iSNP database presented here will be utilized by the METU-SNP application which will be available through web. So, iSNP will provide a fast, error free, up to date and a maintainable database for the METU-SNP application and can also be integrated to other applications for ultimately aiding SNP biomarker discoveries and development of personalized medicine approaches.

#### **A4: Exposing WikiPathways as Linked Open Data**

Andra Waagmeester\*<sup>1</sup>, Helena F. Deus<sup>2</sup>, Chris T. Evelo<sup>1</sup>

<sup>1</sup>Department of Bioinformatics - BiGCaT, Maastricht University, <sup>2</sup>Digital Enterprise Research Institute, National University of Ireland, Galway

Contact: andra.waagmeester@bigcat.unimaas.nl

**Background:** Biology has become a data intensive science. Discovery of new biological facts increasingly relies on the ability to find and match appropriate biological data elements. For instance for functional annotation of genes of interest or for identification of pathways affected by over-expressed genes. Pathways are a convenient, easy to interpret way to describe known biological interactions. Functional and pathway information about genes proteins is typically distributed over a variety of heterogeneous databases and literature. WikiPathways provides community curated pathways. WikiPathways users integrate their knowledge with facts from the literature and biological databases. The curated pathway is then reviewed and possibly corrected or enriched. Different tools (e.g. Pathvisio and Cytoscape) support the integration of WikiPathways-knowledge for additional tasks, such as the integration with personal data sets. Data from WikiPathways is increasingly also used for advanced analysis where it is integrated or compared with other data. Currently, integration with data from heterogeneous biological sources is mostly done manually. This can be a very time consuming task because the curator often first needs to find the available resources, needs to learn about their specific content and qualities and often spends a lot of time to technically combine the two.

**Methods:** Semantic web and Linked Data technologies eliminate the barriers of database silos by relying on a set of standards and best practices for representing and describing data. The architecture of the semantic web relies on the architecture of the web itself for integrating and mapping universal resource identifiers (URI), coupled with basic inference mechanisms to enable matching concepts and properties across data sources. Semantic Web and Linked Data technologies are increasingly being successfully applied as integration engines for linking biological elements. Exposing WP content as Linked Open Data to the Semantic Web, might enable rapid, semi-automate integration with a rich, expansible set of biological resources available from the linked open data cloud, it also allows really fast queries of WikiPathways itself.

**Results and Conclusions:** We have harmonised WP content according to a selected set of vocabularies (Biopax, ChEMBL, etc), common to resources already available as Linked Open Data.

WP content is now available as Linked Open Data for dynamic querying through a SPARQL endpoint: <http://semantics.bigcat.unimaas.nl:8000/sparql>.

## **A5: Implementation of the Protein Fluorescence And Structural Toolkit (PFAST)**

Cynthia N. Prudence\*, Yana K. Reshetnyak

Physics Department, University of Rhode Island, 2 Lippitt Rd., Kingston, RI, 02881, USA

Contact: cprudence@my.uri.edu

**Background:** Fluorescence spectroscopy is a powerful tool for investigating protein structure, conformations and dynamics, since fluorescence properties of tryptophan residues vary widely depending on the tryptophan environment in a given protein. The major goal of the application of tryptophan fluorescence spectroscopy is to interpret fluorescence properties in terms of structural parameters and to predict structural changes in a protein. One of the major obstacles in the analysis of fluorescence data lies in the complex nature of protein fluorescence. An overwhelming majority of proteins contain more than one fluorophore and therefore exhibit multi-component smooth spectra.

**Methods:** To address this ill-posed problem we have implemented two mathematically different algorithms, SIMS (SIMple fitting procedure using the root-Mean-Square criterion) and PHEQ (PHase-plot-based RESolution using Quenchers), to analyze the protein fluorescence spectra. These algorithms allow for a stable decomposition of tryptophan fluorescence spectra into at most three spectral components. We have also designed and implemented algorithms that analyze the structural parameters of the microenvironment of tryptophan residues from the atomic structures of proteins from the Protein Data Bank (PDB). These algorithms reveal a set of structural parameters, which correlates with a set of spectral parameters obtained as a result of the application of the decomposition algorithms.

**Results:** We have integrated the algorithms, introduced new programs to assign tryptophan residues to the spectral-structural classes, and created a web-based toolkit, PFAST: Protein Fluorescence and Structural Toolkit. PFAST contains three modules: 1) FCAT, fluorescence-correlation analysis tool, decomposes protein fluorescence spectra and assigns spectral components to one of five previously established spectral-structural classes; 2) SCAT, structural-correlation analysis tool, calculates the structural parameters of the microenvironment of tryptophan residues from the atomic structures of the proteins from the PDB, and assigns the tryptophan residues to one of five spectral-structural classes; and 3) the PFAST database.

**Conclusions:** FCAT and SCAT are toolkits for the calculating the spectral and structural properties of tryptophan residues in proteins. PFAST contains the first database of tryptophan fluorescence properties obtained by spectral decomposition analysis. This provides information on the spectral properties of individual tryptophan residues or clusters of nearby tryptophan residues, as well as on the assignment of the tryptophan fluorophores to one of the five spectral-structural classes



## Abstracts Flash Oral Presentations B and Poster Session B

### **B1: BioGraph: unsupervised biomedical knowledge discovery via automated hypothesis generation**

Anthony ML Liekens\*, Jeroen De Knijf, Walter Daelemans, Bart Goethals, Peter De Rijk, Jurgen Del-Favero

Universiteit Antwerpen

Contact: [anthony@liekens.net](mailto:anthony@liekens.net)

**Background:** In the high-throughput research to discover biomedical targets, for example in the discovery of disease genes, biomedical researchers have witnessed two important problems of exponential growth over the last few years. First, high-throughput technologies yield outputs that are up to a million-fold higher than ten years ago and we need to somehow sift through this data to find relevant targets for further research. Secondly, the available information on which we can base this target selection is also exponentially growing. Moreover, this knowledge is distributed among the literature and diverse, heterogeneous databases. There's no way for a researcher to keep up with even the most specific area of interest anymore.

**Methods:** In order to cope with these problems of exponential growth we have constructed the BioGraph software platform (<http://www.biograph.be>) [1]. BioGraph allows us to integrate knowledge from various biomedical databases into a common knowledge network. Consequently we use data mining on this network to discover new hidden knowledge by generating a map of paths from biomedical research subjects (e.g., a disease) to potential targets (e.g., susceptibility genes). Assessments of these hypotheses/plausibilities and specificities to their respective source and targets allows for various applications in the identification of promising research targets. Recent developments are in the integration of text mining results in the approach and the development of interpretation methods of expression studies.

**Results and Conclusions:** BioGraph allows for the discovery of relevant targets for specific research contexts, ignoring unspecific noise. These discoveries are supported by rationale to further the functional research on such targets. This rationale is based on automatically generated hypotheses of relevant, possibly indirect paths in the integrated knowledge network, relating source and target concepts with references to the literature. In contrast with the state-of-the-art, our method is unsupervised and consequently does not require prior domain knowledge from the user, resulting in a more robust and user-friendly methodology. BioGraph has been shown to retrospectively confirm recently discovered disease genes and identify potential susceptibility genes, outperforming existing technologies for gene prioritization.

## **B2: Origin and Evolution of the Organellar Release Factor Family**

Isabel Duarte\*, Sander Nabuurs, Ramiro Magno, Martijn Huynen

CMBI, Radboud University, The Netherlands

Contact: i.duarte@cmbi.ru.nl

**Background:** Translation termination is accomplished by proteins of the Class I release factor family (RF) that recognize stop codons and catalyze the release of the newly synthesized peptides from the ribosome. Bacteria have two canonical RFs: RF1 recognizes UAA and UAG, RF2 recognizes UAA and UGA. Despite that these 2 release factor proteins are sufficient for de facto translation termination, the eukaryotic organellar RF protein family, which has evolved from bacterial release factors, has expanded considerably, comprising multiple subfamilies, most of which have not been functionally characterized or formally classified.

**Methods:** Here we integrate multiple sources of information to analyze the remarkable differentiation of the RF family among organelles. We use standard comparative genomics methods to document the origin, phylogenetic distribution and sequence structural features of the mitochondrial and plastidial release factors: mtRF1a, mtRF1, mtRF2a, mtRF2b, mtRF2c, ICT1, C12orf65, pRF1 and pRF2, and review published relevant experimental data.

**Results and Conclusions:** The canonical release factors (mtRF1a, mtRF2, pRF1 and pRF2) and ICT1 are derived from bacterial ancestors, while the others have resulted from gene duplications of a canonical release factor. These new release factor family members have all lost one or more specific motifs relevant for bona fide release factor function but are targeted to the same organelle as their ancestor. We also characterize the subset of classical release factor proteins that bear non-classical PxT/SPF tripeptide motifs, and provide a molecular model based rationale for their retained ability to recognize stop codons. Finally we systematically analyze co-evolution of RFs with the organellar genetic code. Although the presence of a release factor and the organellar genetic code tend to co-evolve, we uncover three taxa that encode an RF2 without using UGA stop codons, and one reverse scenario, where the green algae use UGA stop codons in their mitochondria without having a mitochondrial type RF2. For the latter we propose a \_stop-codon reinvention\_ hypothesis that involves the retargeting of a plastid release factor to the mitochondria.

Overall, translation termination is far from understood, but our organellar RF family systematic classification, characterization and phylogenetic profiling has shed some light on these proteins\_ molecular features, highlighting the most striking attributes of each subfamily, paving the way for imperative experimental functional studies.

### **B3: Protein regulation dynamics analysis in R**

Florian P Breitwieser\*, Alexey Stukalov, Jacques Colinge

Research Center for Molecular Medicine, Austria

Contact: fbreitwieser@cemm.oeaw.ac.at

**Background:** Monitoring quantitative changes of protein abundances and their post-translational modifications (PTM) is of major importance in biomedical research. PTMs are essential in many biological processes and are often deregulated in disease and cancer. Using mass spectrometry and isobaric peptide labeling (iTRAQ and TMT), differences in protein abundance in different samples can be measured by generating reporter ions whose intensities are compared. On the bioinformatics side, quantitative information has to be extracted and represented, and robust statistical models used to decide for regulation of proteins and post-translational modifications.

**Methods:** Technical variation is captured using noise models on a spectrum level, where accuracy of ratios depends on intensity of reporter ions. Biological variation is estimated by modeling the protein null ratio distribution. We calculate p-values to decide for protein regulation: Both the accuracy of the calculated protein abundance ratio, and its extremeness in the variation of the biological background are used. For PTM analysis, the peptides bearing the modifications must be assessed. Less data is available for individual peptides compared to proteins, and using all available information is even more important. We add a statistical layer handling spectra where one or several reporter ions are missing - data which was previously excluded.

**Results:** We have devised statistical models and implemented the package isobar for R and Bioconductor. It implements fundamental proteomics data representation in S4 classes and determines protein groups and peptides shared by multiple proteins. Protein ratios are calculated together with p-values to detect significantly deregulated proteins, which improves over often used fold-change thresholds. User-oriented L TEX and Excel quality-control and analysis reports can easily be generated via scripts. We extended models and reports for post-translationally modified proteins, handling them on the peptide level and using all spectra. We applied our tools to samples of eye fluids and in cancer research, being able to detect more differences.

**Conclusions:** The R package isobar can be used for analysis and visualization of protein expression changes, presence of protein splice variants with shared peptides and dynamics of post-translational modifications. Based on Bioconductor classes, it can be used with available packages to analyze pathways, GO term enrichment and integrate data from genomics experiments. Automated reports provide a convenient way to share data with biologists, reducing the often tedious work of generating reports. Sound statistics enable the reliable detection of significant changes.

**B4: Inheritance analysis and quality control for Next Generation Sequencing data**

Joep de Ligt\*, Lisenka E.L.M. Vissers, Christian Gilissen, Joris A. Veltman, Jayne Y. Hehir-Kwa

Department of Human Genetics, UMC St. Radboud, Nijmegen

Contact: j.ligt@antrg.umcn.nl

**Background:** Massive parallel sequencing of patient-parent trios has proven to be a successful strategy for prioritizing causal variants in rare dominant mendelian diseases. This has contributed to the implementation of whole exome sequencing into diagnostics in Nijmegen. The experimental confirmation of sample relations is essential for sample tracking in gene and variant identification studies that rely on inheritance to prioritize variants.

**Methods:** We have developed the DeNovoCheck algorithm to determine inheritance patterns of patient-parent Next Generation Sequencing (NGS) data and to identify sample relations directly from the NGS data. Our method takes into account the read depth, read errors and different inheritance models. An additional algorithm compares genotype information determined by Affymetrix 250k microarray experiments with those obtained by whole exome sequencing, for sample identification and as an independent quality control. Both algorithms have been tested on trio data from exome sequencing (Agilent SureSelect 50Mb enrichment + SOLiD4 50bp fragment sequencing) projects investigating the role of de novo mutations in intellectual disability.

**Results and Conclusions:** We present two techniques to identify samples, parent child relations and quality measures from both microarray and NGS data. We tested our method on the private variants of 20 trios and detected on average 6 mendelian errors (potential de novo variants) per exome per generation. Simulation of non-paternity of one parent increased this number of mendelian errors to 48, whereas a sample swap of the child resulted in 163 inconsistencies. When comparing exonic SNPs with microarray data (~3200 positions) a child had on average 182 discordant SNPs. The discordance between the microarray and the NGS is largely due to low quality data from one and/or both platforms. After excluding all NGS variants with a read depth less than 10 and a base quality below 10, and a confidence score on the microarray higher than 0.3, the discordance decreased to 64 (std +/-12). In the case of a sample swap the discordance between the microarray and exome data significantly increased to 958 (std +/-21).

In conclusion, microarray data can be successfully applied for sample identification and quality control. Furthermore DeNovoCheck provides accurate inheritance predictions and enables a quick and reliable detection of de novo variants in a both a research and diagnostic setting.

**B5: Reflect: an augmented browsing tool for life scientist**

Janos Binder\*, Sean O'Donoghue, Sune Frankild, Lars Juhl Jensen, Reinhard Schneider

Structural and Computational Unit, EMBL, Germany

Contact: janos.binder@embl.de

**Background:** During the past decade, customization of web browsers became a default feature. However there were only a few extensions available for biologist community and we have built a browser plugin called Reflect that augments life scientist browsing experience. By tagging an article with a single click, any user can easily access further information about a selected term in a popup window.

**Methods:** Reflect tags chemical, proteins and Wikipedia terms using Named Entity Recognition. We provide plugins to our end users those support the most widespread web browsers. This year we are empowering our software with new features like sub-cellular and tissue localization of proteins, while keeping the current abilities e.g.: sequence, structure and interaction network.

**Results:** Reflect has been widely used during the past years in-house project and in external collaborations. Also we provide an extensive API and different websites begin to incorporate its features e.g. it can switched on ScienceDirect website by highlighting keywords and we also developed a SciVerse application. Reflect service gets several thousand hits per day. We aim to keep it simple and interactive in order to support the life scientist community without any complex computational knowledge, while the users can get an overview about the selected terms.

**Conclusions:** This tool helps novel scientist to identify biological terms in a scientific article and provides brief information. We plan to extend it with additional features like better sub-cellular localization.

The software is freely downloadable at: <http://www.reflect.ws>

## Abstracts Poster Session A

### A6: Gene set analysis in the cloud

Lu Zhang\*, Shengchang Gu, Yuan Liu, Bingqiang Wang, Francisco J. Azuaje

Laboratory of Cardiovascular Research, CRP-Santé, Luxembourg

Contact: lu.zhang@crp-sante.lu

**Background:** Cloud computing offers low cost and highly flexible opportunities in bioinformatics. Existing cloud platforms, such as those offered by Amazon Web Services, provide the environment required to deploy computationally expensive algorithms and applications. Such environments allow users to configure and exploit resources on a “pay as you use” basis. Cloud computing applications are increasingly being made available for high-throughput DNA sequencing data. There is a need for publicly-available algorithms that can enable other translational biomedical research applications, such as large-scale gene set analysis of expression data.

**Methods:** We developed a cloud-based application, YunBe, which is written in Java using the MapReduce framework, for gene set analysis. To test this application, we analyzed a human liver gene expression dataset including 466 samples with 31842 transcripts and a synthetic dataset with 1000 samples and 19634 transcripts. As gene sets, we used a canonical pathway list with 880 gene sets from the Molecular Signatures DataBase (MSigDB). We compared YunBe’s execution speeds on Amazon Elastic Compute Cloud (EC2) with a desktop program and another version running on a local cluster.

**Results:** In comparison to a desktop implementation, YunBe reduced the execution time from hours to minutes in both datasets. In the case of the liver dataset, speed-ups of at least 10.9 and 24.1 times were obtained with Amazon EC2 and BGI computational cluster respectively. Major execution time improvements were also observed on the simulated dataset: 8.6 and 16.4 faster with Amazon EC2 and BGI cluster respectively. Moreover, YunBe’s running time scales with nearly linear speed-up over the desktop program performance as the number of cores increases. YunBe’s performance is aided by the highly parallel processing nature of its underlying algorithm. The application is freely accessible on AWS (Jar location: `s3n://lrcv-crp-sante/app/yunbe.jar`). Code and user’s guidelines can be downloaded from <http://tinyurl.com/yunbedownload>.

**Conclusions:** YunBe is a new open-source gene set analysis tool for the cloud. We show how, in comparison to a desktop implementation, YunBe significantly improves execution times. YunBe can accelerate pathway-based biomarker identification through inexpensive and secure distributed computing. Strong cooperation will be required to make other bioinformatics tools cloud-compliant.

## A7: Breaching the surface with HOPE

Jules Kerssemakers

CMBI, Radboud University Nijmegen Medical Centre, The Netherlands

Contact: J.Kerssemakers@cmbi.ru.nl

**Background:** The HOPE system ('Have yOur Protein Explained', <http://www.cmbi.ru.nl/hope>) aims to make detailed bioinformatics research knowledge available (and digestible) to physicians. It automatically predicts the effects of protein mutations by integrating information from leading sources in bioinformatics in a finely tuned decision system. Up till now, this system has primarily used Uniprot annotations and calculations on the protein structure or (automatically generated) homology models. After HOPE's original success, we are now broadening the scope of information sources.

**Methods:** Biological assemblies: PISA: PDBe's PISA (Protein Interfaces, Surfaces and Assemblies) service predicts biological assemblies based on information present in crystal structures. This yields a wealth of information on biological interactions at the residue level, information that is usually not taken into account when analysing mutation effects. Our new module for HOPE, presented here and currently in testing and calibration, will automatically incorporate this important information in HOPE's detailed reports by analysing the structural effects of the mutation on buried surface and the number of atomic contacts, both proxies for final binding energy. Domains: InterPro Much of a protein's function is described by its domains. The definitive domain resource is InterPro, which integrates the 11 dominant domain resources (e.g. Pfam, PANTHER, SMART and TIGR). We are now in the process of integrating the InterPro software into our HOPE analysis pipeline to tap this information, further expanding the use of high quality, but hard-to-reach information in HOPE.

**Results:** Work on both new modules is still in progress, but including PISA biological assemblies will solve several false negatives on our test set, while the InterPro annotations will add a whole, previously unused, field of 'background knowledge' into the prediction process.

**Conclusions:** Bioinformatics can add tremendous value in the medical world, but everyday physicians do not have the time to familiarise themselves with the tools or to do lengthy investigations. With HOPE, we can do this automatically, bringing state-of-art bioinformatics analysis within reach for every doctor. The presented modules are an incremental, but important step forward.

### **A8: Towards a Standard for Cooperative Interactions**

Kim Van Roey,\* Henning Hermjakob, Samuel Kerrien, Toby J. Gibson

Structural and Computational Biology Unit, European Molecular Biology Laboratory,  
Heidelberg, Germany"

Contact: roey@embl.de

**Background:** Cells must continuously monitor external and internal cues, integrate the variety of signals they perceive, and translate these inputs into proper outputs. This requires reliable and robust signal transduction, which is mediated by intricate and interlinked networks of pathways and processes that are tightly regulated. Assembly of the dynamic macromolecular complexes that modulate these pathways depends on multiple transient, low-affinity interactions, many of which are highly cooperative, with distinct binding events affecting each other either positively or negatively. Such cooperative interactions provide the dynamic plasticity that is required for cells to integrate multiple input signals, robustly and reliably transmit information, and rapidly generate appropriate responses. However, despite the central importance of cooperativity in these systems, it is missing from all current formalisms for describing molecular interactions.

**Methods:** Cooperative interaction data were collected from the literature and used to explore the information required to incorporate cooperative interactions within such formalisms. The current PSI-MI standard for molecular interactions was used as the basis for an extended standard that is able to capture this additional information.

**Results:** From our analysis of cooperative interactions extracted from the literature, we developed a draft data-exchange format and controlled vocabulary that can capture the relevant features of these interactions. Once we are confident to have the computational ability to adequately describe cooperative interactions, a relational database will be developed or adapted to store such data, and a prototype will be populated with the data extracted from the literature. The Minimum Information About a Cooperative Interaction (MIACI) will be defined and serve as a guideline for experimentalists to unambiguously describe cooperative interactions.

**Conclusions:** Together, these tools will facilitate systematic capture, comparison, exchange and verification of cooperative interaction data.



**A9: Gene expression evolution on the emergence of pathogenicity in Ascomycetes**

Aminael Sanchez-Rodriguez\*, Riet De Smet, Kristof Engelen, Qiang Fu, Yan Wu,  
Kathleen Marchal

Centre of Microbial and Plant Genetics, Department of Microbial and Molecular  
Systems. K.U.Leuven, Belgium

Contact: [aminael.sanchezrodriguez@biw.kuleuven.be](mailto:aminael.sanchezrodriguez@biw.kuleuven.be)

**Background:** The Ascomycetes form the largest phylum in the fungal kingdom. They are of special interest due to their broad spectrum of life styles including both plant and human pathogens. Several comparative genomic studies tried to explain their pathogenic potential or their ability to cause disease by studying differences in the coding potential between pathogenic and non-pathogenic Ascomycetes. From those studies it became clear that most of the protein-coding genes needed for pathogenicity were already present in an ancestor common to both pathogenic and non-pathogenic Ascomycetes. However, what remains unclear is to what extent alterations in expression behavior possibly as a result of mutations in non-coding regions tuned this ancestral coding potential to better contribute to the pathogenicity phenotype.

**Methods:** To evaluate how changes in expression of both direct orthologs and paralogs affect the origin of phenotypic traits in Ascomycetes we built an expression compendium for the non-pathogen *N. crassa* and the pathogens *M. grisea* and *F. graminearum* containing respectively datasets). We compared expression behavior across and within species and applied a collection of evolutionary models to the expression data of the largest gene families on the studied species.

**Results and Conclusions:** We found that the expression behavior of pathogenicity related genes show a significant anti-correlation to the one of their direct ortholog in non-pathogens. The speed by which expression behavior changes among paralogs match with their physical location on the genome for non-pathogenic species while in pathogens is the interaction with their host the driven force of paralogs expression changes.

### **A10: Prediction of bacterial relationships in the human microbiome**

Karoline Faust\*, J. Fah Sathirapongsasuti, Curtis Huttenhower, Jeroen Raes

Bioinformatics and (Eco-) Systems Biology, VIB-Vrije Universiteit Brussel, Belgium

Contact: karoline\_faust@yahoo.de

**Background:** Metagenomic sequencing projects are accumulating abundance data for microbial organisms in a wide variety of environments, including the human body. These data enable ecological studies of microbiota that could not be carried out previously. In macro-ecological data sets, non-random patterns of species distributions were found that reflect ecological relationships, such as the checkerboard pattern, which indicates competition. The analysis of microbial abundance data revealed similar non-random patterns for microorganisms. Recently, the Human Microbiome Consortium has compiled a massive data set of microbial sequences in up to 18 human body sites. Our goal is to predict from these data ecological relationships between microbial taxa in the human body.

**Methods:** We computed pair-wise scores of taxa relationship strength, using a variety of correlation, distance and similarity measures. In addition, we carried out a sparse linear regression to predict the abundance of a target taxon from source taxa abundances. For each scoring method and taxa pair, we computed a p-value from a background score distribution and combined method-specific p-values with Fisher's formula to obtain the final score. Scores were computed for taxa pairs residing in the same as well as in two different body sites.

**Results:** We obtained a network of bacterial relationships within and across body sites, which reproduces known microbial communities for the vagina, the gut and the sub-lingival dental plaque. The network is modular and scale-free, with hub taxa located in the oral cavity.

A functional analysis illustrated that pairs of closely related taxa with similar functional capacity co-occur within the same body area (i.e. oral cavity), but not within the same body site, whereas taxa pairs whose functional similarity is closer than expected from their phylogenetic distance tend to exclude each other.

**Conclusions:** Bacterial taxa form relationships mostly within the same body site or similar body sites (i.e. various oral sites), whereas only few relationships exist between taxa in different body areas (i.e. skin versus gut). This shows that taxa specialize to their respective body area niche. The oral hub taxa may be drivers of alternative oral communities (such as those observed for vagina and gut).

**A11: Union makes strength: Building baseline Tracks from 69 open access full human genomes**

BITS, VIB - Belgium

Stephane Plaisance\*, Mark Veugelers

Contact: stephane.plaisance@ugent.be

**Background:** Scientists can buy a ‘complete’ human genome sequence through a sequence service provider with quick delivery of data within a few weeks at a reasonable cost. ‘Complete Genomics’ (abbreviated as CG), not only performs library preparation and NGS, but also developed a very consistent annotation and analysis pipeline to deliver high quality data to its customers.

We used data from 69 open access healthy HAPMAP genomes released by CG to assemble a number of baseline ‘Tracks’ inspired from the universally adopted UCSC tables. Our Tracks are catalogs of specific features present in CG genomes, which allow substantial enrichment when analyzing private genomes, and are therefore an important resource to bring down validation costs of data obtained from newly sequenced genomes.

**Methods:** Data presented here was obtained from 69 CG ‘var files’ available from the CG ftp repository. These files report results obtained by CG after mapping reads to the hg18 (hg19 also available) human reference and calling SNVs, short-insertions, -deletions, and -substitutions. We used cgatools ‘listvariants’ and ‘testvariants’ (v1.4) to produce a global database file from the 69 ‘var files’. Various in-house developed parsing scripts were then applied to this large table to extract and reformat data for the different ‘Tracks’.

**Results:** Our current Tracks include:

- Genome regions where non-ambiguous read mapping OR calling are impossible (no-calls)
- Systematic variants’ found in a majority of CG genomes and likely resulting from reference or mapping/calling issues.
- The ‘69-genome variome’ reporting all variations found in 69 genomes and their allele frequencies.
- The bit-called fraction’ from 69 genomes at each base position of the reference genome (experimental evidence of the CG ‘mappability’).

**Conclusions:** Disease variant discovery is largely based on the ability to identify features specific to patient genomes. Experimental design has been shown to play a key role in the ability to pinpoint disease variants. However, easy access to Tracks of aggregate, pre-processed data from multiple genomes allows a reduction of the number of candidate variants to a list compatible with Sanger-sequencing validation. Combining novelty filtering with ‘Biologist-readable’ annotations will make full-genome data more accessible to expert biologists.

### **A12: Using Hilbert curves to visualize structural variations with Meander**

Georgios A. Pavlopoulos\*, Alejandro Sifrim, Jan Aerts

Faculty of Electrical Engineering - ESAT/SCD, Katholieke Universiteit Leuven,  
Leuven, Belgium

Contact: georgios.pavlopoulos@esat.kuleuven.be

**Background:** In the past several years, the interest for browsing and analyzing structural variations in a genome increases exponentially. While sequencing techniques always improve and become cheaper over time, vast amounts of data get continuously produced. The analysis, the interpretation and the visualization of such an overload of information emerges and still remains a bottleneck.

**Methods and Results:** In this article we present Meander as java standalone application to visualize read-depth genome coverage. Meander uses a 2D plane of  $512 \times 512 = 262.144$  pixels. DNA is computationally split into 262.144 buckets, each one holding the average coverage among the nucleotides that belong to the specific bucket. Each pixel on the Hilbert curve represents the read-depth coverage of the bucket. The coverage value is initially mapped to a color gray scale. Thus, the higher the coverage is, the darker the pixel appears and vice versa. In addition to the Hilbert representation, Meander also uses a linear representation of 512 pixels in length to show read-depth coverage at a lower resolution using bar charts. Each pixel/bar represents the average coverage of the nucleotide of each of the 512 buckets. The higher the bar height is, the higher the coverage. Meander is able to visualize data both linearly and as Hilbert curve at 5 different zoom-levels. For higher performance, files holding information about the read-depth coverage at different resolutions are pre-calculated. To directly compare the coverage between two genomes (sample vs reference) we use the log ratio. We use two different color schemes (red-green and blue-yellow) to visualize the ratio. Given a range of structural variations as they are calculated by external software applications, Meander is able to overlay such information by highlighting the areas on the Hilbert curve using rich color schemes.

**Conclusions:** Meander is tool to visualize read-depth coverage information and investigate inter-chromosome structural variations (currently deletions and duplications). Users are able to directly compare genomes among each other and overlay structural variations predicted by other software. It is a highly interactive standalone java application that utilizes a 2D Hilbert space to visualize the genome coverage. Currently it supports browsing at 5 different zoom and resolution levels. While Meander is currently limited to visualize intra-chromosomal variations we aim to support whole genome browsing using both read-depth and split read information. Meander will get further extended to support both balanced and unbalanced structural variations.

### A13: Genomic Profiling of HMGN

Arjan van der Velde

National Center for Biotechnology Information (NCBI/NIH), United States

Contact: a.g.vandervelde@student.vu.nl

**Background:** Epigenetic changes play critical roles in many cellular processes including regulation of gene expression. Several epigenetic alterations are involved in gene regulation, with members of the high-mobility group (HMG) protein superfamily acting as key players in chromatin remodeling. HMG proteins are associated with activation of genes and involved in maintenance of open/accessible chromatin, competing with the converse effect of histone H1. The HMG superfamily consists of three families (HMGA, HMGB and HMGN) with distinct functions in eukaryotes. Altered expression of these proteins lead to developmental abnormalities. The HMGN family consists of five members (HMGN1-5), which are unique within the HMG superfamily in their ability to bind nucleosomal core particles directly through their nucleosomal binding domains without DNA sequence specificity. HMGN proteins are implicated in maintenance of open/active chromatin through their inhibitory effect on ATP dependent chromatin remodeling enzymes. However, it is still largely unknown how the different members of the HMGN family are involved in epigenetic regulation. Therefore we investigated the genomic landscape of HMGN proteins in wild-type and knockout mice B-cells using a large ChIP-Seq dataset, suggesting a possible (indirect) interaction between HMGN1, HMGN2 and HMGN3.

**Methods:** Using next-generation sequencing techniques a ChIP-Seq dataset has been generated for HMGN1, HMGN2, HMGN3, HMGN5 and several knock-outs. In order to perform genome-wide profiling on these proteins we developed a pipeline to validate replicates, map the sequenced DNA to the mouse genome and determine in which chromatin regions HMGN proteins bind. Our pipeline includes LAST, SICER and a combination of R (Bioconductor) and Python programs, which were used to analyze HMGN binding around transcription start sites and throughout coding regions of genes.

**Results:** HMGN1, 2, 3, 5 resulted preferentially localized to the gene body, particularly promoter regions. Binding sites identified by SICER in wild-type HMGN1, HMGN2 and HMGN3 overlap to a considerable degree, while differences in binding patterns were detected in HMGN2 samples with HMGN1 and HMGN3 present versus samples in which HMGN1 and both HMGN1 and HMGN3 were knocked out. We also confirmed systematic biases inherent to the ChIP-Seq technique that are currently actively studied.

**Conclusions:** Our results suggest a possible (indirect) interaction between HMGN2, HMGN1 and HMGN3. The degree of overlap in HMGN samples and the differences in HMGN2 binding between the samples give rise to a whole new set of hypotheses including cooperative behavior of the HMGNs and recruitment of HMGNs by common transcription factors.

**A14: Proteins in the orthology twilight zone. The Ortho-Profile iterative method and the experimental function confirmation.**

Radek Szklarczyk\*, Bas F.J. Wanschers, Thomas D. Cuypers, Leo G. Nijtmans,  
Martijn A. Huynen

CMBI, Radboud University Nijmegen Medical Centre

Contact: radek@cmbi.ru.nl

**Background:** Orthology is a central tenet of comparative genomics and its identification is instrumental to protein function prediction. Although major advances have been made to determine orthology relations among a set of homologous proteins, they depend on the comparison of individual sequences, disregarding homology relations that can only be detected by comparing sequence profiles.

**Methods:** We have developed a sensitive orthology prediction method (Ortho-Profile) that uses best reciprocal hits at the level of 1) sequence, 2) sequence profiles and 3) Hidden Markov Models to infer orthology. The method identifies orthologs of multiple assembly factors of respiratory chain complexes, previously unnoticed due to the short length and fast sequence evolution rate of genes encoding assembly factors.

**Results and Conclusions:** The Ortho-Profile method predicts 598 human orthologs of mitochondrial proteins from *Saccharomyces cerevisiae* and *Schizosaccharomyces pombe* with 94% accuracy. Based on this analysis, we obtain 174 novel candidate mitochondrial proteins in human. We selected 11 predicted cytochrome c oxidase assembly proteins for experimental validation, confirming the mitochondrial localization for five proteins and physical interactions with cytochrome c oxidase (COX)-associated proteins also for five proteins. Next to the functional assays supporting the proteins' role in COX assembly, we identified a homozygous disease mutation in C2orf64. This ortholog of *S. cerevisiae* gene PET191 is responsible for impaired COX assembly in human and causes fatal neonatal cardiomyopathy in two affected siblings.

**A15: Specificity of allostery networks within the SH3 domain family**

Ana Zafra Ruano, José Couceiro, Javier Ruiz Sanz, Joost Schymkowitz, Frederic Rousseau, Irene Luque, Tom Lenaerts\*

Département d'Informatique, Université Libre de Bruxelles, Belgium

Contact: tlenaert@ulb.ac.be

**Background:** Allostery, roughly defined as the mechanism responsible for changes in the shape of a protein surface site caused by changes at another site, provides a mechanism to exchange information, in a precise manner, within and between proteins. Next to large structural changes, it also covers long-range information exchange throughout protein structures induced by the reorientation of only the residue sidechains. Predicting precisely the residues involved in the process is of significant importance.

**Methods:** We proposed the idea that mutual information may be used to identify and quantify the network of allosteric residues. We derive it from NMR structural data by measuring the strength of the conformational coupling between each pair of residues through Monte Carlo sampling, taking into account the backbone variation provided by the structural ensemble. We apply this method here to show for the first time the biological significance of our predictions, the robustness of our method and the evolutionary significance of the identified network in SH3 domain family.

**Results:** Through our analysis of the bound and unbound Src SH3 domain, we identify a plausible allosteric network. We show that these results are robust by comparing this network to the one derived similarly for the close homolog Fyn SH3. Using differential scanning calorimetry and isothermal titration calorimetry, we determine that a subset of the network residues, not participating in the binding surface, affects the affinity for the peptide in vitro. In addition, we show that these mutants have a modest effect on the phosphorylation level of Src in cellulo. Finally, we show that the allosteric network predicted for Fyn and Src SH3 are similar to the recently identified evolutionary sectors for this domain. However, analysis of the family of SH3 domains reveals that only the closest homologues of Src and Fyn share the same allosteric pattern, exposing functional differences between members of the same domain family, which are supported by chimera experiments discussed in literature.

**Conclusions:** In light of these results we conclude that allosteric networks should be derived for the different members of a domain family since such detailed analysis will provide a better insight into the actual functional differences inherent to these domains. In the long run this information will not only provide a better understanding of the allosteric mechanisms controlling cell functioning, but also provide unambiguous information on how to engineer particular domain functionality within the context of a complete protein structure.

**A16: Identifying common structural DNA properties in transcription factor binding site sets of the LacI-GalR family**

Meysman Pieter\*, Marchal Kathleen, Engelen Kristof

M2S, K.U.Leuven, Belgium

pieter.meysman@biw.kuleuven.be

**Background:** It is well known that transcription factors can induce deformations in their DNA-binding sites upon complex formation. However, few attempts have been made to investigate the extent to which induced structural deformations in the DNA molecule are conserved between different members of the same transcription factor family. The LacI-GalR family is an extensively studied transcription factor family and provides an excellent case to evaluate the viability of identifying familial recognition patterns using binding site modeling.

**Methods:** The CRoSSeD methodology is a Conditional Random Field model tuned for the representation of DNA structural profiles and sequence motifs in the binding sites of transcription factors. These models allow for direct comparison of the recognition characteristics of different LacI-GalR family members and the creation of a family-wide binding site model.

**Results:** Using the CRoSSeD methodology, we were able to extract common features in the binding sites of different LacI-GalR family members. The most significant feature identified in this way was located at the center of the binding sites, which is also the most likely location for an induced DNA deformation following an amino acid interdigitation. This feature was related further to specific elements present in the protein structure and was used to identify and characterize deviant family members. A general family-wide binding site model was constructed and applied to screen for unknown member binding sites.

**Conclusions:** There seem to be recognition characteristics that are conserved throughout most of the LacI-GalR family which could be extracted by the CRoSSeD methodology. Likely this method can be expanded towards any transcription factor family and could be used to separate recognition characteristics due to a common bindingsmodi or which are specific for a given transcription factors.



**A17: Modelling the dynamics of chronic myeloid leukemia under therapy**

Tom Lenaerts\*, Fausto Castagnetti, Arne Traulsen, Jorge M. Pacheco, Gianantonio Rosti, David Dingli

Département d'Informatique, Université Libre de Bruxelles, Belgium

Contact: tlenaert@ulb.ac.be

**Background:** Understanding the molecular mechanisms behind the appearance and clonal expansion of BCR-ABL expressing cells in chronic myeloid leukemia (CML) has transformed therapy and prognosis for this disease. Rationally designed tyrosine kinase inhibitors (TKI) such as imatinib and nilotinib lead to deep and durable responses and have reduced the risk of progression to blast crisis. Nevertheless, multiple questions remain like i) whether it is possible to stop TKI treatment and if yes when to stop, ii) whether nilotinib or imatinib should be given as first line treatment and iii) whether patient outcome may be predicted from initial TKI treatment results.

**Methods:** We aim to provide answers to these questions through the use of computational model of the hematopoietic system that may be used to understand CML dynamics under TKI therapy. This model contains 3 essential parameters, i.e. the differentiation rate of the cancer cells ( $\epsilon_{CML}$ ), the differentiation rate of the treated cancer cells ( $\epsilon_{TKI}$ ) and fraction of the cells affected by treatment ( $z_{TKI}$ ), for which values are obtained by fitting the model to the BCR-ABL transcript data of patient cohorts using a non-linear least squares approach with constraints based on established observations.

**Results:** So far we have shown that, when taking into account that the hematopoietic model is partially stochastic, the majority of the virtual patients in our simulations (~84%) no longer possess the leukemic stem cell that initiated the disease, making the progenitors and not the stem cells the drivers of the disease. An important implication of this is that the different TKI, which affect progenitor cells, may actually be capable of curing the disease. Additionally, we have shown that treatment differences between TKI are explainable by the differences in the values of model parameters: Nilotinib produces a higher fitness disadvantage for the CML progenitor cells than imatinib producing in this way the faster and deeper response observable in transcript data. Finally, analysis of individual patient transcript levels indicate that our model may provide help in answering the third question, i.e. whether we can predict patient outcome from initial response. These results show that current risk scores are not sufficient to decide TKI therapy, requiring close monitoring of patient.

**Conclusions:** Together these results show that our computational model provides a clear added value in our understanding of hematopoietic diseases like CML. Additionally, it may provide in the long-term a decision support tool to guide individual patient treatment.

**A19: PROCAR-SEQ An analysis and visualization framework for next-generation sequencing based quantification of prokaryotic communities**

Joachim De Schrijver\*, Pieter-Jan Volders, Frederiek-Maarten Kerckhof, Dagmar Obbels, Elie Verleyen, Wim Vyverman, Tim De Meyer, and Wim Van Criekinge

Laboratory of Computational Genomics and Bioinformatics (BioBix), Ghent University, Ghent, Belgium

Contact: joachim.deschrijver@ugent.be

**Background:** 16S ribosomal DNA (rDNA) based PCR followed by sequencing of these PCR fragments is the preferred method of quantifying prokaryotic microbial communities. Limitations of the sequencing technology (mainly the throughput) were until recently the limiting factor to quantify and compare large amounts of different samples or communities. However, recent developments in high-throughput sequencing techniques allow easier quantification of microbial communities and have led to a spectacular increase of insights into the composition and functionalities of microbial communities.

The development of these recent sequencing techniques went hand in hand with the development of new analysis tools. In the last couple of years, several tools such as MOTHUR, PyroNoise/AmpliconNoise, VITCOMIC and Qiime have been developed to analyse pyrosequencing data sets and to assess microbial diversity.

**Methods:** PROCAR-SEQ is a modular framework consisting of several different modules, which uses a MySQL database to retrieve raw data, to store intermediary data, and to store final analysis results. Final results are visualized by the visualization module, which fetches all required data from the database as well. Sequences are processed using AmpliconNoise and mapped onto the reference using BWA-SW. 16S rDNA reference sequences were downloaded from the Ribosomal Database Project (RDP) website. A complete taxonomic overview of all prokaryotes present in RDP was downloaded from the RDP website. Each taxonomic entry is quantified (using a custom perl pipeline) on each taxonomic level (ylum, order, genus...). An interactive interface was made using HTML/PHP and the Google Maps API. The web-based toolkit includes a prokaryote taxonomical browser, a geographical visualisation tool and various graphing tools.

**Results:** For a user-selected taxonomic clade (domain, phylum, class, order, family or genus according to RDP), the relative frequency in each sample is visually depicted by a circle on the geographic location of the sample. For each sample, the rRNA content is visualized with a chart depicting the relative abundance of the selected taxonomic group from high to low. Thus users can compare different samples with each other (on the required taxonomic level) or analyze the composition of a single sample in detail.

**Conclusions:** We developed PROCAR-SEQ, a tool to quantify and visualize microbial communities using GS-FLX next-generation sequencing (NGS) data. An example of the possibilities is available at [http://athos.ugent.be/metagenom-x/?p=map\\_test](http://athos.ugent.be/metagenom-x/?p=map_test).

**A20: The Fc receptor complex from human neutrophils.**

Florentinus AK, Jankowski A, Petrenko V, Bowden P, Marshall JG.\*

Department of Chemistry and Biology, Ryerson University, Canada

Contact: 4marshal@ryerson.ca

**Background:** The Fc receptor complex and its associated phagocytic cytoskeleton machinery were captured from the surface of live cells by IgG coated microbeads and identified by mass spectrometry.

**Methods:** The random and independently sampled intensity values of peptides were similar in the control and IgG samples analyzed by LC-ESI-MS/MS. After log transformation, the parent and fragment intensity values showed a normal distribution where more than 99.9% of the data was well above the background noise.

**Results:** Some proteins showed significant differences in intensity between the IgG and control samples by ANOVA followed by the Tukey-Kramer honestly significant difference test. However many proteins were specific to the IgG beads or the control beads. The set of detected cytoskeleton proteins, binding proteins and enzymes detected on the IgG beads were used to predict the network of actin-associated regulatory factors. Signaling factors/proteins such as PIK3, PLC, GTPases (such CDC42, Rho GAPs/GEFs), annexins and inositol triphosphate receptors were all identified as being specific to the activated receptor complex by mass spectrometry. In addition, the tyrosine kinase Fak was detected with the IgG coated beads.

**Conclusions:** Hence, for the first time, an activated receptor complex and its associated cytoskeleton and regulatory proteins were captured from the surface of live human primary cells. The mass spectrometry data provided specific isoform information and extended the protein-protein interaction model for the Fc receptor in human leukocytes.

**A21: Proteomic identification of differentially expressed proteins in curcumin-treated prostate cancer cells**

Anthoula Gaigneaux\*, Marie-Hélène Teiten, Sébastien Chateauvieux, Anja Billing, Jenny Renaut, Claude P. Müller, Mario Dicato, Marc Diederich

Laboratoire de Biologie Moléculaire et Cellulaire du Cancer

Contact: anthoula.gaigneaux@lbmcc.lu

**Background:** Prostate cancer is the most common cancer in men in the western world. Lifestyle, diet and environmental as well as genetic factors promote malignant transformation of healthy prostatic epithelium. Due to the high prevalence and the slow progressive development of prostate cancer, primary prevention appears as an attractive strategy to eradicate prostate cancer. During the last decade, curcumin (diferuloylmethane), a natural compound extracted from Turmeric (*Curcuma Longa*), was described to be a potent chemopreventive agent as it exhibits anti-inflammatory, anti-carcinogenic, anti-proliferative, anti-angiogenic and anti-oxidant properties in various cancer cells. The present study was designed to identify proteins involved in anti-cancer activity of curcumin in androgen dependent and independent human prostate cancer cells.

**Methods:** A 2D-based proteomic analysis (2D-DIGE) was performed to identify differentially expressed proteins in nuclear and cytosolic fractions extracted from 22Rv1 and PC-3 cell lines, treated with 20  $\mu$ M curcumin during 24 h. Gels comparing treated and untreated fractions were performed in triplicate. Protein identification was obtained by MALDI-TOF-MS analysis of selected spots. Several bioinformatics analyses were performed in order to find biological categories associated with identified proteins, as well as proteins that could regulate them.

**Results:** 2D-DIGE revealed 425 differentially expressed spots after curcumin treatment in either 22Rv1 and/or PC-3 cells, among which 93 proteins were identified. Enrichment analyses showed that proteins modulated by curcumin were implicated in GO categories and KEGG pathways related to protein folding (Hsps family members, PP2R1A), RNA splicing (RBM17, DDX39), and cell death (HMGB1, NPM1). Network analysis was performed to find potential regulators of affected proteins and identified several candidates, including MYC and HSF-1, and miRNAs. Contextualisation of these results with published data highlighted relationships with the androgen receptor, which was further shown to be down-regulated in androgen-dependent 22Rv1 cells upon curcumin treatment.

**Conclusions:** Taken together, these data underlined the chemopreventive potential of curcumin by showing that curcumin modulates the expression of proteins that potentially contribute to prostate carcinogenesis.

**A22: Divergence in the length of unstructured protein termini drives the evolution of protein half-life**

Robin van der Lee\*, Kai Kruse, Benjamin Lang, Jörg Gsponer, Natalia Sánchez de Groot, Monika Fuxreiter, M. Madan Babu

MRC Laboratory of Molecular Biology, Cambridge, United Kingdom - Centre for Molecular and Biomolecular Informatics, Radboud University Nijmegen Medical Centre, The Netherlands

Contact: r.vanderlee@cmbi.ru.nl

**Background:** Protein degradation is the end point of gene expression. The presence of unstructured regions (i.e. intrinsic disorder) in protein sequences is important for proteasome-mediated degradation of specific proteins. However, genome-scale design principles of the relationship between intrinsic disorder and protein degradation in vivo remain unknown.

**Methods:** The availability of sequence information, robust disorder prediction algorithms and experimentally determined in vivo protein half-life for the yeast proteome provides us with a unique opportunity to investigate this question. We integrated information on protein half-life with the position and length of intrinsic disorder and analyzed it using appropriate statistical tests.

**Results:** We show that proteins with longer terminal or proteasome-cleavable internal disorder tend to have shorter half-life in yeast. Strikingly, an investigation of paralogous proteins revealed that upon gene duplication, divergence in the length of terminal disorder results in altered half-life of paralogs.

**Conclusions:** The results suggest that variation in the length of intrinsic disorder may serve as a remarkably simple means to evolve protein half-life. Since altered protein half-life can influence cellular homeostasis, we suggest that mechanisms that generate diversity in the length and position of intrinsic disorder of key pleiotropic genes may serve as an underappreciated source of genetic variation that may have important phenotypic consequences.

**A23: PuMaQC: R-based pipeline for the search, import and QC/QA of public microarray data**

Joana P. Corte-Real, Petr V. Nazarov\*, Arnaud Muller, Tony Kaoma and Laurent Vallar

Microarray Center, CRP-Sante, Luxembourg

Contact: petr.nazarov@crp-sante.lu

**Background:** Data-driven studies such as inference of gene regulatory networks and translational cancer research normally require large amounts of transcriptomic data. One simple and cost free solution comes from importing microarray data from public repository databases such as NCBI Gene Expression Omnibus (GEO), integrating hundreds of thousand experiments. Despite the existence of the MIAME guidelines for standard microarray information, there is still a lack of information related to the quality of submitted data. Given that low quality samples can add noise and impair the statistical and biological significance of microarray analysis, quality control and quality assessment (QC/QA) becomes an important step when using public microarray data. Taking this into account we have developed R-based PuMaQC (Public Microarray Quality Control) pipeline.

**Methods:** PuMaQC is a robust, easy to use, all-in-one pipeline for public microarray data handling based on 3 sequential steps: i) search for raw Affymetrix data in GEO, ii) import and preprocessing of CEL files; and iii) QC/QA with identification and removal of low quality arrays. The pipeline incorporates functions from GEOmetadb, GEOquery, arrayQualityMetrics R/Bioconductor packages and uses Affymetrix Power Tools (APT) for raw data extraction and normalization. We have included the possibility to filter out unwanted samples at step (i), and a Gpl-platform dictionary that allows broadening sample search to several related GEO platforms (Gpl).

**Results:** To test PuMaQC we have applied it to 3 possible cases when searching for healthy human lung samples generated with Affymetrix HG-U133plus2 chips:

1. All lung-related samples existing for GPL570.
  2. Similar to case 1, but filtering out cancer- and embryo-related samples.
  3. Similar to case 2 but broadening search to all Gpl related to HG-U133plus2 chips.
- The search for human lung related samples returned a total of 1370 found GEO samples (Gsm) (Case 1). By filtering out cancer related samples we were able to exclude 1313 unfitting samples, leaving a total of 57 arrays (Case 2), hence avoiding and exhaustive manual curation of query results. The incorporation of a Gpl-platform translation dictionary (Case 3) doubled the number of found arrays (105 in case 3).

**Conclusions:** PuMaQC pipeline allows for performing effective search, import and QC of public Affymetrix microarray data, with identification and removal of outliers. PuMaQC is a simple-to-use, relatively fast, but powerful tool, which makes it attractable to both bioinformaticians and biologists. See <http://sablabs.net/PuMaQC> for the details.

**A24: Identification of crosstalk strength in signalling networks by optimization of Probabilistic Boolean Network models**

Thomas Sauter\*, Panuwat Trairatphisan

Life Sciences Research Unit, University of Luxembourg

Contact: thomas.sauter@uni.lu

**Background:** The dynamic behaviour of a signalling pathway is usually also dependent on crosstalk effects from other signalling pathways therefore generating strongly connected networks. Deciphering this complexity applying detailed mathematical modelling with ordinary differential equations is very often hampered by the low availability and/or identifiability of the kinetic parameters. On the other hand the widely applied Boolean Networks allow a description of the crosstalk structure, but fail in the quantitative determination of the crosstalk strength, because of their inherent qualitative nature.

**Methods:** We therefore used Probabilistic Boolean Network (PBN) models to describe signalling crosstalk in a simplified but quantitative manner, and to allow for a robust determination of crosstalk strength based on the integration of a candidate interaction network with experimental data. PBNs thereby combine the appealing characteristics of rule based Boolean modelling with a probabilistic description in the context of. They have so far mainly been applied to gene regulatory networks but are also ideally suited for the description of signalling networks with uncertain crosstalk.

**Results and Conclusions:** An automated Matlab based computational toolbox (optPBN) was developed to estimate the respective probabilities (representing the strengths of the crosstalk reactions) based on experimental data. In contrast to classical discrete Boolean approaches, continuous values of the experimental data are used directly. optPBN allows to automatically generate PBN models based on a candidate network and to easily import the experimental data. A combined optimization problem is generated and efficiently solved with a particle swarm algorithm. The quality of estimation is assessed by statistical analysis following a Bootstrapping approach.

optPBN was successfully tested with a variety of signalling models, including pro- and anti-apoptotic signalling in hepatocytes and altered growth factor signalling in cancers. It allows indicating the crosstalk strength in the networks based on quantitative proteomics data therefore pointing to disease specific sensitive network points.

**A25: Accuracy of information flow predictions within the PTP1E PDZ2 domain**

Elisa Cilia\*, Geerten W. Vuister, Tom Lenaerts

Département d'Informatique, Université Libre de Bruxelles, Belgium

Contact: ecilia@ulb.ac.be

**Background:** PDZ domains have been extensively analysed with the aim of identifying long-range allosteric effects caused by binding events. PDZ2 methyl-sidechain relaxation experiments in combination with crystallographic analysis have shown that sidechain dynamics play the major role in the propagation of the information throughout this domain structure as opposed to backbone structural changes.

**Methods:** We applied our information theoretical approach to predict which residues are allosterically relevant in the PDZ2 domains of human (hPDZ2) and mouse (mPDZ2) protein tyrosine phosphatase 1E. The algorithm computes the changes in the mutual information among the residue sidechain distributions, sampled from the NMR bound and unbound structures. We assessed the quality of our predictions with respect to the results of NMR relaxation experiments reported in the literature for the hPDZ2 binding to the RA-GEF-2, and we compared the predictive quality with other computational methods. We built a network of direct allosteric couplings among residues in the domain, to identify allosteric pathways, and we investigated the potential of network flow algorithms in explaining long-range effects between distally located residues.

**Results:** Starting from the NMR relaxation results we show that current computational approaches have a predictive quality comparable to that of a random model. We show that our method produces more accurate predictions. We also show that there is a significant overlap between the allosteric networks identified for the hPDZ2 and mPDZ2, even though the latter experiences structural changes in the second  $\alpha$ -helix upon binding to APC-derived peptide. Our network of allosteric couplings for the human variant reveals two main information flow pathways that partially confirm previous findings.

**Conclusions:** Our information theoretical approach not only gives us good quality predictions of the hPDZ2 residues involved in allosteric mechanisms, it also allows us to reason about the flow of information going through the domain structure, and to build residue networks revealing where the major dynamical changes are taking place. The identification of these flows will allow to explicitly study which interventions (mutations) may increase or decrease flow, improving in this way our general scientific understanding of allosteric mechanisms inside protein structures.



**A26: DALAS : an R-Java desktop application for Affymetrix exon array data analysis**

Tony Kaoma\*, Christelle Ghoneim, Arnaud Muller, Petr Nazarov, Laurent Vallar

Microarray Centre, CRP-Sante, Luxembourg

Contact: tony.kaoma@crp-sante.lu

**Background:** Affymetrix Exon Array (AEA) is a sophisticated microarray platform dedicated to alternative splicing detection. Although powerful, AEA generate data that are associated with a significant level of false positives. Analysis of such arrays thus requires complex statistical and bioinformatics methods to get a list of robust candidates for biological validation. Although several user-friendly tools have been developed to analyze AEA data, R (particularly R/Bioconductor and recently R/Aroma.Affymetrix) remains the most used. Because of requirements in programming and statistics, R is not easily accessible to biologists. To overcome this limitation, we developed DALAS (Detection of Alternative Splicing) - a Java-GUI desktop application which used R in background as a calculator engine.

**Methods:** DALAS was designed to completely fit the general pipeline established for AEA analysis. This pipeline consists of preprocessing and quality control, pre-filtering, detection of differential events, post-filtering, downstream analysis and visualization. We began by listing all R methods, which can be used at each step. Then, we selected the most relevant method according to the literature and organized them into a workflow from which we designed our software using R as a design pattern. To implement our tool, we used Java in order to make DALAS flexible and able to run on any standard operating system. In addition, Java provides a lot of libraries allowing development of advanced user-friendly GUI and chart.

**Results:** DALAS is a Java desktop application for the analysis of AEA data from raw CEL files to downstream analysis. It offers a collection of R-methods recommended for the analysis of AEA. More than 5 statistical methods and various filters are available. The exons identified through this process are then annotated using reference public databases (RefSeq, Ensembl, KEGG, PubMed, etc) and analyzed using gene set enrichment-based approach. Relevant exon candidates can be easily selected and visualized (with DomainGraph, IGV) for further study. Through a careful benchmarking study, DALAS was shown to outperform other similar existing tools by considerably reducing false positive rates in exon datasets.

**Conclusions:** DALAS is a suitable tool to analyze AEA data. With a user-friendly interface and an intuitive workflow, DALAS can be used regardless of background in programming and statistics. Its design makes it extendable to handle data from other “omics” platforms.

**A27: Cis-regulation of toxin clusters in *Fusarium*.**

Valeria Montis<sup>1</sup>, Francesca Cardinale, Ivan Visentin, Marco Beyer<sup>2</sup>, Hoffmann Lucien<sup>2</sup>, Harold Corby Kistler<sup>3</sup>, Matias Pasquali\*<sup>2</sup>

<sup>1</sup>Dept of Plant Physiology, University of Torino, <sup>2</sup>CRP- Gabriel Lippmann, EVA Department, <sup>3</sup>USDA ARS Cereal Disease Laboratory, St Paul USA

Contact: pasquali@lippmann.lu

**Background:** *Fusarium* species are able to produce an array of known and unknown metabolites. *Fusarium* plant pathogens produce toxins that accumulate in the edible plant parts contaminating food and feed. Current European regulations determine limits for some fusariotoxins such as deoxynivalenol and fumonisins. These two toxins are produced by *F. graminearum* and *F. verticillioides* by two main gene clusters (the tri and fum cluster, respectively) conferring ability to synthesize, intracellularly detoxify and export the toxin. Understanding the mechanisms driving toxin synthesis and regulation is therefore important for further developing containment measures.

**Methods:** Here the two key regulatory Transcription Factor Binding Sites (TFBSs) specific for the tri cluster and the fum cluster were identified by overrepresentation analyses. Moreover for *F. graminearum* full genome expression studies were used to identify genes controlled by the main transcriptional regulator of the cluster. In *F. verticillioides* in vivo and in planta experiments were carried out to confirm the biological function of the newly identified binding site.

**Results:** In *F. graminearum*, the TFBS identified confirmed a previous EMSA study. The TFBS also was found upstream of the genes of the primary metabolism necessary for toxin synthesis. Comparative genomics showed that this is true only for the species carrying the cluster therefore suggesting that co-evolution of the genome and the cluster has occurred.

In *F. verticillioides* a new cis-regulatory sequence was identified. When experimentally modified, the newly discovered TFBS resulted in a decreased transcriptional activity of the fumonisin cluster PKS gene (FUM1), suggesting a direct role of the predicted binding site in fumonisin regulation. The TFBS is partially conserved in *F. oxysporum* and also in the phylogenetically distant *Aspergillus niger* (both able to produce fumonisins), suggesting the sequence is bound by a transcriptional factor specific for the cluster. Structural and bioinformatics data suggest it may be the zinc-finger protein FUM21 that has a slight modification of the DNA-binding site between *F. verticillioides*, *F. oxysporum* and *A. niger* which could account for the slight different preferences in the TFBS sequence.

**Conclusions:** After acquisition of a toxin biosynthetic cluster two events were observed. In *F. graminearum* the genome appears to adapt to the cluster specific transcription factor while in the fumonisin producing species (*A. niger*-*F. verticillioides*-*F. oxysporum*). The TFBS has coevolved with the zinc finger protein binding it.

**A28: Analysis of the X!TANDEM correlation of the HuPO blood consortium results by SQL and SAS**

Bowden P, Beavis R, Marshall JG\*

Department of Chemistry and Biology, Ryerson University, Canada

Contact: 4marshal@ryerson.ca

**Background:** The Human Proteome Organization (HuPO) Plasma Proteome Initiative coordinated the efforts of some 35 laboratories international to analyze the peptides and proteins of human blood. Here we show the results of large proteomic experiments can be completely analyzed using only the standard Structure Query Language (SQL) and Statistical Analysis System (SAS) software packages.

**Methods:** A goodness of fit test may be used to assign tandem mass spectra of peptides to amino acid sequences and to directly calculate the expected probability of mis-identification. The product of the peptide expectation values directly yields the probability that the parent protein has been mis-identified. A relational database could capture the mass spectral data, the best fit results, and permit subsequent calculations by a general statistical analysis system. The many files of the HuPO blood protein data correlated by X!TANDEM against the proteins of ENSEMBL were collected into a relational database using SQL server and statically analyzed using SAS.

**Results:** A redundant set of 247,077 proteins and peptides were correlated by X!TANDEM, and that was collapsed to a set of 34,956 peptides from 13,379 distinct proteins. About 6875 distinct proteins were only represented by a single distinct peptide, 2866 proteins showed 2 distinct peptides, and 3454 proteins showed at least three distinct peptides by X!TANDEM. More than 99% of the peptides were associated with proteins that had cumulative expectation values, i.e. probability of false positive identification, of one in one hundred or less.

**Conclusions:** The distribution of peptides per protein from X!TANDEM was significantly different than those expected from random assignment of peptides. Hence the HuPO Plasma Proteome Initiative was successful in identifying the proteins of human blood with good confidence.

**A29: Quantitative statistical analysis of blood proteins from liquid chromatography, electrospray ionization and tandem mass spectrometry**

Bowden P, Zhu P, McDonnell M, Thiele H, Marshall JG\*

Department of Chemistry and Biology, Ryerson University, Canada

Contact: 4marshal@ryerson.ca

**Background:** It will be important to determine if the parent and fragment ion intensity results of liquid chromatography, electrospray ionization and tandem mass spectrometry (LC-ESI-MS/MS) experiments have been randomly and independently sampled from a normal population for the purpose of statistical analysis by general linear models and ANOVA.

**Methods:** The parent and fragment ion  $m/z$  and intensity data in the mascot generic files from liquid chromatography, electrospray ionization and tandem mass spectrometry of human blood were parsed into a Structured Query Language (SQL) database and were matched with protein and peptide sequences provided by the X!TANDEM algorithm. The many parent and fragment ion intensity values from 302 peptides sequences of 119 proteins were transformed, tested for normality and analyzed using the generic Statistical Analysis System (SAS).

**Results:** Transformation of both parent and fragment intensity values by logarithmic functions yielded intensity distributions that closely approximate the log normal distribution. ANOVA models of the transformed parent and fragment intensity values showed significant effects of treatments, proteins, and peptides, as well as parent versus fragment ion types, with a low probability of false positive results. Transformed parent and fragment intensity values were compared over all sample treatments, proteins or peptides by the Tukey-Kramer Honestly Significant Difference (HSD) test.

**Conclusions:** The approach analyzing log transformed ion intensity values mapped to peptides, proteins and treatments by ANOVA provided a complete and quantitative statistical analysis of LC-ESI-MS/MS data from human blood without the use of accurate mass tags, isotopic labels or chromatographic retention times.

**A30: Analyzing gene and protein expression variance in cellular pathways using high-throughput experimental data**

Enrico Glaab\*, Reinhard Schneider

Luxembourg Centre for Systems Biomedicine

Contact: enrico.glaab@uni.lu

**Background:** Finding significant differences between the expression levels of genes or proteins across diverse biological conditions is one of the primary goals in the analysis of functional genomics data. However, existing methods for identifying differentially expressed genes or sets of genes by comparing measures of the average expression across predefined sample groups do not detect differential variance in the expression levels across genes in cellular pathways. Corresponding pathway deregulations occur frequently in microarray gene or protein expression data, but so far, no software tool has been available to systematically exploit these deregulation patterns for biological data interpretation.

**Methods:** We present a new web-application, PathVar, for microarray data analysis to identify and prioritize pathways with changes in the pathway expression variance across samples (unsupervised mode) or predefined sample groups (supervised mode). In the supervised analysis mode, the software ranks pathway-representing gene/protein sets in terms of the differences of the variance in the within-pathway expression levels across labeled sample groups using both parametric and non-parametric statistical tests. Alternatively, in the unsupervised mode, three feature rankings are obtained from the extracted matrix of pathway expression variances (rows = pathways, columns = samples) by computing the absolute variances across the samples, the magnitude of the loadings in a sparse principal component analysis, and an entropy score proposed recently in the literature. Apart from identifying new pathway deregulation patterns, the tool exploits the extracted patterns by combining different machine learning methods to find clusters of similar samples and/or build sample classification models.

**Results:** When evaluating PathVar on two labeled microarray cancer datasets and cellular pathways from the KEGG database and additionally comparing the median gene expression levels in the pathways across the sample classes, PathVar identifies new significantly deregulated pathways, which are not identified by a conventional comparison of the median expression levels. These include known cancer-associated KEGG pathways, like the angiogenesis-related “VEGF signaling pathway” and the inflammation-related “Natural killer cell mediated cytotoxicity” process. Moreover, when using the extracted pathway expression variance matrix to train machine learning models for sample classification, the obtained models reach cross-validated accuracies in the range between 70% and 100%.

**Conclusions:** PathVar is an easy-to-use web-application that identifies statistically significant pathway deregulations, different from those detected by classical methods for comparing averaged expression levels. This enables the software to generate pathway-based clustering and classification models that enable a new interpretation of microarray data. The web-application is freely available at <http://pathvar.embl.de>

**A31: COLOMBOS: Access Port for Cross-Platform Bacterial Expression Compendia**

Kristof Engelen, Qiang Fu, Pieter Meysman\*, Aminael Sanchez-Rodriguez, Riet De Smet, Karen Lemmens, Ana Carolina Fierro, Kathleen Marchal

Department Of Microbial And Molecular Systems (M2S), KULeuven, Belgium

Contact: pieter.meysman@biw.kuleuven.be

**Background:** Microarrays are the main technology for large-scale transcriptional gene expression profiling, but the large bodies of data available in public databases are not useful as is due to the large heterogeneity. There are several initiatives that attempt to bundle these data into expression compendia, but such resources for bacterial organisms are scarce and limited to integration of experiments from the same platform or to indirect integration of per experiment analysis results.

**Methods and Results:** We have constructed comprehensive organism-specific cross-platform expression compendia for three bacterial model organisms (*Escherichia coli*, *Bacillus subtilis*, and *Salmonella enterica* serovar Typhimurium) together with an access portal, dubbed COLOMBOS, which provides a suite of tools for exploring, analyzing, and visualizing the data within these compendia. It is freely available at <http://bioi.biw.kuleuven.be/colombos>. The compendia also incorporate extensive annotations for both genes and experimental conditions; these heterogeneous data are functionally integrated in the COLOMBOS analysis tools to interactively browse and query the compendia not only for specific genes or experiments, but also metabolic pathways, transcriptional regulation mechanisms, experimental conditions, biological processes, etc. Additionally we have invested in the development of a compendia creation and management system: automated retrieval and parsing of experiments from GEO and ArrayExpress, guided sample annotation, and homogenization consisting of various normalization pipelines. (All working on the same backend DB schema that COLOMBOS runs on.)

**Conclusions:** We have created cross-platform expression compendia for several bacterial organisms and developed a complementary access port COLOMBOS, which also serves as a convenient expression analysis tool to extract useful biological information. This work is relevant to a large community of microbiologists by facilitating the use of publicly available microarray experiments to support their research.

### A32: Receptor-Ligand Prediction through Machine Learning

Ernesto Iacucci\*, D. Popovic, L.-C. Tranchevent, B. De Moor, and Y. Moreau

KULeuven, ESAT/SCD

Contact: ernesto.iacucci@gmail.com

**Background:** Intercellular communication is mediated by the interaction of circulating ligands and cellular receptors. These interactions often initiate important biological processes such as tissue development, homeostasis, and stress response. Identification of receptor-ligand pairs is thus an important task, facilitated by computational prediction. We consider that all protein pairs may be assigned to one of two classes, the interacting class and the non-interacting class. All proteins have measurements from various data sources such as expression profiles and sequence information. The association between the measurements of two proteins can be represented by various similarity measures used to construct a similarity feature.

**Methods:** As several data sources, similarity measures, and classifiers exist, it's not clear which combination will function best at addressing the interaction prediction task described above. In this study, we benchmark the performance of several popular classifiers across several different data sources and similarity measures to find the best combination for this prediction task.

**Results:** Through benchmarking of Database of Ligand-Receptor Partners (DLRP) protein receptor-ligand dataset across several classifiers we find the best performing classifier to be the random forest with a sensitivity of 0.87 and a specificity of 0.79. The most informative feature was found to be Kegg as its' removal from the trials caused a drop in performance of about 15%. The most useful similarity measure, when the random forest is used, was found to be the absolute cosine function as it was the best performing measure for four of the six features.

**Conclusions:** The results from this work suggest several key findings. First, there exists a major advantage in balancing the training data, particularly with the lowest performing classifiers, suggesting that they are sensitive to imbalanced data. Second, there is a high amount of mutual dependence between some features, suggesting that they are not all necessary for the interaction prediction task. Finally, as the best performing feature was the Kegg membership, though it seems that the performance of the other features combined works very well also.

### A33: Metabolic Modeling of Human Adipogenesis

Mafalda Galhardo\*, Merja Heinäniemi, Thomas Sauter

LSRU, University of Luxembourg, Luxembourg

Contact: mafalda.galhardo@uni.lu

**Background:** Alterations in metabolism sustain both normal cellular adaptation to perturbations and disease-associated states. Opposed to data on cellular metabolism, gene expression data is easily and commonly obtained (e.g. microarrays). Such data availability motivates the assessment of how well can gene expression data represent metabolic fluxes, a task performed by combining them with a human metabolic model (Recon1) through gene-protein-reaction associations and computing for metabolic flux distributions consistent with both gene expression and network structure. Adipogenesis is the process through which precursor cells differentiate into mature adipocyte cells with the capability of producing and storing lipids (fats). During adipogenesis, many alterations in gene expression have been reported while the key metabolic changes remain less well characterized.

**Methods:** Microarray data from human pre-adipocytes and adipocytes (SGBS cell line) were first discretized into 3-format values accounting for lowly (-1), moderately (0) and highly (1) expressed metabolic genes. Different methods for discretization were employed in order to test how much the resultant flux predictions would differ. Selection of the values for each of the 3 discrete categories was based on thresholds obtained using the median, mean or quartiles of all expression values. The human general metabolic model Recon1 was obtained from the BiGG database (<http://bigg.ucsd.edu/>). Shlomi's method for flux prediction was implemented in Matlab® and the results were analyzed using own implementation.

**Results:** The different methods employed to discretize the expression values lead to different flux predictions. A single threshold for assigning discrete categories (median) clearly differed from methods in which two thresholds were used. Comparison of the flux predictions between pre-adipocytes and adipocytes revealed a trend of up-regulation, rather than down-regulation, of pathways in adipocytes. Pathways involved in lipid metabolism were the ones predicted to have the most differences in the two cell types, with cholesterol synthesis pathway shifting from inactive to totally active in adipocytes. Fatty acid metabolism and triacylglycerol synthesis also appeared up-regulated in adipocytes.

**Conclusions:** Our work highlights the dependency of metabolic flux prediction on the method used for converting continuous expression values into discrete tri-valued format. Nevertheless, the method was able to predict differences on metabolic flux in lipid metabolism pathways between pre-adipocytes and adipocytes that are biologically plausible.



### A34: Merging partially labelled trees

Anthony Labarre\*, Sicco Verwer

Department of Computer Science, Katholieke Universiteit Leuven, Belgium

Contact: Anthony.Labarre@cs.kuleuven.be

**Background:** Intraspecific studies often make use of haplotype networks instead of gene genealogies to represent the evolution of a set of genes. Cassens et al. designed a network reconstruction method for that purpose, based on the global maximum parsimony principle, which consists in (1) generating genealogies explaining the history of a given set of genes, and then (2) finding a graph that contains all those genealogies as subgraphs and is as small as possible. They proposed a heuristic for the latter step, and found that their approach performed quite well with respect to other widely-used approaches on simulated data, sometimes outperforming them. However, their algorithm makes a number of arbitrary choices, produces solutions whose quality depends on the order in which the merging process is performed, and is a heuristic with an implicit objective function.

**Methods:** We introduce a formal model for the second step of the network reconstruction method proposed above, and study it both from a theoretical and from a practical point of view. We explore the possibilities offered by SAT solvers to provide an exact solution to our problem, and examine the performances of this approach on various generated datasets.

**Results:** We prove that our problem is computationally difficult in a well-defined sense, which is usually a commonly accepted justification for designing approximate and alternative solutions to efficiently solving the problem to optimality. Nevertheless, we focus our efforts on the search for an efficient exact solution, and rely to that end on a system developed by Wittocx et al. We study the performance of this approach on simulated data, and compare the results obtained in that way with those yielded by a greedy-type alternative. We observe that the greedy approach is a surprisingly good approximation, and that the exact approach, while eventually limited by the inherent hardness of the problem, provides an efficient practical solution for reasonably large instances.

**Conclusions:** The solution we propose is very appealing, because of (1) its ease of use, (2) its efficiency, and (3) the possibility of interrupting the search for an optimal solution while retaining the ability to save the best result obtained so far. Future perspectives include evolving the prototype into a user-friendly application integrating with other popular phylogenetic software, as well as helping determine useful additional criteria for discriminating between different optimal solutions.

**A35: Evaluation of novel SNP prioritization and sub-set selection approaches for GWAS**

Yeşim Aydın Son

Bioinformatics Graduate Program, Middle East Technical University, 06800, Ankara, Turkey and Health Informatics Department, Middle East Technical University, 06800, Ankara, Turkey

Contact: yesim@metu.edu.tr

**Background:** Translational and clinical research to develop new personalized medicine approaches requires identification of predictive and diagnostic biomarkers. Genome-Wide Association Studies (GWAS) of Single Nucleotide Polymorphisms (SNPs), are among the promising approaches for the identification of disease causing variants for biomarker discovery. The high-dimension of the SNP genotyping data presents a challenge for the understanding of the genotype and its possible implications for the etiology of the diseases and also for the identification of the representative SNPs to design the follow up studies for the validation of the associations. Two major bottlenecks of standard GWAS approaches are the prioritization of statistically significant results and selecting representative SNP subsets for the conditions under study. Data mining methodology that is based on finding hidden and key patterns over huge databases have the highest potential for extracting the knowledge from genomic datasets and to select the sub-set of SNPs that are representative and informative for clinical applications.

**Methods:** We are investigating data mining and optimization tools that are available in other scientific fields, such as engineering and finance, for the dimension reduction of high-throughput data that can be utilized for the development of prioritization and selection tools to be integrated to GWAS. Variety of computational artificial intelligence techniques, such as Decision Trees, Artificial Neural Network, Bayesian Classification methods and hybrid models are being studied and their performances are benchmarked according to the efficiency of the methods in predicting disease conditions. The access to the various dbGaP data (such as Alzheimer's disease, autoimmune disorders, cancer and schizophrenia) is allowing us to test the performance of these computational approaches developed with different disease conditions on a wide platform.

**Results and Conclusions:** The investigation of analytical approaches is essential for the improvement and development of computational tools for the interpretation of GWAS results and identification of specific disease associated SNPs. Acceleration in the applications of high-throughput technologies, along with the development of algorithms and methodologies for the analysis of GWAS is expected to yield in new personalized medicine approaches in near future, for the prediction, early detection, prevention and intervention of diseases and to assess drug efficacy to guide dosing and avoid adverse reactions.

**A36: The influence on protein structure of single nucleotide polymorphisms identified in *Mycobacterium tuberculosis*.**

E Vandermarliere\*, T Muth, J Blackburn, L Martens

Elieen Vandermarliere

Department of Biochemistry, Ghent University, Belgium and Department of Medical Protein Research, VIB, Belgium

Cobtact: elien.vandermarliere@ugent.be

**Background:** *Mycobacterium tuberculosis* is the causative agent of tuberculosis, a disease that affects about one-third of the world population. Only a minor group of the infected persons develops an active disease, which is mainly manifested as a pulmonary infection. Like most microorganisms, *M. tuberculosis* displays strain diversity. This diversity is generated within the species through mutation, deletion, duplication and recombination events. But, unlike for many other bacterial pathogens, gene exchange is rare. This method of diversity resulted in the evolution of distinct clonal lineages of *M. tuberculosis* which are not only associated with variation in virulence resulting in a diverse outcome of infection but also with particular geographic regions and human populations.

In this study, we focus on diversity generated by single nucleotide polymorphisms (SNP) resulting in non-synonymous mutations by analyzing the effect of these SNP on protein structure. The SNP in focus are identified from MS/MS spectra.

**Methods:** All *M. tuberculosis* structures available in the PDB were extracted. For each residue within these proteins, the secondary structure and solvent accessible surface were determined.

The SNPs are retrieved with the aid of PEPNOVO from MS/MS spectra from *Mycobacterium* species. The obtained peptides were subsequently matched against the *M. tuberculosis* genome using BLAST followed by identification of the mutated residues.

**Results:** A total of 400 non-redundant structures of *M. tuberculosis* were retrieved from the PDB. This list contains both housekeeping proteins and virulence factors hence a broad spectrum of functions is covered.

From the MS/MS spectra, 957 151 peptides were retrieved of which only 32 128 peptides could fully be aligned with a protein sequence. 925 023 peptides containing one or more mutations were identified.

**Conclusions:** This study not only provides information on the position in structure of SNPs in *M. tuberculosis*. But it also shows the possibilities of using mass spectrometry to help identify the link between strain genotype and resulting different phenotypes. In a next step then, the link between the different outcome of the disease and the differences in genetics between the human populations could be unraveled in this way.

**A37: Comparative genomics of GlnR-mediated transcription regulation in Gram-positive bacteria**

Tom Groot Kormelink<sup>\*1,2,3,5</sup>, Eric Koenders<sup>5</sup>, Yanick Hagemeijer<sup>5</sup>, Lex Overmars<sup>2,4,5</sup>, Roland J. Siezen<sup>1,2,4,5</sup>, Willem M. de Vos<sup>2,3</sup> and Christof Francke<sup>1,2,4,5</sup>

<sup>1</sup>Kluyver Centre for Genomics of Industrial Fermentation, P.O. Box 5057, 2600 GA Delft, The Netherlands. <sup>2</sup>TI Food and Nutrition, P.O.Box 557, 6700AN Wageningen, The Netherlands. <sup>3</sup>Laboratory of Microbiology, Wageningen University & Research Centre, Dreijenplein 10, 6700 HB Wageningen, The Netherlands. <sup>4</sup>Netherlands Bioinformatics Centre, P.O. Box 9101, 6500 HB Nijmegen, The Netherlands. <sup>5</sup>Centre for Molecular and Biomolecular Informatics, NCMLS, Radboud University Nijmegen Medical Centre, P.O.Box 9101, 6500HB Nijmegen, The Netherlands.

Contact: tom@cmbi.ru.nl

**Background:** The assimilation and re-distribution of nitrogen within a cell is essentially controlled within the central metabolic conversions between alpha-ketoglutarate, glutamate and glutamine. The enzymes that catalyze these conversions are glutamine synthase, glutaminase, glutamate dehydrogenase and glutamine alpha-ketoglutarate aminotransferase. Transcription control of the central nitrogen metabolism in Gram-positive bacteria is mediated by three transcription factors: CodY, GlnR and TnrA. GlnR is known to repress gene transcription during growth with excess nitrogen.

**Methods:** We reconstructed the GlnR regulon in Gram-positive bacteria using comparative genomics techniques. The upstream region of the conserved glnRA operon was retrieved for all analyzed species and the promoter region was aligned. Then the conserved sequence upstream of the promoter was collected and used to identify potential binding sites in the genomes using a novel motif search tool called Similar Motif Search.

**Results:** We found a conserved regulatory association of GlnR with glutamine synthetase, and the active transport of ammonium, glutamine and glutamate. In addition, we observed various, previously not reported connections, for instance glutamate dehydrogenase in Streptococcaceae, purine catabolism and the reduction of nitrite in Bacillaceae, and aspartate/asparagine deamination in Lactobacillaceae.

**Conclusions:** Our analyses imply GlnR-mediated regulation in constraining the import of ammonia/amino-containing compounds and the production of intracellular ammonia under conditions of high nitrogen availability. Such a role fits with the intrinsic need for tight control of ammonia levels to limit futile cycling, and/or alternatively, to limit the changes in internal pH.

**A38: Use of domain knowledge for dimension reduction: Application to drug side effects**

Emmanuel Bresso\*, Sidahmed Benabderrahmane, Malika Smail-Tabbone, Gino Marchetti, Arnaud Sinan Karaboga, Michel Souchet, Amedeo Napoli and Marie-Dominique Devignes

LORIA CNRS, Nancy University, INRIA Nancy Grand-Est and Harmonic Pharma (SAS), France

Contact: bressoem@loria.fr

**Background:** Dimension reduction of datasets is an important aspect of data mining research. Indeed, high dimensionality can impair the execution of most data mining programs, especially symbolic methods that deal with nominal attribute values. High-dimension datasets can also lead to the production of numerous and complex patterns which are difficult for experts to interpret. When structured terminologies, or "ontologies", are used for representing attributes, it becomes possible to exploit domain knowledge for dimension reduction without loss of information. The Life Sciences constitute a suitable domain for testing such approaches because numerous structured terminologies are available.

**Methods:** The recently described IntelliGO semantic similarity measure is applied to quantify pair-wise term similarity in a terminology structured as a rooted directed acyclic graph. This allows semantic clustering of terms to be applied in order to use the resulting term clusters as descriptors for data representation. The approach is tested with a set of drugs and their associated side effects collected from the SIDER database. Terms describing side effects belong to the MedDRA (Medical Dictionary for Regulatory Activities) terminology.

**Results:** The 1,288 MedDRA terms involved in the SIDER database were clustered to an optimal collection of 112 term clusters (TCs). Datasets including up to 170 drugs are constructed either with the individual terms or with the term clusters. Searching for closed frequent itemsets yields more abundant and less redundant itemsets with this reduced data representation. The discovery of potentially useful rules to discriminate between two drug categories on the basis of their side effects was possible in a reasonable computation time only with the reduced data representation.

**Conclusions:** The IntelliGO semantic measure was used successfully on the MedDRA terminology to obtain more than a ten-fold reduction of an attribute list. The results of two data mining experiments illustrate the advantage of using this reduced representation. Knowledge-based dimension reduction by semantic clustering can be applied to any dataset in which the attributes belong to a structured terminology.

**A39: "Identify me", says the alien peptide. Combining publicly available mS/MS repositories and clustering tools to identify peptides from non-model organisms.**

Gerben Menschaert\*, Eisuke Hayakawa, Wim Van Crielinge, Geert Baggerman.

Laboratory of Computational Genomics and Bioinformatics (BioBix), Ghent University, Ghent, Belgium.

Contact: Gerben.Menschaert@ugent.be

**Background:** Many genomes of non-model organisms are yet to be annotated. Peptidomics research on aforementioned organisms therefore cannot adopt the commonly used database-driven identification strategy, leaving the tougher de novo sequencing approach as the only alternative in a peptidomics bioinformatics pipeline. The reported tool uses the growing resource of publicly available fragmentation spectra and sequences of model organisms to elucidate the identity of peptides of experimental spectra of non-annotated species. Clustering algorithms are implemented to infer the identity of unknown peak lists based on their publicly available counterparts. The tool can cope with post-translational modifications and amino acid substitutions.

**Methods:** The HomClus-tool is a modular tool consisting of several different modules that are substitutable. First an input parser enables importing many different dataset, e.g. PRIDE experiment data, sequence data from protein resources such as UniProt-KB, or annotated, in-house MS/MS peak lists of peptidomics studies. Afterwards, a clustering module tries to group these “known” spectra with the experimental MS/MS peak lists. Next, an identification module tries to explain the experimental fragmentation spectra by introducing amino acid mutations (up to 3) and post-translational modifications (up to 10), generating many possible solutions. Subsequently, the scoring module scores all candidates (many scoring schemes are implemented) and calculates an expectation value based on the scoring distribution to validate the peptide prediction. The e-value calculation is based on extreme value distribution statistics. Finally, the ranked solutions (based on score) are presented to the researcher for manual inspection.

**Results:** We applied this tool on two locusts (*Schistocerca gregaria* and *Locusta migratoria*) LC-MALDI-TOF-TOF datasets. Compared to a Mascot database search (using the available UniProt-KB proteins of these species), we were able to double the amount of peptide identifications for both spectral sets. Known bio-active peptides from *Drosophila melanogaster* (i.e. fragmentations spectra generated in silico thereof) were used as starting point for clustering, trying to reveal their experimental homologues counterparts.

**Conclusions:** We developed the HomClus-tool to identify MS/MS peak lists of bio-active peptides based on annotated fragmentation spectra of homologues sequences. Next to revealing the peptide identity of bio-active peptides of non-model organisms, this tool could also be successfully adopted in identifying peptide sequences of other types of peptides, e.g. toxins from spiders, scorpions, or cone snails, or also tryptic peptides that are hard to characterize.

**A40: A method to detect mono- and biallelic DNA-methylation from MBD-seq using single nucleotide polymorphisms**

Sandra Steyaert\*, Ayla De Paepe, Geert Trooskens, Simon De Nil, Wim Van Criekinge, Tim De Meyer

Department of Mathematical Modelling, Statistics and Bioinformatics, Ghent University, Ghent, Belgium

Contact: sandra.steyaert@ugent.be

**Background:** Erroneous monoallelic expression is possibly related to certain inherited disorders. DNA-methylation plays a significant role in the regulation of monoallelic expression. The choice of an allele to be used for monoallelic expression can be random or predetermined by imprinting. Here we present a pipeline to screen for genes that exhibit monoallelic DNA-methylation and thus might regulate monoallelic expression. The methylome profile, obtained from a large set of MBD-seq data, together with the SNP profiles from the sequence reads are combined in a statistical framework and allows the detection of regions with monoallelic DNA-methylation.

**Methods:** MBD-seq, which combines enrichment of methylated DNA-fragments by methyl-binding domain (MBD) based affinity purification with massively parallel sequencing (Illumina GAIIx, paired end), was used to profile the DNA-methylation pattern of 229 human samples. The obtained non-duplicate, uniquely mappable sequence reads were screened for SNPs. We developed a new statistical methodology in the R statistical environment that uses these SNP data to detect loci with significant monoallelic DNA-methylation. This methodology was applied for each individual SNP-locus that had been observed with an adequately high frequency, thereby reducing the effect of sequencing errors.

In summary, for each single SNP-locus, the Hardy-Weinberg theorem can be applied on the observed allele frequencies to estimate the expected frequencies of samples that are either mono- or biallelically methylated. Using a permutation approach it can be evaluated whether the observed frequency of samples featured by biallelic methylation is lower than randomly expected. This permutation approach uses the Bayes theorem to adjust for low coverages, i.e. if a specific locus is biallelically methylated in a given sample, it is still possible that only one allele is detected if coverage is low. The permutation test yields p-values, and for loci with  $p < 0.05$ , the presence of monoallelic DNA-methylation is called significant.

**Results:** Starting from MBD-seq data, we present a pipeline to screen for significant monoallelic DNA-methylation based on the SNP-profiles within the different methylomes.

**Conclusions:** With state-of-the-art methods, it is now possible to screen for genes that display monoallelic DNA-methylation. Furthermore this information would give us the opportunity to track down genes that are involved in various non-Mendelian inherited genetic disorders. In a next step, we will study the effect of the functional position (e.g. promoter, exon, intron) on the presence of mono-allelic DNA-methylation. Finally, it is our intention to combine the results with RNA-seq data to further evaluate the effect on gene-expression.

**A41: Multivariate Framework for Biomarker Discovery**

Yousef El Aalamat\*, Dusan Popovic, Etienne Waelkens, Bart de Moor

Katholieke Universiteit Leuven, Dept. of Electrical Engineering, (ESAT), SCD-SISTA (BIOI), Kasteelpark Arenberg 10, B-3001 Leuven, Belgium

Contact: yousef.elaalamat@esat.kuleuven.be

**Background:** The rapid advance in high throughput technologies in bio-sciences results in generation of high dimensional datasets, and, as a consequence, feature selection is becoming one of the most important steps in data analysis. It has been widely used in many research areas ranging from statistics to medical sciences. However the traditional methods for feature selection within a univariate framework have shown their limited capability for biomarker discovery and decision support as the found subset are often not consistent/reproducible in follow-up studies due to the fact that association between features is neglected. In this context we developed a new framework to tackle these issues.

**Methods:** We propose an approach that further improves on the standard techniques for feature selection by taking into account the association between features and addressing the consistency issue of the feature subset. This is a computationally intensive method, based on multivariate logistic regression that aims for high classification accuracy next to improved robustness. During each iteration of the algorithm, the best biomarkers are extracted from a bootstrap replicate of the original data set, after which the models containing these biomarkers are finally selected.

**Results and Conclusions:** To demonstrate the potential of the proposed feature selection method, we tested it on a real-life dataset composed of 353 subjects from each the expressions of the blood plasma proteins were measured. It was shown that this approach outperforms the standard feature selection techniques by selecting these features in a multivariate framework and resolving the issue of consistency of the selected subset.



### **A42: Preprocessing approach for single-cell array Comparative Genomic Hybridization**

J. Cheng\*, P. Konings, E. Vanneste, T. Voet, J. Vermeesch, Y. Moreau

Department of Electrical Engineering, ESAT-SCD, Katholieke Universiteit Leuven, Belgium; IBBT-K.U.Leuven Future Health Department, Belgium; Center for Human Genetics, Katholieke Universiteit Leuven

Contact: jiqui.cheng@esat.kuleuven.be

**Background:** Single-cell array Comparative Genomic Hybridization (CGH) is aimed to detect aneuploidies from a single cell instead of a pool of cells. It has the particular property that the test sample is the amplified DNA extracted from a single cell and the reference sample is the non-amplified genomic DNA from multiple cells. Consequently, the artefacts are induced which obscure the real copy number variation (CNV) detection. Unfortunately, the current available preprocessing methods do not take these artefacts into account and give inaccurate results.

**Methods:** We developed a channel-clone normalization approach based on channel and clone-specific correction to preprocess the single-cell array CGH. The channel-clone normalization consists of three steps: standardization of the intensities for each channel, genome composition artefacts correction and recurrent genome artifacts correction. We used simulated data as well as the real single-cell array CGH data to show the impact of the channel-clone normalization approach. We simulated 15 samples including 23 known artificial aberrations. In addition, 7 Epstein-Barr virus (EBV) transformed lymphoblastoid cells were analyzed. True positive rate (TPR) and false positive rate (FPR) of the CNV detection were used to evaluate the performance of channel-clone normalization approach.

**Results:** Among the 23 simulated CNVs, the TPRs using global loess, CGHnormaliter, poplowess, and channel-clone are 0.97, 0.94, 0.92 and 0.96, whereas the FDRs are 0.06, 0.08, 0.08 and 0 respectively. The channel-clone approach outperforms the other normalization methods in the simulation study. For the 7 EBV single cells, the TPRs using global loess, CGHnormaliter, poplowess, Haarseg and channel clone normalization are 0.66, 0.71, 0.71, 0.66 and 0.98 while the FPRs are 0.13, 0.09, 0.15, 0.05 and 0.006. It is obvious that the channel-clone normalization approach clearly improves the performance of single-cell CNV detection.

**Conclusions:** Our channel-clone normalization approach is designed originally for single-cell array CGH experiments. However, it can be extended for the array experiments that suffer from interchannel variation or genome artefacts.

**A43: Analysis of critical transitions in Parkinson's disease;**

Christophe Trefois\*, Paul Antony, Aidos Baumuratov, Sandra Koeglsberger, Olga Boyd, Rudi Balling

Luxembourg Centre for Systems Biomedicine, Luxembourg

Contact: christophe.trefois@uni.lu

**Background:** Parkinson's disease is the most common neurodegenerative movement disorder and is clinically characterized by resting tremor, bradykinesia and cogwheel rigidity. The disease affects 1-2% of the global population with prevalence in the people above 65 years of age. The main pathological hallmark of Parkinson's disease is a progressive loss of dopaminergic neurons in the substantia nigra. Therefore, one important challenge is to improve the understanding of regime shifts between health and disease states.

Improving predictions of critical transitions triggering the onset of parkinsonian phenotypes could contribute to the improvement of preventive treatments.

**Methods:** Based on cellular models, we will use the mathematical concept of critical transitions to create a toolbox for predicting tipping points towards cellular Parkinson's disease phenotypes, e.g. mitochondrial dysfunction. Experimentally, we will induce and analyze critical transitions in the SH-SY5Y cell line. To do this, we will apply Parkinson's disease relevant chemical and genetic perturbations and analyze multiple scales of the resulting temporal system behavior. We will combine high content imaging with genetic and biochemical data. A significant informatics challenge arises from the aim to perform the analysis of high time-resolved 3D imaging data. We are therefore developing an automated image analysis pipeline that relies on latest technologies and techniques, such as 3D deconvolution and 3D particle tracking. This pipeline will be applied to study parameters, such as mitochondrial dynamics, which include for instance velocity, morphology, and spatial organization.

**Results and Conclusion:** So far, we established the experimental protocols for differentiation and microscopy of the SH-SY5Y cell line. In addition, we established automation protocols for time-resolved qRT-PCR experiments. After preliminary tests, we found that the robot was significantly more consistent compared to manual pipetting, especially when scaling up to 384 well plates. The toolbox developed in the first year of my PhD will provide the basis for the continuation of experiments targeting the analysis of tipping points around Parkinson's disease.

## Abstracts Poster Session B

### **B6: Prediction of cardiac perturbations of non-cardiovascular drugs with My-DTome**

Lu Zhang\*, Yvan Devaux, Daniel R. Wagner, Francisco J. Azuaje

Laboratory of Cardiovascular Research, Public Research Centre for Health (CRP-Santé)

Contact: lu.zhang@crp-sante.lu

**Background:** Cardiac complications of various non-cardiovascular drugs have been reported over the past decade. Traditionally this has been achieved through systematic reviews of clinical trials. However, molecular mechanisms underlying most of cardiac complications are still unknown. Furthermore, there is a need to develop new approaches to predict cardiac effects of non-cardiovascular drugs in both early discovery and post-marketing stages of drug development. Here we report how My-DTome (The Myocardial Infarction Drug-Target Interactome) database can be used to predict cardiac perturbations of non-cardiovascular drugs.

**Methods:** To generate My-DTome database, all drugs approved by European Medicines Agency (EMA) for use in myocardial infarction (MI) and compounds annotated by Pharmacogenomics Knowledge Base (PharmGKB) related to MI were used to retrieve other interacting drugs, drug targets and interacting proteins from different public databases. We built a database hosted on a MySQL server, which can be queried on the Web.

**Results:** My-DTome database comprises 2097 entities (drugs and proteins) and 3958 relationships (interactions), and is available on [www.my-dtome.lu](http://www.my-dtome.lu). Users can search for drug-drug, drug-target and target-protein interactions. To illustrate how My-DTome can help to predict cardiac effects of non-cardiovascular drugs, we considered two antidiabetic drugs, which are associated with cardiovascular complications. In My-DTome, we found different interaction routes between these drugs and significant perturbations of processes regulating heart function, remodeling and blood pressure. We also show mechanisms underlying the cardiac effects of imatinib, a tyrosine kinase inhibitor used in cancer. In My-DTome, imatinib is involved in a module that is statistically associated with heart-specific pathways, such as arrhythmogenic right ventricular cardiomyopathy. Moreover, we characterize its relationship to WT1 (Wilms tumor 1), which is important in heart development.

**Conclusions:** My-DTome provides a systems-based, unbiased strategy to identify cardiovascular complications of non-cardiovascular drugs, and provides insights into molecular mechanisms controlling these side effects.

### **B8: Charting The Methylome**

Geert Trooskens\*, Tim de Meyer, Simon Denil, Wim Van Crielinge

Ghent University

Contact: geert.trooskens@ugent.be

**Background:** Epigenetics, with DNA-methylation as its most stable feature, translates the genetic background into a particular phenotype. Massively parallel sequencing technologies opened up new possibilities for genome-wide profiling of DNA-methylation. Particularly Methyl Binding Domain capturing based Sequencing (MethylCap-Seq) is a low-cost, high-resolution technology to uncover DNA-methylation in a truly genome-wide manner and is becoming increasingly popular.

**Methods:** To chart the map of the methylome, we used raw MethylCap-Seq data of 80 different samples, including different healthy tissues, cell lines and tumor samples. Since no normalization procedures are applied, artefacts are avoided. A Poisson background model is used to identify significantly methylated regions. A conservative set of rules was derived that identifies adjacent methylation prone regions in a single region.

**Results:** Based on this methodology, we provide a reference map of ~1.5 million methylation cores. Together they make up about 10.4% of the human genome and 40% of the approximately 28 million human CpGs dinucleotides. Validation by a different MBD kit and targeted bisulfite sequencing data indicates that the Map of the Human Methylome is approximately 95% complete.

**Conclusions:** We found that although CpG-islands (CGIs) and exon regions are highly enriched in methylation cores with high methylation levels, they show less variability between samples compared to promotor, intergenic and intronic regions. The accuracy of the methylome map will increase with more samples from different tissues and diseases. Comparing the map of the methylome with expression and other data such as histone marks will enable functional annotation of the methylation prone regions, providing a better understanding of the mechanisms involved in epigenetic regulation. This approach is a flexible methodology that can be ported to other genome wide high-throughput methods such as third generation sequencing technologies.

**B9: Predicting the cholinesterases binding sites for Plant derived inhibitors: template to design a drug for Symptomatic treatment of alzheimer's disease**

Venkatachalam Lakshmi\*, Rathanam Boopathy

DRDO-BU Center for Life Sciences, Bharathiar University, Coimbatore, India \_  
641046

Contact: mishlabio@gmail.com

**Background:** Central cholinergic system is considered as most important neurotransmitter system involved in regulation of cognitive functions. Cholinergic neuronal loss in hippocampal region is main feature of Alzheimer's disease (AD) and enhancement of central cholinergic activity is presently the mainstay of pharmacotherapy of senile Alzheimer type of Dementia. Cholinesterase inhibitors (ChEIs) are the only drugs so far approved for AD treatment. Therapeutically used inhibitors include natural phytoconstituents and synthetic compounds, synthesized based on template natural phytoconstituents. Also ChEIs have potential roles in Vascular Dementia, Parkinson's Disease and Down Syndrome treatment.

**Methods:**

- i) In silico docking study of plant based inhibitors onto human acetylcholinesterase (AChE)protein.
- ii) In silico docking study of plant based inhibitors onto human butyrylcholinesterase (BChE)protein.

**Results:**

- i) The docking algorithm, LigandFit differentiates active ligands as inhibitors from inactive ones.
- ii) Of all scoring functions namely DockScore, LigScore1, LigScore2, Jain, PLP and PMF, DockScore is the efficient scoring tool in predicting best binding inhibitor.
- iii) Ser203, His447, Arg463, Tyr124, Trp86, Glu202, Phe295, Phe297 and Tyr337 are the residues important in inhibitor binding to AChE.
- iv) Ser198, Glu197, Asp70, Thr120, His438, Trp82, Gly116, Gly115, Phe329 and Tyr332 residues are important in inhibitor binding to BChE.

**Conclusions:** The structure of best docked inhibitors can be used as template for designing new, selective and powerful AD drugs.

### **B10: Mining the rest-fraction in RNA-seq experiments**

Simon Denil\*, Tina Kyndt, Annelies Haegeman, Geert Trooskens, Tim De Meyer,  
Wim Van Crielinge, Godelieve Gheysen

Laboratory of Computational Genomics and Bioinformatics (BioBix), Ghent  
University, Ghent, Belgium

Contact: [simon.denil@ugent.be](mailto:simon.denil@ugent.be)

**Background:** The increasing availability and affordability of sequencing has led to widespread acceptance of sequencing based experimentation in a variety of biological and clinical settings. These experiments vary from genomic and epigenomic applications to expression based analyses (ie RNA-seq). In the latter context one of the most cited properties used to argue the superiority of sequencing based technologies is its potential to detect novel transcripts both in model and non-model organisms for which complete genomes are rarely available. In practice this process is often limited to transcripts constructed from reads which readily map to a reference genome. There are however several means available to identify novel transcripts beyond the typically used methodologies.

**Methods:** We demonstrate this premise using a RNA-seq dataset generated to study the influence of nematodes (root parasites) on the host's gene expression pattern (*Oryza sativa*, rice). Raw sequence reads were processed in a pipeline consisting of "TopHat" (mapping), "Cufflinks"(reference based transcript assembly), "Velvet" (de novo and guided assembly), "baySeq" (expression profiling) and a few Perl scripts. After initial mapping and transcript assembly we remapped the reads using the additional information contained in the putative novel transcripts. After the second round of mapping all unmapped reads were used for de novo transcriptome assembly. Putative novel transcripts (nTARs) were BLASTed against the *Oryza sativa* genome and NCBI nucleotide databases. Parasite transcriptome contig assemblies obtained from long sequencing reads were available as a means of external validation.

**Results:** Using only open-source tools we were able to establish expression profiles for 34,356 known genes and over 8,000 putative nTARs. Sets of both known genes and putative nTARs demonstrated highly specific transcription profiles related to the plants\_ physiology and the infection process. We were further able to detect transcripts most likely originating from the parasite transcriptome which was partially captured due to the nature of the sample processing. Many of these could be detected without the aid of the contig assemblies.

**Conclusions:** With this poster we present a model workflow that allows researchers to extract biologically relevant information from sequencing data beyond results obtained by typical bio-informatics approaches. Using existing, publicly available analysis tools we demonstrate that qualitative if not quantitative information may be gained from the data rest-fraction after it has undergone "standard" analyses.

### **B11: Functional classification and co-expression analysis of genetically imprinted Genes**

M.Hamed\*, Siba Ismael, Martina Polsky, Volkhard Helms

Chair of Computational Biology, Saarland University Germany

Contact: mhamed@bioinformatik.uni-saarland.de

**Background:** Imprinted genes play important roles in development and growth both pre- and postnatally by acting in fetal and placental tissues (Morison et al. (2005)). Interestingly, there appears to exist a general pattern whereby maternally expressed genes tend to limit embryonic growth and paternally expressed genes tend to promote growth. (Reik and Walter, 2001)

As the phenomenon of genomic imprinting is an important evolutionary facet of mammals with placentas, it is of great interest to identify which sorts of cellular and developmental processes of developing and/or mature organisms are subject to control by imprinted genes. Additionally, to identify the other bi-allelic genes which are involved in the same developmental biological processes and being regulated by the same transcription factors.

**Methods:** We aimed in this study at characterizing the cellular roles of imprinted genes in an unbiased, data-driven approach. For this, we used the gene annotations of the Gene Ontology (GO). First, we analyzed which terms of the Gene Ontology are enriched in the full set of all imprinted genes when compared to the set of all human genes. Then it is of particular interest to analyze which of these functions are controlled by the sets of maternally and paternally expressed separately. Furthermore, we looked at the enrichment for transcriptional factors targets to determine which bi-allelic genes. Then we investigated what cellular functions are enriched in the Bi-allelic genes, which share same binding motifs and being regulated by same transcription factors of imprinted genes.

Finally, we performed a computational analysis of microarray expression data of imprinted genes in human in a variety of adult issues and tested the co-expression with the normal bi-allelic genes.

**B12: Micro peptides as a new class of bio-active peptides in Eukaryotes**

Jeroen Crappé\*, Gerben Menschaert, Geert Trooskens, Joachim Deschrijver, Geert Baggerman, Wim Vancrickinge

Biobix, Ghent University

Contact: jeroen.crappe@ugent.be

**Background:** For a long time it was assumed, partially due to informatics and statistical limitations, that a protein-coding gene had to be at least 100 AA in length. In recent years it came to knowledge that genes exist smaller than 100 AA, encoding for small peptides. Those small peptides, from here on defined as micro-peptides are translated directly from their small open reading frames (sORF). In 2007 an evolutionary conserved sORF gene, polished rice or tarsal-less was identified in *Drosophila melanogaster*, playing its role in early developmental stages. Thanks to advances in sequencing and bioinformatics tools, it is now possible to scan the genome of different species unceasingly deep, e.g. in a search for this type of small peptides.

**Methods:** The *Drosophila melanogaster* genome was scanned for sORFs with the sORFfinder tool. It makes use of a hidden markov model predicting the coding potential of possible open reading frames genome-wide. Furthermore bi-directional mRNA sequencing of specific embryonic and larval stages of *Drosophila melanogaster* was performed. An in-house designed pipeline (Perl and MySQL based) subsequently analyses the available sORFs on their potential of encoding micro-peptides. In this pipeline, properties of already discovered micro-peptides are taken into account: sequence conservation in other closely related *Drosophila* subspecies for example seem to play an important role. The predicted micro-peptides are finally ranked by means of a customized score, combining all derived properties,

**Results:** Based on the aforementioned pipeline, a list of possible micro-peptides with high coding potential was constructed. All predicted micro-peptides are highly conserved on both DNA and AA level, and moreover have a favorable synonymous versus non-synonymous mutation rate. Next, they are supported by experimental evidence by means of (bidirectional) RNAseq data or Ensembl ncRNA gene annotations.

**Conclusions:** Micro-peptide research is still in its infancy. Parallel to the discovery of more micro-peptides, our knowledge will also grow. It will become easier to discover and annotate new members of this new class of bio-active peptides. We strongly believe that micro-peptides herald important functions and are, in the same way as microRNAs, an important but long time overlooked class of bio-active molecules.



**B13: Probing the active site of Cholinesterases for its associated aryl acylamidase activity: molecular docking of substrate analogues**

Venkatachalam Lakshmi\*, Rathanam Boopathy

DRDO-BU Center for Life Sciences, Bharathiar University, Coimbatore, India \_  
641046

Contact: mishlabio@gmail.com

**Background:** Cholinesterases (ChEs) in general refer to butyrylcholinesterase (BChE) and acetylcholinesterase (AChE). In addition to their esterase activity, they also show a genuine aryl acylamidase (AAA) activity. This catalytic function is observed in vitro by certain aromatic amides hydrolysis. A wide variety of substrate analogues of ortho nitroacetanilide were employed for assaying AAA activity of ChEs. The biological significance of this catalysis is unknown. Although AAA activity is widely accepted as a secondary catalytic function, its mediation by the same catalytic triad is not demonstrated for want of their enzyme-substrate crystal structure. While esterase catalytic triad: Glu 327, His 438 and Ser 198 of BChE, Glu 334, His 447 and Ser 203 of AChE have been characterized, that of AAA is pending.

**Methods:** In the present study an attempt has been made to probe the amino acid residues involved in AAA activity using in silico docking (DiscoveryStudio2.0) of several substrates on BChE and AChE.

**Results and Conclusions:**

The conclusions drawn by observing LigScore1, Ligscore2, Piecewise Linear Potential, Jain and Potential of Mean Force values are:

- 1) LigScore1 was the most efficient scoring function in predicting favourable and unfavourable docked poses of ligands towards identifying the active substrates. Receiver operating characteristics curve analysis validated Ligscore1 as efficient tool in predicting the correct docked poses with ChEs as receptor protein.
- 2) Ser 198 of BChE and Ser 203 of AChE was the only nucleophile of catalytic triad to which favourable docked poses of all active substrates interacted with close proximity.
- 3) In silico mutation analysis showed Ala 328, Trp 82, Trp 231 and Phe 398 to be essential for substrate binding in the active site of BChE.
- 4) In silico pKa calculations predicted Glu 197 and His 438 of BChE, Glu 202 and His 447 of AChE to be the residues involved in AAA active site, suggesting that the His residue is commonly shared for both esterase and amidase activity, while the predicted Glu is exclusively required for AAA activity. The Glu of esterase activity (Glu 327 of BChE, Glu 334 of AChE) is thus excluded for the AAA activity of ChEs.
- 5) We predict the necessity and absolute essentiality of the nitro group at ortho position of all active substrates that initiates a hydrogen bond formation with Gly residues within the active site of ChEs and helps in aligning scissile amide bond in the active site for their hydrolysis.

**B14: Predicting the Interactions between G-Protein Coupled Receptors:  
Computational and Experimental Approaches**

Mehmet Emreiahin, Tolga Can, Cagdas D. SON\*

Biyological Sciences Department, METU, Turkey

Contact: cson@metu.edu.tr

**Background:** G protein-coupled receptors (GPCRs) are membrane proteins that mediate physiological response to a diverse array of stimuli. In humans they mediate the action of hundreds of peptide hormones, sensory stimuli, odorants, neurotransmitters, and chemokines. GPCRs also are targets for ~50% of all currently marketed pharmaceuticals. These receptors have traditionally been thought to act as monomeric units. However, recent evidence suggests that GPCRs may form dimers as part of their normal trafficking and function. While the formation of GPCR dimers/oligomers has been reported to play important roles in regulating receptor expression, ligand binding, and second messenger activation, less is known about how and where GPCR dimerization occurs. We hypothesize that dimerization occurs early after biosynthesis in the endoplasmic reticulum, suggesting that it has a primary role in receptor maturation.

**Methods:** We are addressing this issue using confocal microscopy and split green fluorescence protein (GFP) to monitor GPCR interactions within discrete intracellular compartments of intact living cells as an experimental approach. Time-lapse confocal imaging is being used to determine the ability of two fragments of enhanced GFP (EGFP) to reassemble and fluoresce when fused to interacting GPCRs. We also take advantage of the real-time monitoring of fluorescence resonance energy transfer between monomeric (m)EGFP and mCherry tagged full length G-protein coupled alpha-factor receptor within discrete regions of the endoplasmic reticulum, Golgi and plasma membrane to understand the molecular mechanisms important in receptor dimerization. Meanwhile we have a computational approach to address the same problem. We are analyzing the known GPCR interactions in the literature and trying to figure out the important structural elements that play a role in dimerization. Both projects are going in tandem, the results we obtain from computational studies are being tested by our experimental setup and the prediction methods are being revised using the results from the experiments carried out in the wet LAB.

**Results and Conclusions:** The results of this project will contribute to our understanding of how and where GPCRs dimerize and importance of receptor dimerization in ligand induced receptor activation.

**B15: “Last In First Out”, gain and loss of the Intra Flagellar Transport components**

John van Dam\*, Martijn Huynen

Centre for Molecular and Biomolecular Informatics, Nijmegen Centre for Molecular Life Sciences, Radboud University Nijmegen Medical Centre, Nijmegen, the Netherlands

Contact: [jvandam@cmbi.ru.nl](mailto:jvandam@cmbi.ru.nl)

**Background:** The Intra Flagellar Transport (IFT) is an ancient protein complex that facilitates active and selective trafficking of proteins and molecules across the length of the eukaryotic cilium. IFT components are often involved in ciliary dysfunction, which causes a plethora of syndromes and complex diseases, collectively called ciliopathies. Although a variety of coarse grained sub complexes have been defined experimentally, the structure of the IFT is still largely unknown. Reconstructing the evolution of the various subcomponents provide insight into the evolution and structural organization of the IFT.

**Methods:** We use comparative genomics to investigate the origin of the IFT and use phylogenetic profiles to elucidate intra complex relationships between the various components of the IFT protein complexes.

**Results:** IFT components are highly conserved in ciliated species and originated from before the Last Eukaryotic Common Ancestor. Homologous relationships between subunits show a checkered origin via gene duplication of the IFT-A, IFT-B and BBSome complexes that comprise the IFT. Asymmetry in phylogenetic occurrences indicates that loss of the cilium in some species occurred in distinct steps in time and provides clues into the structural composition of these large protein complexes.

**Conclusions:** The IFT is an evolutionary conserved critical component of the eukaryotic cilium. Even though the IFT we observe in current-day species was already present in the Last Eukaryotic Common Ancestor. We elucidate the gradual acquisition of the IFT components and the inverse gradual loss in species that have lost the cilium. Therefore the observed modularity in IFT complexes behaves much like a LIFO structure. Fine grained physical or functional relations between IFT subunits is reflected in asymmetric phylogenetic profiles between these components and in the future should allow us to resolve the structural organization of the IFT complex.

**B16: Are REPs genetic insulators that enable differential regulation of gene expression in bacteria?**

Lex Overmars<sup>\*1,2,4</sup>, Tom Groot Kormelink<sup>1,2,3</sup>, Roland Siezen<sup>1,2,4</sup> and Christof Francke<sup>1,2,3,4</sup>

<sup>1</sup>TI Food and Nutrition, The Netherlands, <sup>2</sup>Radboud University Medical Centre, Centre for Molecular and Biomolecular Informatics, The Netherlands, <sup>3</sup>Kluyver Centre for Genomics of Industrial Fermentation

Contact: l.overmars@cmbi.ru.nl

**Background:** Repetitive Extragenic Palindromic elements (REPs) are short palindromic sequences that were first identified in *E. coli* and closely related enteric bacteria. Recently, REPs were identified in more diverse bacterial taxa. REPs exhibit some remarkable characteristics. They (i) are almost exclusively found in the intergenic space, in which they are often arranged in repeats called BIMEs, (ii) occur in high abundance, up to 500-1000 REPs in some species, occupying a substantial portion of the intergenic space, (iii) are highly conserved within a genome. Various functional links have been proposed in literature. For instance, it was shown that REPs play a role in stabilizing mRNA, i.e. that REP-containing transcripts are less prone to degradation. Specific REPs were also shown to act as binding sequence for either DNA polymerase or DNA gyrase. However, none of the related literature provides a common functional denominator for the complete set, let alone a satisfying mechanism of action. We therefore decided to investigate the commonality using a comparative genomics approach.

**Results:** *E. coli* REPs were identified using a redefined motif based on a conserved 29bp sequence. We observed a biased distribution of REPs with respect to the ORFs, in which REPs are not found between divergent gene-pairs and predominantly located between convergent gene-pairs. A set of 950 publicly available microarrays was used to explore the discordant effects of REPs on transcription under various conditions. This analysis revealed a clear trend towards a decrease in correlation of expression between the genes on either side of a REP versus other neighboring genes. We demonstrate that in 40 out of 950 microarrays the gene-REP-gene pairs behaved significantly different from pairs that lack a REP. These arrays represented the transcriptional response to a certain kind of stress, i.e. related to biofilm formation, aerobiosis and stress specific sigma factors. In addition, we found that REP-containing operons frequently possess an alternative promoter which is known to modulate the stress responses.

**Conclusions:** This study shows that REPs potentially have a global role in regulation of differential expression. Our results imply that REPs enable differential expression specifically in cases where elevated pressure on DNA supercoiling can arise, i.e. expression of convergent gene-pairs and transcription regulated by an alternative promoter. Using a large set of arrays under various conditions we can also couple our findings to physiology; we propose that the phenomenon of REP-enabled differential expression is linked to the bacterial response to various stresses.

**B17: Automated comparative genome annotation in prokaryotes: a halt to error propagation?**

Thomas H. A. Ederveen<sup>\*1</sup>, Amy de Bruin<sup>2</sup>, Brechtje Hoegen<sup>2</sup>, Bernadet Renckens<sup>1</sup>,  
Roland J. Siezen<sup>1,4,5</sup>, Sacha A. F. T. van Hijum<sup>1,3,4,5</sup>

<sup>1</sup>Centre for Molecular and Biomolecular Informatics, Radboud University Medical Centre, the Netherlands. <sup>2</sup>HAN University of Applied Sciences, the Netherlands.

<sup>3</sup>NIZO food research, the Netherlands

Contact: T.Ederveen@cmbi.ru.nl

**Background:** With next-generation sequencing, genome sequences are generated faster than the capacity to adequately annotate these assembled genome sequences. Automated genome annotation engines facilitate genome annotation. However, these engines suffer from inaccuracy in gene coordinate- and function prediction compared to manually annotated and/or curated efforts by scientists. In public sequence databases, many erroneous gene annotations exist making database error propagation a great concern as annotation engines rely on these databases for their predictions. Many automated annotation pipelines exist, but the choice of what engine to use seems an arbitrary one.

**Methods:** In this work, our goal is to significantly reduce the number of genes that need to be manually checked and curated. To this end, we combine for a given genome sequence the results of multiple automated annotation engines into a consensus gene coordinate- and function prediction. We hypothesize that this consensus prediction will be more accurate compared to the use of solely one individual engine. We present a pipeline for processing data from popular, publicly available automated genome annotation engines for prokaryotes: BASys, IGS, ISGA, RAST and xBASE. New engines can be added relatively straight-forward. It compares coordinate- and function prediction calls made by the various engines and provides the user with weighted gene coordinate- and function predictions. We use rules based on prior-knowledge in addition to majority voting. These rules are based on errors that frequently occur with a given annotation engine for certain genes. They allow favoring certain “trusted” predictions based on prior-knowledge over those provided by majority voting. These *\_trusted\_* predictions are currently based on three phylogenetically diverse bacteria with manually curated genome annotations: *Moraxella catarrhalis* RH4, *Lactobacillus plantarum* WCFS1 and *Lactococcus lactis* KF147.

**Results and Conclusions:** By analysis of annotations from different engines for these three strains we can: (i) (for specific gene sets) identify automated annotation engine specific biases in their method of gene coordinate- and function prediction; (ii) build up a vocabulary of annotation terms that are equivalent across different engines, termed a translation list; and, (iii) by data analysis and pattern recognition, derive specific prior-knowledge based rules that allow circumventing error propagation and if relevant, remove annotation engine specific bias. Our comparative annotation pipeline provides more accurate gene coordinate- and function predictions, leaving the curator with only a subset of genes to be manually checked and/or curated.

**B18: Dynamic co-regulation of miRNAs and mRNAs following cytokine stimulation of melanoma cells**

Susanne Reinsbach\*<sup>1</sup>, Peter V. Nazarov<sup>2</sup>, Martina Schmitt<sup>1</sup>, Demetra Philippidou<sup>1</sup>,  
Nathalie Nicot<sup>2</sup>, Iris Behrmann<sup>1</sup>, Laurent Vallar<sup>2</sup>, Stephanie Kreis<sup>1</sup>

<sup>1</sup>Universtiy of Luxembourg, L-1511, Luxembourg; <sup>2</sup>Microarray Center, CRP-Santé,  
L-1526 Luxembourg

Contact: susanne.reinsbach.001@student.uni.lu

**Background:** MicroRNAs (miRNA), a class of small non-coding RNAs, are ubiquitously expressed in almost every tissue and play essential roles in a variety of processes including cellular differentiation, proliferation and apoptosis. In recent years, miRNAs have emerged as key post-transcriptional regulators of gene expression. A single miRNA can control the expression levels of many mRNA target genes and vice versa a given mRNA can be regulated by several miRNAs. To elucidate the exact role of individual miRNAs or groups of related miRNAs in a given cell, their regulated target mRNAs need to be identified. In this study, we have investigated the dynamic regulation of miRNAs and mRNAs following cytokine stimulation of melanoma cells.

**Methods:** We performed a detailed time course experiment to generate miRNA and mRNA expression profiles using microarray technology. Briefly, we activated the transcription factor STAT1 by IFN-gamma; stimulation to identify STAT1-regulated miRNAs and target genes. The focus was on temporal expression patterns of individual miRNAs as well as on groups of miRNA that are co-regulated. We used different bioinformatics methods for data analysis available as packages in R/Bioconductor such as clustering of time series data (Mfuzz), a Bayesian based method (BETR), the package timecourse as well as the software EDGE and independent component analysis (PearsonICA). Furthermore, co-expression analyses using CoExpress was performed and the miRNA-mRNA negative correlation events were compared with results of TargetScan - a tool for prediction miRNA target genes.

**Results:** We identified 23 differentially expressed miRNAs regulated over time, which were potentially directly or indirectly regulated by the STAT1 transcription factor. Remarkably, ICA outperformed other methods in elucidating dynamically regulated groups of differentially expressed miRNA. Co-expression analysis allowed for detecting sub-networks consisting of mRNA and miRNA with their correlated expression profiles. Among negative miRNA-mRNA regulations, some (178 miRNAs) were found to also be predicted by TargetScan. The majority of co-regulated miRNAs and mRNAs have the tendency to be up- or down-regulated gradually over time (between 24 and 72h).

**Conclusions:** We were able to identify differentially expressed miRNAs as well as mRNAs in time series microarray experiments. Different groups of miRNAs exist that are gradually up-regulated following a transcription stimulus; others are coming up late suggesting additional indirect regulatory loops. Furthermore, we showed that ICA is a promising new approach suitable for the time-series microarray data analysis while co-expression analysis of inversely correlated expression profiles can be used for improved miRNA target genes predictions.

**B19: Critical assessment of candidate gene prioritization methods**

Daniela Bîrnigen<sup>1,2</sup>, Léon-Charles Tranchevent<sup>\*1,2</sup>, Francisco Bonachela-Capdevila<sup>3</sup>, Koenraad Devriendt<sup>4</sup>, Bart de Moor<sup>1,2</sup>, Patrick De Causmaecker<sup>3</sup>, and Yves Moreau<sup>1,2</sup>

<sup>1</sup>Department of Electrical Engineering, KULeuven, BE; <sup>2</sup>IBBT-K.U.Leuven Future Health Department, BE; <sup>3</sup>Department of Computer Science, KULeuven, BE; <sup>4</sup>Center for Human Genetics, KULeuven, BE

Contact: ltranche@esat.kuleuven.be

**Background:** Gene prioritization aims at identifying the most promising candidate genes among a larger pool of candidates so as to maximize the yield and biological relevance of further downstream validation experiments and functional studies. During the past few years, several gene prioritization methods have been defined and some of them have been implemented and made available through freely available web tools.

**Methods:** In this study, we aim at comparing the predictive performance of eight publicly available prioritization methods on novel data. We have performed an analysis in which 42 recently reported disease gene associations from literature are used to benchmark these tools before the underlying databases are updated. Our approach mimics a novel discovery, and therefore the estimation of the performance is more realistic than when benchmarking through cross-validation on retrospective data.

**Results:** Our benchmark indicates that although the observed performance is slightly lower than for benchmarks on retrospective data, several methods can still efficiently identify the novel disease genes. There are however marked differences, and methods that rely on more advanced data integration schemes appear more powerful.

**Conclusions:** Although our validation is not very large, it can be observed that computational methods can efficiently identify novel disease genes. It shows that these quickly maturing methods can be used to provide an additional line of evidence when demonstrating the association of a gene to a disease or its biological function.

**B20: A kernel based framework for cross-species candidate gene prioritization**

Shi Yu<sup>1,2</sup>, Léon-Charles Tranchevent<sup>1,2</sup>, Sonia Leach<sup>1,2</sup>, Pooya Zakeri<sup>\*1,2</sup>, Bart De Moor<sup>1,2</sup> and Yves Moreau<sup>1,2</sup>

<sup>1</sup>Department of Electrical Engineering ESAT-SCD, Katholieke Universiteit, Leuven, Leuven, Belgium., <sup>2</sup>IBBT-K.U.Leuven Future Health Department, Leuven, Belgium

Contact: Pooya.Zakeri@esat.kuleuven.be

**Background:** In biology, there is often the need to prioritize large list of candidate genes to further assay only the most promising candidate genes with respect to a biological process of interest. In the recent years, many computational approaches have been developed to tackle this problem efficiently by merging multiple genomic data sources.

**Methods:** We define the prioritization problem as a MKL task, each kernel is a normalized linear kernel and corresponds to a single genomic data source. The optimization task is then solved using a one class SVM (1-SVM) algorithm. Basically, the training genes are used to model the biological process under study (i.e., to define the separating hyperplane), the candidate genes are then scored based on their distance to this hyperplane. The prioritization is performed independently for each species through homology. Results are then integrated through a Noisy-Or like model.

**Results:** As a proof of concept, we have benchmarked our kernel based method with 14 biological pathways and 28 diseases using a leave-one-out cross-validation procedure. Altogether, results indicate that our cross-species model is conceptually valid, and that combining genomic data cross-species can enhance the gene prioritization performance. In particular, our kernel based method outperforms our previous method based on order statistics.

**Conclusions:** We present a gene prioritization method based on the use of kernel methods and prove that it outperforms our previous method based on order statistics. In addition, the method supports data integration over multiple related species. We have also developed a web based interface termed MerKator that implements this strategy and proposes candidate gene prioritization for 5 species.



**B21: A novel approach for the identification of miRNAs**

Bart Aelterman\*, Peter De Rijk, Jurgen Del-Favero

Applied Molecular Genomics Unit, VIB Department of Molecular Genetics, VIB,  
University of Antwerp, Belgium

Contact: bart.aelterman@gmail.com

**Background:** miRNAs are small (~22 nt) noncoding RNAs that are important regulators of gene expression. The latest release of miRBase (17.0) contains 1424 human miRNA sequences of which several are already implicated in human diseases such as Tourette's syndrome, autism, schizophrenia, .... The goal of this study is to develop a new computational approach to identify novel human miRNAs in a high-throughput manner.

**Methods and Results:** We developed a sensitive miRNA prediction tool that searches for local stable RNA hairpins and predicted novel miRNA genes with this tool in the entire human genome. In order to validate these predictions, we developed an in silico high-throughput validation method using data from 88 publicly available small RNA Sequencing runs from 20 different human tissues. Data from these sequencing runs are adapter trimmed and quality filtered. Next, the RNA sequences are mapped on the miRNA gene predictions using BLAST. As millions of predictions were made, an implementation was required that would enable parallel computing on a computer cluster. Validated miRNAs were expected to be covered by at least 10 reads coming from one arm of the hairpin. Additional filters were applied to reduce false positives due to sequence similarity with other transcriptionally active regions. Also the filters described by the miRBase authors were applied here. With these tools, we are able to identify a substantial number of novel miRNAs and generate insights into their expression patterns. Also processes such as the addition of non-template nucleotides and presence of isomirs can be studied.

**Conclusions:** We were able to identify a substantial number of novel miRNAs expressed in different human tissues using a novel computational approach. Many of these are tissue specific and that the intragenic complexity due to post-transcriptional modifications and processing is much higher than previously expected. Our results can have a significant impact on human genetics as many of these novel miRNAs could be novel candidate genes for diseases or biomarkers for diseased tissue states.

**B22: Arrayanalysis.org: friendly solutions for standardised microarray analysis**

Lars Eijssen<sup>\*1</sup>, Magali Jaillard<sup>1</sup>, Michiel Adriaens<sup>1</sup>, Anwesh Dutta<sup>1</sup>, Martina Kutmon<sup>1</sup>, Philip de Groot<sup>2</sup>, Chris Evelo<sup>1</sup>

<sup>1</sup>Department of Bioinformatics, BiGCaT, P.O. Box 616, 6200 MD Maastricht University, NL; <sup>2</sup>Nutrition, Metabolism and Genomics Group, P.O. Box 8129, 6700 AA Wageningen, NL

Contact: l.eijssen@maastrichtuniversity.nl

**Background:** Together with the firm establishment of microarray technology in systems biology research and the accumulation of microarray experiments in public repositories, an extensive set of bioinformatics procedures has been developed to deal with the data, including quality control (QC) and processing of signals as well as interpretation of study outcome. This holds especially for Affymetrix gene expression chips. In contrast to this, there is currently a lack of a standardised, accessible, and user-friendly open-source software tool. On one hand, Affymetrix and other companies provide many integrated commercial and closed source software tools. On the other hand, many public contributions are available from the scientific community, but lack of an environment that bundles all available functionality compromises accessibility for non-expert users. This even leads to a tendency among researchers to not adopt available methods and perform suboptimal analyses.

**Methods:** One of the largest collections of public bioinformatics contributions in the microarray field is the Bioconductor repository, to be used conjointly with the open-source statistical scripting language R. We automated methods available from several libraries within this repository, extended with home-built functionality, in a workflow performing QC, pre-processing, statistical analysis, and pathway analysis of Affymetrix expression sets.

**Results:** We implemented a complete workflow that can be accessed using the user-friendly interface provided at [www.arrayanalysis.org](http://www.arrayanalysis.org). The site offers user guides, technical documentation, example datasets and explanation of the output. We also provide open source R wrapper functions for running the modules locally. The QC step generates about twenty images assessing sample, hybridisation and overall signal quality, signal comparability and biases diagnostics, and array correlation. The pre-processing step includes gene re-annotation and proposes several normalisation methods. The statistics module computes the expression changes of genes between experimental groups of choice and their significance. The pathway module provides lists of the most affected cellular processes and visualisation of the data on those based on PathVisio [[www.pathvisio.org](http://www.pathvisio.org)] functionality. Availability of the integrated workflow facilitates implementation of current standards, and introduction of new standards in the future. It assists newcomers to the field in processing their microarray datasets. For experienced data analysts, it facilitates running all the established functionality in one go.

**Conclusions:** We provide the scientific community with an easy-accessible, open-source and extensively documented workflow for the QC, processing, and interpretation of Affymetrix expression datasets. Availability of such a method enhances the application and standardisation of up-to-date microarray data analysis.

**B23: Comparative studies of genome-wide DNA methylation arrays and gene expression data: a breast cancer as an example;**

Singhal Sandeep\*, K., Desmedt Christine, Ignatiadis Michail, Sotiriou Christos, Michiels Stefan

Breast Cancer Translational Research Laboratory, Institut Jules Bordet, Université Libre de Bruxelles, Bruxelles, Belgium

Contact: sandeep.singhal@bordet.be

**Background:** Epigenetics refers to heritable changes in gene expression that occur without alteration in DNA sequence. DNA methylation plays an important role in X-chromosome inactivation, tumor-suppressor gene silencing, chromosomal instability and even carcinogenesis. Being a recent area not much information is available and - at least for the time being - mechanisms that control how, when and where genes are expressed is still not well understood. Our objective was to review statistical issues arising from the low-level analysis, such as preprocessing of DNA methylation data, up to the high-level analyses such as pathway analysis and the study of the complex association between gene expression and DNA methylation.

**Methods and Results:** We highlight some statistical methods for genome wide methylation data analysis with an emphasis on what new information is gained from several breast cancer studies for which both DNA methylation array and gene expression array data are publicly available. First, we describe the basic characteristics of methylation array data and preprocessing procedures. Due to the non-normal distribution of methylation data, nonparametric approaches seem better tailored as compared to many parametric methods typically used in gene expression studies. We illustrate some of the nonparametric approaches that can be used for the specific goals of different studies: class prediction, class comparison and class discovery, and for elucidating relationships between DNA methylation data and gene expression. For illustration, the proposed non-parametric class prediction method has a low cross validation error for classifying breast cancer subtypes. Finally, when measuring the methylation signal as a function of the distance to transcription start site, we demonstrate that for the majority of genes, their methylation and gene expression are highly inversely correlated within the promoter region.

**Conclusions:** In conclusion, the nonparametric methods presented in this overview seem more appropriate to the analysis of DNA methylation data and can be straightforwardly applied to other studies in order to pin down novel cancer genes whose expression is altered by DNA methylation alone.

**B24: Classification of MCAD deficiency using tandem MS neonatal screening data**

Tim Van den Bulcke

biomina, Antwerp University Hospital, Belgium

Contact: tim.van.den.bulcke@uza.be

**Background:** Newborn screening programs for severe metabolic disorders using tandem mass spectrometry are widely used. Medium-Chain Acyl-CoA dehydrogenase deficiency (MCADD) is the most prevalent mitochondrial fatty acid oxidation defect (1:15,000 newborns) and it has been proven that early detection of this metabolic disease decreases mortality and improves the outcome. In previous studies, data mining methods on derivatized tandem MS datasets have shown high classification accuracies. However, no machine learning methods currently have been applied to datasets based on non-derivatized screening methods.

**Methods:** A dataset with 44,159 blood samples was collected using a non-derivatized screening method as part of a systematic newborn screening by the PCMA screening center (Belgium). Twelve MCADD cases were present in this partially MCADD-enriched dataset. We extended three data mining methods, namely C4.5 decision trees, logistic regression and ridge logistic regression, with a parameter and threshold optimization method and evaluated their applicability as a diagnostic support tool. Within a stratified crossvalidation setting, a grid search was performed for each model for a wide range of model parameters, included variables and classification thresholds. This setup succeeded to meet the following clinical goals: to identify all MCADD cases, to have an interpretable model with few variables and to minimize the number of false positives.

**Results:** The best performing model used ridge logistic regression and achieved a sensitivity of 100%, a specificity of 99.987% and a positive predictive value (PPV) of 33.90% (recalibrated for a real population), obtained in a stratified cross-validation setting. Current state of the art methods achieved PPV values of 3.79% and 6.31%. The results were further validated on an independent test set (sensitivity 100%, PPV 28.67%).

**Conclusions:** Using a method that combines ridge logistic regression with variable selection and threshold optimization, a significantly improved performance was achieved compared to the current state-of-the-art, while retaining more interpretability and requiring less variables. These results indicate the potential value of data mining methods as a diagnostic support tool.

**B25: Capture of an activated receptor complex from the surface of live cells by affinity receptor chromatography**

Jankowski A, Zhu P, Marshall JG.\*

Department of Chemistry and Biology, Faculty of Engineering, Architecture, and Science, Ryerson University, Toronto, Ont., Canada.

Contact: 4marshal@ryerson.ca

**Background:** Cell surface receptors and their associated signaling pathways on the plasma membrane are key targets in understanding cellular responses. However, the isolation and identification of receptor complexes has been elusive.

**Methods:** The Fc receptor was captured from the surface of live cells using microbeads coated with the receptor's cognate ligand, gamma globulin (IgG), and analyzed by liquid chromatography and tandem mass spectrometry (LC-MS/MS) alongside several controls. The resulting lists of proteins and protein sequences were compared between IgG beads and control beads by Structured Query Language (SQL) and the The Basic Local Alignment Search Tool (BLAST).

**Results:** Live-cell affinity receptor chromatography (LARC) resulted in a partially nonredundant list of 288 proteins that were specific to the Fc receptor complex. The proteins identified were in close agreement with previously determined factors in the Fc receptor complex as demonstrated by genetic and biochemical methods and revealed novel complex members. Confocal microscopy was used to confirm recruitment of SRC, SYK, PLC, PKC, PI3K, SHIP, TEC, CDC42, RAP, PAK, GAP, GEF, GRP, and CRK to the receptor complex upon activation by the same ligand microbeads. The expression of mutants and silencing RNA against specific isoforms were used to demonstrate a functional role for novel members of the Fc receptor complex, including RHOG (RAS homologue member G), p115 RhoGEF (protein of 115-kDa RAS homologue guanine exchange factor), and CRKL (CRK-like). The recruitment of AKT pleckstrin homology (PH) domain green fluorescent protein (GFP) was used to quantify the production of phosphorylated inositol at the activated receptor complex.

**Conclusions:** We conclude that it is feasible to capture an activated receptor complex from the surface of live cells using ligand-coated microbeads for identification of members of a receptor complex or pathway by LC-ESI-MS/MS.

**B26: A Custom-Designed Peak Picking Algorithm for Mass Spectral Imaging Data**

Nico Verbeeck\*, Raf Van de Plas, Etienne Waelkens, Bart De Moor

ESAT-SCD, Katholieke Universiteit Leuven, Belgium

Contact: nico.verbeeck@esat.kuleuven.be

**Background:** A good peak picking strategy is one of the most crucial steps in the pre-processing of mass spectrometry data. However, until recently, peak picking on mass spectral imaging (MSI) data has mainly used algorithms that are also employed in standard mass spectrometry. While performing reasonably, these algorithms neglect a key advantage of mass spectral imaging: spatial information. Therefore, we present a new approach that improves on these results by using the available spatial information in an intuitive way, namely by cross-comparing the presence of peaks found in a target pixel with the information in the neighboring pixels.

**Methods:** Our method starts with the common observation that a target pixel will most often be at least partly surrounded by pixels with similar mass spectra. Our hypothesis is that by comparing our target pixel with these similar neighboring pixels, we can detect which peaks are real peaks, and which ones are due to noise. In order to compare the pixels, we transform our spectra to the wavelet space using the discrete wavelet transform (DWT), which has already been successfully applied for peak picking algorithms in standard mass spectrometry. In wavelet space, features of the target spectrum that are not supported by the neighboring spectra are discarded as noise, resulting in much cleaner spectra for subsequent peak picking.

**Results:** We illustrate the performance of our new peak picking method on a computer-generated artificial MSI dataset. The artificial dataset consists of 400 pixels, with mixtures of 3 mass spectra. Each spectrum has 60 very small to large peaks randomly distributed over the  $m/z$ -range. Gaussian noise was added to each spectrum. When comparing our method to standard peak picking, we see a strong decrease in the number of false negative peaks, from an average of 45,6 peaks per pixel to 3,6 peaks per pixel. This shows that our method can preserve peaks that would otherwise be lost in the noise. A major advantage of our method is that only localized measurements are used. At no point is there a need to load the entire data set into memory, thus memory overflow issues, an increasingly common problem in MSI data processing, are avoided.

**Conclusions:** We present a peak picking method that is custom designed for MSI data. Our method improves on standard peak picking methods by integrating spatial information.

**B27: A non-parametric method to assess the presence of significant DNA-methylation in enrichment-based NGS data**

Klaas Mensaert\*, Geert Trooskens, Wim Van Criekinge, Olivier Thas, Tim De Meyer

Dept. Mathematical Modelling, Statistics and Bioinformatics, Ghent University,  
Belgium

Contact: klaas.mensaert@ugent.be

**Background:** MBD-seq is a next-generation sequencing based method that allows to profile DNA-methylation in a genome-wide manner. First, methylated DNA-fragments are captured by Methyl-CpG Binding Domain (MBD)-based affinity purification. Second, the captured fragments are sequenced by NGS. Third, obtained reads are mapped on the reference genome and coverages are summarized for each Methylation Core (MC), i.e. independently methylated region in the genome. Previously, a Poisson model was used to test for significant methylation. The null-hypothesis assumes that all measurements in the MCs are noise. For an MC with adequately high coverage we reject it to follow this Poisson distribution, and the corresponding MC is called significantly methylated. However, this method is conservative, as its null-hypothesis assumes that no MC contains methylation, while most of the signal is from actual methylation. In addition, the Poisson model requires that maximum coverage per MC should be used instead of total coverage, to avoid the impact of MC length variance. This further decreases the methods power. Notably, analyses of other enrichment-based sequencing experiments (e.g. ChIP-seq) are typically also based on this Poisson model, and therefore suffer from the same weaknesses.

**Methods:** Here, another null-hypothesis is used: if no methylation is present, the coverage of the MCs will have the same distribution as comparable regions outside the MCs. These regions are named null-cores and are alike in features that give rise to bias: length, GC% and mappability. The genome was in silico fragmented and mapped to itself in order to identify the mappable regions. A sliding window algorithm was used to sample all non-overlapping null-cores for each MC, in the mappable regions outside the MCs. Total and maximal coverage were used to construct the null-distribution for each MC. These distributions were then used to select a threshold intensity for which we can reject the null-hypothesis with a 5% type I error rate.

**Results:** A workflow was developed to determine the significance of methylation for each MC. We compared this method to the Poisson model-based approach for a set of genes, revealing clear differences.

**Conclusions:** We developed a workflow to determine the significance of methylation for MBD-seq data. However, for some MCs, insufficient null-cores could be sampled to accurately determine the significance. Therefore a model will be constructed to predict the null-core distribution in function of length and GC%. This methodology may also be used for other enrichment-based NGS-methodologies such as ChIP-seq.

**B28: Finding the differences: whole genome sequencing and analysis of a mono-zygotic twin discordant for schizophrenia**

Peter De Rijk\*, Joke Reumers, Anthony Liekens, Maarten Van Den Bossche,  
Diether Lambrechts, Jorgen Del-Favero

VIB DMG / University of Antwerp

Contact: Peter.DeRijk@molgen.vib-ua.be

**Background:** As prices for bulk sequencing are plummeting, whole genome sequencing is becoming more and more available to researchers, and the potential applications in genomic medicine are enormous. The computational analysis however remains a challenge. The huge lists of resulting variants contain a substantial number of genotyping errors. Validation of genome sequencing results can easily become the most time-consuming and expensive part of the analysis. In this study, we sequenced and compared the genomes of a mono-zygotic twin of which one has schizophrenia and the other is healthy. We expected a very small number of true differences because of somatic mutations, which we tried to find among a much larger number of differences caused by sequencing errors.

**Methods:** We developed a set of tools for the comparison and analysis of multiple genomes. Variants can be extensively annotated with information influencing sequence quality, as well as functional information such as effect on known or predicted genes or the presence in known regulatory or conserved elements. A selection tool allows adaptive filtering based on all types of annotation. These tools were used to analyse the twin genomes.

**Results:** The number of differences found in the unfiltered set of variants was too large to validate, especially taking into account that nearly all of them are caused by sequencing errors. We optimized a set of filters by using the twin genomes as semi-replicates, assuming shared variants are “true variants” and discordant variants are “errors”, allowing us to assess the effect of different filters and quality cut-offs. The resulting cumulative set of filters reduced the per SNV error rate in the remaining variants 290-fold. After strict filtering and Sanger sequencing, 2 differences could be validated.

This strict cumulative filtering did of course remove a considerable fraction of the genome (32%) from consideration. For other types of studies, filter combinations and settings can be used, removing less from the genome, at the expense of a smaller reduction in error-rate.

**Conclusions:** The software tools developed in this study that allow the comparison and analysis of multiple genomes are available at [genomecomb.sf.net](http://genomecomb.sf.net). The software and the filtering approach allowed us to identify the first validated genetic differences in a mono-zygotic twin by whole-genome comparison. One variant is located in a gene desert. Some clues indicate that the location of the other in a LINE-1 may be relevant to the disease, warranting further future study.



## **B29: Evaluating the use of clustering trees for protein subfamily identification**

Eduardo de Paula Costa<sup>1</sup>, Celine Vens<sup>1</sup>, Hendrik Blockeel<sup>\*1,2</sup>

<sup>1</sup>Dept. of Computer Science, Katholieke Universiteit Leuven, Belgium. <sup>2</sup>Leiden  
Institute of Advanced Computer Science, Universiteit Leiden, The Netherlands.

Contact: Hendrik.Blockeel@cs.kuleuven.be

**Background:** The identification of subfamilies within a protein family is a challenging problem in bioinformatics. Being able to identify different subfamilies is important for protein function prediction, since proteins within a subfamily usually share a function that is not common to the entire family. Phylogenomic analysis has been used to tackle this task, as an alternative to the error-prone homology based methods. Phylogenomic methods first build a phylogenetic or hierarchical tree for a protein family, and then extract clusters from this tree that correspond to subfamilies.

**Methods:** Previously, we have proposed a top-down divisive clustering method for clustering protein sequences. It makes clusters that minimize the total branch length of the resulting clustering tree, and in that spirit is similar to Neighbor-Joining. Our method is based on a decision tree learner, and the tests that appear in the nodes check the occurrence of specific amino acids at certain positions. In this work we evaluate how well our trees can be cut into clusters that correspond to protein subfamilies. To extract the protein subfamilies from the resulting phylogenetic tree, we apply a post-pruning procedure that minimizes encoding cost, which is similar to SCI-PHY, a state-of-the-art method in phylogenomic analysis.

**Results:** We tested our method on eight datasets, corresponding to extensively curated protein families. We compared our results with those of SCI-PHY. Because the quality of the resulting clusters largely depends on the underlying tree structure that is being cut, we first evaluate the quality of the trees as a whole. For this purpose, we apply the Fitch algorithm to estimate the minimal number of “subfamily switches” at the internal branches of the tree. Comparing the Fitch scores for our trees to those of SCI-PHY, we observe 6 wins for our method and 2 ties.

When comparing the obtained clusters to the true subfamilies using four evaluation measures, our results are better than the SCI-PHY results for half of the cases. As these results depend on the level of granularity defined over the subfamilies, we also evaluate the results using category utility, which allows an objective comparison, and results in 6 out of 8 wins for our method.

**Conclusions:** The clusterings produced by our method are at least comparable to those produced by SCI-PHY. Additionally, our method produces better underlying trees, and identifies subfamily-specific positions. Finally, these positions can be used as tests for the classification of new protein sequences.

**B30: Analyzing NF- $\kappa$ B signaling networks in UVB treated Cutaneous T-Cell Lymphoma cell-lines using Gene expression profiling and Ingenuity Pathway Analysis**

Amit Kumar\*, Thomas Sauter, Silvia Racolta, Dagmar Kulms, Petr Nazarov, Laurent Vallar

Life Sciences Research Unit, University of Luxembourg, 162a, Avenue de la Faiencerie, L-1511, Luxembourg

Contact: amit.kumar@uni.lu

**Background:** In past few decades, NF- $\kappa$ B is considered as a key signaling player linked to mal-functionality which plays crucial role in inflammation and cancer related disease. Usually UV radiation gives rise to critical events by genetic alternation in the cells whereas it could also be helpful in controlling oncological activities in NF- $\kappa$ B activation and regulation for its target genes. We analysis differential gene expression responses in UVB treated and irradiated cells as well as to compare gene expression profiles for disease progression in two different stages of CTCL (SeAx and MyLa) to two normal cells T-cell lymphocytes (Jurkat) and Keratinocytes (HaCaT).

**Methods:** We studied the UVB radiation effects in Cutaneous T-Cell Lymphoma (CTCL) via four cell lines SeAx, MyLa, Jurkat and HaCaT by performing whole cell transcriptional analysis through micro-array experiments of 24 samples (four cell lines in two conditions and 3 replicates each) using Affymetrix Human Gene 1.0 ST Array chip. We mainly studied how NF- $\kappa$ B Signaling pathway and its upstream and downstream genes are affected by UVB radiation in different cell-lines. We used R Statistical programming for microarray data analysis and Ingenuity Pathway analysis for network analysis.

**Results:** We observed high significant differential gene expression Jurkat and MyLa (249 and 138 genes respectively) and low in SeAx and HaCaT (76 and 88 respectively). In HaCat, up-regulation is dominant to down-regulation whereas, In rest cell-lines down-regulation is dominant over up-regulation by effect of UVB radiations. In NF $\kappa$ B signaling genes MAP3K8 and PI3K3R3 are found down-regulated in MyLa and both controls NF- $\kappa$ B activation. Also, NF- $\kappa$ B target genes are majorly affected by UVB radiation and always up-regulated except in Jurkat where both up and down regulation of genes is observed. We also investigated other signaling pathways and functional activities associated to differentially expressed genes using Ingenuity core analysis

**Conclusions:** We observed cell-specific UVB radiation responses in all cell-lines. We also see contrastically differential gene expression patterns within four cell-lines which provide more biological insights in understanding CTCL and its progression from one stage to another stage. Our results shows that UVB radiation does not always produce same effects in different cells so possibly UV radiation could be at least effective in controlling NF- $\kappa$ B regulation in early stages (MyLa) rather than later stages (SeAx) of CTCL. Further studies are required to understand NF- $\kappa$ B regulation within healthy cells and different CTCL stage to identify exact mechanism of regulation UVB radiation.

**B31: Diabetes and Parkinson's - two old friends? Unraveling potential connections between diseases by text-mining techniques**

Maria Biryukov\*, Serge Eifes, Janos Binder, Venkata P. Satagopam, Reinhard Schneider

onScale Solutions, Germany

Contact: barmaliska@gmail.com

**Background:** It has been hypothesized that Parkinson's disease (PD) and Diabetes Mellitus (DM) share molecular mechanisms. However this relationship has not been confirmed yet by the biomedical community. The goal of this ongoing study is to investigate the relationship between PD and diabetes using text mining techniques and tools.

**Methods:** The main idea is the following: build the individual disease profiles, and check them for overlap. To achieve this, PD and DM topic signatures (TS) are constructed. TS are vectors of disease related <term -- weight> pairs, where weight indicates the strength of correlation. The assumption behind is that if two terms co-occur in one text unit more often than a chance predicts, they are semantically related. To estimate the term weight and set the threshold we apply the likelihood ratio. Roughly 100,000 full-text articles for DM and 50,000 articles for PD have been analyzed. We use Reflect to annotate the corpus with diseases and proteins. Thus, the resulting TS represent PD and DM in terms of highly correlated human proteins. Synonymy and ambiguity in protein annotation is tackled by introducing the weighting scheme that ensures contextualization of the ambiguously tagged proteins with respect to given disease.

**Results:** The resulting sets of TS are ranked lists of proteins that highly correlate to PD or DM, and represent the disease profile. To address the question of potential relatedness between the two diseases we compare the TS vectors and calculate the Pearson correlation coefficient PCC.  $PCC = 0.0124$  does not support the hypothesis of PD and DM correlation. This is not unexpected given the very different nature of these two diseases. Nevertheless, we take the comparison one step further and extract the pathways that involve top ranked genes from the TSs. A set of overlapping pathways, highly ranked in PD and DM has been detected. These pathways are related to cancer, neurotrophin and MAPK signaling, apoptosis as well as neuroactive ligand and receptor interactions.

**Conclusions:** Our approach is domain-independent and proves suitable for profiling of other diseases. It allowed us to highlight the overlapping pathways between PD and DM. To better delineate the mechanistic commonalities between DM and PD, we plan to differentiate between types I and II of DM. Statistical contextualization helped to alleviate the problem of ambiguity in protein annotation and can be used in Reflect to improve its performance.

### **B32: DTSpine: DTProbLog for pathway inference from cause-effect experiments**

Joris Renkens\*, Guy Van den Broeck, Siegfried Nijssen, Kathleen Marchal

K.U.Leuven

Contact: joris.renkens@cs.kuleuven.be

**Background:** Cells react to external stimuli through the activation of regulatory pathways. The SPINE framework, proposed by Ourfali et al. in 2007, aims at explaining gene expression experiments by identifying which proteins act as activators or repressors in pathways. This is done by using data on gene knockouts and molecular interactions. The identification problem is formulated in an integer linear program after which a commercial solver is used to solve the problem. The SPINE framework is probabilistic in nature as physical interactions are assigned a probability. The necessary calculations for computing probabilities are encoded in the integer linear program. Unfortunately, this linear program grows quickly when the amount of physical interactions becomes large, necessitating the use of approximate ILP solvers. Within the SPINE framework a heuristic was used for such a solver.

**Methods:** We formulate the same problem in DTProbLog, a general framework for solving probabilistic decision problems. The resulting framework is called DTSpine. Given the general nature of the DTProbLog framework, it provides a wide range of optimized algorithms for solving probabilistic decision problems, both exact and heuristic in nature. This framework allows us to develop a suitable representation of the decision problem and compare a wide range of solution strategies.

**Results:** SPINE is tested on both a small and a large network. The same tests are executed with several settings in DTSpine. The efficiency of the specialized algorithms allows DTSpine to calculate an exact solution for a small network. This results in an accuracy of 99% for predicting knockout effects. SPINE needs to use a heuristic and only obtains suboptimal results. In the case of the large network, DTSpine accurately approximates probabilities for higher maximal path lengths than SPINE. Both SPINE and DTSpine limit the path length so calculating the probability remains tractable.

**Conclusions:** Using DTProbLog for solving the problem of explaining gene expression experiments, has several advantages. First, the calculation of probabilities is much easier from a user's perspective. Furthermore, the algorithms provided are specialized for the task and thus very efficient. Lastly, DTProbLog provides several ways of approximating the solution, which allows accurate inference with large amounts of data. Both approximations for calculating the probabilities as well as finding a solution can be used.

**B33: Top-down, bottom-up and middle out perspectives to model mitochondrial dysfunction and ROS generation in relation to neurodegenerative diseases**

Alexey Kolodkin<sup>\*1</sup>, Hans V. Westerhoff<sup>2</sup>, Antonio del Sol Mesa<sup>1</sup> and Rudi Balling<sup>1</sup>

<sup>1</sup>Luxembourg Centre for Systems Biomedicine, University of Luxembourg, Luxembourg; <sup>2</sup>Molecular Cell Physiology, VU University Amsterdam, the Netherlands

Contact: alexey.kolodkin@uni.lu

**Background:** Systems biology aims to explain properties of living organisms in terms of interactions between macromolecules, thus, linking the layer of interacting biomolecules with the systemic functioning of the organism emerging from those interactions. It does so by reconstructing this emergence in a computer model. In the projection onto biomedicine, such models become important tools for the development of predictive, preventive, personalized, and participatory (P4) medicine.

**Methods:** There are three strategies to link the network of interacting biomolecules with the functioning of the organism as a whole: top-down, bottom-up and middle-out. Top-down modeling starts with the systemic behavior: first, one determines how the often complicated systemic function of interest varies with conditions, or with time. From the observations one induces hypothetical structures that can be responsible for this function - a so called a data-driven, “digital” approach. The bottom-up (mechanisms-based) strategy starts with describing the actual mechanism in terms of mathematical equations. Then one assigns experimentally determined values to model parameters and verifies the model by comparing its systemic behavior with the behavior of a real system: from known or assumed properties of the components one deduces system functions. A middle-out strategy starts modeling the behavior of a single organ or a single functional system in terms of interactions between entities of the lower, but not necessarily molecular levels of organization and then moves up to the whole organism level or down to smaller components such as biomolecules.

**Results and Conclusions:** The final destination of these three approaches is the same. When the underlying level of interacting elements in the middle-out approach reaches the level of physico-chemical interactions between biomolecules and when the systemic function is extended to the whole organism, the model should be equivalent to one obtained by use of the bottom up or top-down strategy. Analogously, the perfect top-down parameterization would make for a model with the same functionality as a model built using the bottom-up approach. In the unreachable limit it does not matter which approach is used; the final aim is a unique computer replica of the living organism for computing Life on the basis of the complete biochemical, genomic, transcriptomic, proteomic, metabolomic and cell-physiomic information. However, which approach could deliver most spin-offs early on, en route to that ideal model? The poster addresses this question on the example of perspectives to model mitochondrial dysfunction and ROS generation in the relation to neurodegenerative diseases.

**B34: The impact of copy number variation in mental retardation: exploratory analysis**

Peter Konings\*, Joris R. Vermeesch, Yves Moreau

ESAT, KU Leuven, Belgium

Contact: peter.konings@esat.kuleuven.be

**Background:** Copy Number Variation explains an important part of phenotypic variation. The Center for Human Genetics at Leuven University Hospital has been systematically collecting genotypes and phenotypes of patients with mental retardation for diagnostic and research purposes.

**Methods:** 1207 patients were genotyped on OGT 105 and 180K microarrays, phenotyped using the LDDb classification and stored in the Cartagenia Bench platform. We performed an exploratory analysis of this database and presented summary statistics graphically.

**Results and Conclusions:** Quality assessment revealed the high internal consistency and quality of the data. Exploratory data analysis can provide leads to novel discoveries. It is clear that large datasets such as this one will play an important role in further research on the impact of Copy Number Variation.

**B35: Clinical Data Miner - an Electronic Data Capture software framework that improves interrater agreement**

Arnaud Installé\*, Dirk Timmerman, Thierry Van den Bosch, Bart De Moor

ESAT - SCD, Katholieke Universiteit Leuven, Belgium

Contact: [arnaud.install@esat.kuleuven.be](mailto:arnaud.install@esat.kuleuven.be)

**Background:** Clinical trials are used to increase medical knowledge. To that end, recorded variables must be of sufficient quality. For variables that are collected based on the interpretation of imaging modalities, this can pose a problem. Indeed, the meaning of variables can be interpreted differently by different clinicians potentially from different educational backgrounds. We implemented an Electronic Data Capture (EDC) software framework that allows to add pictograms as \hints\ to Case Report Forms (CRFs), enabling a more consistent interpretation of the variables, leading to better interrater agreement. To complement this, building on this software framework we implemented a module that can be used to conduct interrater agreement studies based on a set of imaging modalities.

**Methods:** The software framework was developed in Java, using software development best practices such as Test-Driven Development and Continuous Integration. Currently, a web-based user interface is available. However, its architecture is highly modular, enabling different user interface implementations, such as desktop or smartphone interfaces.

**Results:** The framework's high modularity allows not only changing the user interface, but also integration in hospital IT environments, or to define other types of studies than clinical trials or interrater agreement studies. The framework has been used to conduct two interrater agreement studies, which showed that showing pictograms in a CRF indeed does affect interrater agreement. It is also currently used to conduct the multi-centric studies conducted by the International Endometrial Tumor Analysis group.

**Conclusions:** The possibility to show pictograms as “hints” in CRFs, as well as the ease with which interrater agreement studies can be conducted in this software framework provides clinicians with all necessary tools to improve data quality in the evaluation of imaging modalities. Its high modularity allows its repurposing for a wide range of contexts, be they clinical or academic.

**B36: Generate pseudo expression scores from ChIP-seq data**

Filip Pattyn\*, Frank Westermann, Frank Speleman, Jo Vandesompele

Center for Medical Genetics, Ghent University, Belgium

Contact: filip.pattyn@ugent.be

**Background:** The combination of Chromatin Immunoprecipitation with next-generation sequencing (ChIP-seq) has shifted the study of DNA-protein interactions from single locus to genome wide assessments. Although binding events are clearly visible after peak calling, it remains a challenge to annotate the peaks with genomic information. The simplest approach filters for genes with at least one peak within a predefined area. A lot of valuable information is omitted with this type of binary methods. Therefore we developed a continuous peak scoring method called Pseudo Expression Scoring (PSES) to include more peak details to allow the calculation of a single score for every type of “ChIPed” protein per gene.

**Methods:** The PSES algorithm uses peak genomic coordinates outputted by a peak calling program in combination with peak intensity values. The score equals the multiplication of the peak width and the peak intensity. For every peak, the distance from the nearest downstream gene is calculated. Depending on the type of DNA-protein interaction, a specific distance correction factor is used to exponentially lower the weight of more distant peaks. The biological function of the studied DNA-interacting protein determines how the distance calculation is done. On one hand, peak distances calculated for peaks from binding events of a typical transcription factor, are done by using the transcription start site as a reference point. An additional factor is used in the distance correction calculation to make a distinction between conservative transcription factors that are normally binding to the proximal gene promoter region and transcription factors binding to a wider range of the promoter region. Finally, the sum of all relevant scores from the peaks in the vicinity of a gene is calculated and used as a single reference value for a gene.

**Results and Conclusions:** This method allows generating ordered lists of target genes, which are more relevant and easier to handle in downstream processing like functional annotation, pathway analysis, and motif analysis. The combination of scores from ChIP-seqs of different DNA-binding proteins allows predicting gene expression values.



**B37: Introducing conformational variability into X-ray structures based on experimental NMR data**

Wim Vranken\*, Alex Volkov, Tom Lenaerts, Nico van Nuland

Department of Structural Biology, VIB and Structural Biology Brussels, Vrije  
Universiteit Brussel, Brussel, Belgium

Contact: wvranken@vub.ac.be

**Background:** For structural bioinformatics purposes such as protein-ligand docking, it is desirable that multiple structural models are available for the protein, and that the variability in these models reflects the solution dynamics of the protein. Obtaining such models is not always straightforward, especially for structures obtained only by x-ray crystallography, and typically involves long and computationally expensive molecular dynamics simulations.

**Methods:** Based on the premise that the experimentally derived NMR distance restraints encode the dynamics of proteins in solution (as well as uncertainties in the data itself), we apply information from a statistical analysis of NMR distance restraints to filter interatomic distances derived from x-ray structures. This filtered set of distances is then used to rapidly generate ensembles of structures following a procedure similar to an NMR structure calculation.

**Results:** The ensembles generated with this method for 16 PDB entries, each with variation of a set of parameters, are compared to available x-ray and NMR structures as well as molecular dynamics simulations. They relate well to the original information both in their conformational space and when comparing to experimental data from NMR. We discuss several in silico applications.

**Conclusions:** The method quickly introduces controlled and expected variation into a single x-ray structure, and can be useful in in silico docking and point mutation studies.

**B38: Transcriptor: a web-tool for mining whole-genome transcriptomes of prokaryotes**

Tilman Todt\*, Roland J. Siezen, Sacha A.F.T. van Hijum

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

Contact: [tilman.todt@han.nl](mailto:tilman.todt@han.nl)

**Background:** High-resolution microbial transcriptomes allow gaining new insights into genomic elements, such as operons, untranslated regions and transcription start sites. They additionally reveal potential regulatory roles of small RNAs and antisense transcripts. To date a few web-tools have been developed dedicated to the analysis of prokaryotic whole-genome transcriptomes. They allow analysis of tiling array data and/or RNA-Seq. Here, we describe Transcriptor, an easy to use web-tool for whole-genome transcriptome analysis designed for tiling array data and RNA-Seq of prokaryotes.

**Methods:** Transcriptor automatically detects transcriptionally active regions (TARs) in each sample provided and compiles one coherent transcript list over different samples (replicates, experiments, etc) specifying different types of transcripts, namely coding, anti-sense and non-coding.

**Results:** Transcriptor provides the user with a transcriptome landscape derived from the expression data and allows the user to e.g., identify conditionally expressed transcripts. It returns results in multiple standard output formats (e.g., GFF format, Wiggle format and tab-delimited format) which are suitable for visualization and interpretation of these complex datasets in third party software, such as statistical analysis in R and wiggle track visualization using the UCSC genome browser. Unique features of the tool include (i) for tiling array data: the estimation of thresholds from probe signal distributions, (ii) a robust method for locally improved transcript boundaries and (iii) the creation of a coherent transcript list over multiple samples / experiments that is suitable to perform comparative expression analysis in order to e.g., identify conditionally expressed sRNAs.

**Conclusions:** Transcriptor is a web-tool that allows straight-forward analysis of whole transcriptome data of prokaryotes. It enables researches to easily detect and compare different types of transcripts over different experiments.

**B39: Network analysis of differential expression for drug target prioritization**

Griet Laenen\*, Lieven Thorrez, Yves Moreau

ESAT, K.U.Leuven, Belgium

Contact: griet.laenen@esat.kuleuven.be

**Background:** The pharmaceutical industry is facing unprecedented productivity challenges. Attrition rates have risen sharply, especially in late-phase clinical trials. A large bottleneck in drug development is the identification of a compound's mode of action and its off-target effects. DNA microarray technology enables a genome-wide analysis of the transcriptional response to a compound treatment, and thus can provide valuable information on a compound's targets and its physiological effects prior to clinical trials. However, whole-genome expression profiles do not immediately distinguish the genes targeted by a compound from genes that respond indirectly to changes in the activity of the targets. To overcome this problem there is an urgent need for in silico methods considering these expression values at the level of a molecular network, rather than in isolation.

**Methods:** Since most drugs act at the protein level, the expression levels of the target gene(s) will often not be affected following treatment. Rather, genes functioning downstream of the target gene are those whose expression will be changed. Considering the problem from the perspective of a protein-protein interaction or functional association network, genes can therefore be prioritized as possible drug targets based on the level of differential expression in their neighborhood. To this end differential expression values are propagated over the network, based on the confidence scores of the interactions/associations, using a discrete approximation of the heat kernel diffusion ranking introduced by Chung and Yau. This iterative diffusion can be regarded as a 2-step random walk through the network.

**Results:** We illustrate our method by applying it to gene expression data following treatment with drugs exhibiting a well-defined mode of action. A first such drug is infliximab, a tumor necrosis factor (TNF)-binding monoclonal antibody marketed under the brand name Remicade. A genome-wide infliximab target prioritization based on our method, ranks the target TNF in the top 1%. Also the androgen receptor, primary target of bicalutamide and the glucocorticoid receptor NR3C1, primary target of methylprednisolone are ranked in the top 1% using the same approach.

**Conclusions:** In conclusion we can state that we have developed a powerful in silico method for genome-wide drug target prioritization, with at the core a protein-protein interaction or functional association network and a drug-specific expression data set. A comparison study of different protein networks and kernel methods, as well as the integration of structure-activity information are planned for the future to further improve this method.

**B40: Tree-based machine learning methods for Zebrafish Image Analysis**

Olivier Stern\*, Nathalie Jeanray, Raphaël Marée, Jessica Aceto, Benoît Pruvot, Pierre Geurts, Marc Muller, Louis Wehenkel

GIGA-Systems Biology & Chemical Biology, GIGA-R and Dept. EE CS, University of Liege, Belgium

Contact: olivier.stern@ulg.ac.be

**Background:** The zebrafish is a well-known model organism used for biological studies on development, gene function and toxicology. Traditionally, effects of biological experiments on zebrafish embryos are evaluated manually through microscopic observation. However, due to the large number of experimental protocols, chemical substances, acquisition modalities and the recent availability of high-throughput imaging equipments, visual inspection of zebrafish images by experts is a limiting factor in large-scale studies. These considerations led us to consider supervised machine learning methods to automate the extraction of useful, quantitative information from these images.

**Methods:** In this work, we consider two tasks: zebrafish embryo phenotype classification and morphometric measurements of the cartilage skeleton. We propose to tackle these two tasks by exploiting experts' annotations with extremely randomized tree methods combined with extraction of subwindows within images. More specifically, the first task aims at classifying zebrafish embryos exposed to various chemical substances at different concentrations according to the malformations observed (e.g. pericardial edemas, curved tails, growth delay). We use image processing methods to first standardize images, then recent variants of the generic method presented in Marée et al 2007 were evaluated systematically on batches of microscopy images. We obtain more than 90% recognition rates using cross-validation protocols. The second task aims at partitioning elements of the zebrafish larvae and performing various developmental measures of the cartilage such as skeleton surface, distances and angles between specific points. Here, we use subwindows and extremely randomized trees to build a multiple output pixel classifier model Dumont et al. 2009 to segment cartilage and we propose a multiple output regression model to jointly detect multiple points of interests within images, hence perform measurements.

**Results and Conclusions:** Our approach will be evaluated on large image sets of larvae to confirm encouraging, preliminary results obtained on tens of images.

**B41: Inferring gene association network from gene expression data using quantitative association rules**

Maria Martinez-Ballesteros, Isabel Nepomuceno-Chamorro\*, José C. Riquelme

Lenguajes y Sistemas Informaticos, Universidad de Sevilla, Spain

Contact: inepomuceno@us.es

**Background:** We use an unsupervised learning approach to deal with the problem of network reconstruction from gene expression profiles applying a data mining method based on quantitative association rules which define gene interrelations.

**Methods:** A multi-objective evolutionary algorithm for mining quantitative association rules has been proposed in this work. The approach is based on the well-known NSGA-II algorithm and has determined the intervals that form the rules without discretizing the gene expression values as a first step of the process.

**Results:** We applied our methodology to the yeast cell cycle (Spellman et al). The same training and validation experiments used by Soinov et al. were analyzed to achieve a comparison between our method and the decision-tree-based method presented by Soinov. The rules inferred by the decision-tree-based method (simultaneous rules) were also inferred by our approach, with the exception of seven rules. The biological relevance of the rules inferred by our approach was verified by analyzing whether such rules reflect functional properties relating to the different cell cycle phase G1, S, G2, M and M\G1. Two rules are consistent with the knowledge that the maximum of CLB2 transcription is in G2 phase, whereas CLN1, CLN2, CLB5 and CLB6 all have their expression maximum in G1. Two other rules obtained are in agreement with CLB2 and CLB1 being expressed simultaneously in G2. Three other rules are in agreement with the literature: transcription of SWI5 and CLB1 is G2/M specific and activated in late S phase; the expression pattern of SWI5 is similar to that of CLB1 and CLB2 and the peak of mRNA concentration of SWI5 is in G2. The rules where genes MBP1, CDC34 and SKP1 are consequent can be explained by the fact that their activities as parts of the MBF and SCF complexes are completely separated in time. Two other rules are consistent with the knowledge that CDC20 is transcribed in late S/G2 phase and its product is required for metaphase-to-anaphase transition, whereas CLN2 and CLN1 have their transcription maximum in G1.

**Conclusions:** The network is consistent with the knowledge store in the literature. Furthermore, the method can be improved by adding prior knowledge. Our method constitute an interactive expert system for gene association networks, where the expert decides when to stop adding new gene expression profiles and what biological meaning represent the network.

**B42: Inferring gene co-expression networks with Biclustering based on linear correlations among genes**

Juan A. Nepomuceno\*, Isabel Nepomuceno-Chamorro, Alicia Troncoso, Jesus Aguilar-Ruiz

Lenguaje y Sistemas Informaticos, Universidad de Sevilla, Espana

Contact: janepo@us.es

**Background:** The extraction of regulatory modules from gene expression profiles is one of the most important tasks in gene expression data analysis. One of the major drawbacks in this field is the huge amount of data. Gene network methods search regulatory modules. They are usually based on the idea that if two genes show similar expression profiles, they are supposed to follow the same regulatory regime. This idea assumes that co-expression genes means co-regulation. Biclustering is an Unsupervised Data Mining technique that searches for local patterns in the gene expression data matrix. Traditional Clustering is not adequate because genes are considered along all experimental conditions and interesting local relations could not be contemplated.

**Methods:** The proposed methodology to infer gene co-expression networks combines a Biclustering procedure with a network extraction procedure. The Biclustering algorithm reports a set of biclusters through the optimization of a fitness function. This algorithm is based on Scatter Search schemes. Scatter Search is a population-based metaheuristic where a set of individuals that represent trial solutions evolves in order to find optimal solutions. The fitness function optimized is based on linear correlations among genes. Local dependencies among genes are determined for each bicluster. Hence, the expression profile of every gene is only considered respect the conditions that determine the bicluster. These local patterns among genes are used to build a gene interaction graph using the correlation-based method.

**Results:** We have used as benchmark data set the well-known Yeast cell cycle CDC28 data, that studies yeast *Saccharomyces cerevisiae* along temporal conditions of two completes cycles of cell cycle. The microarray matrix is composed by 6131 genes and 17 conditions that represent samples each 10 minutes in the microarray experiment. We report the obtained results from the proposed methodology and the study considering the enrichment analysis based on Gene Ontology data base.

**Conclusions:** In this work in progress a methodology to infer genes co-expression networks has been presented. As previous step, a Biclustering algorithm has been used in order to obtain biclusters. Later, a network extraction procedure based on the linear correlations among genes of such biclusters has been provided. Experiments have been reported and a detailed analysis of one of the obtained networks will be shown.

**B43: Comparative analyses imply that the enigmatic sigma factor 54 is a central controller of the bacterial exterior**

Christof Francke\*, Tom Groot, Kormelink, Yanick Hagemeijer, Lex Overmars,  
Vincent Sluijter, Roy Moezelaar and Roland J. Siezen

TI Food and Nutrition, Wageningen; Kluyver Center for Genomics of Industrial  
Fermentation, Delft; Wageningen University RC; NBIC, Nijmegen; CMBI, RU  
Nijmegen MC, the Netherlands

Contact: c.francke@cmbi.ru.nl

**Background:** Sigma-54 is a central regulator in many pathogenic bacteria and has been linked to a multitude of cellular processes like nitrogen assimilation and important functional traits such as motility, virulence, and biofilm formation. Until now it has remained obscure whether these phenomena and the control by Sigma-54 share an underlying theme.

**Methods and Results:** We have uncovered the commonality by performing a range of comparative genome analyses. A) The presence of Sigma-54 and its associated activators was determined for all sequenced prokaryotes. We observed a phylum-dependent distribution that is suggestive of an evolutionary relationship between Sigma-54 and lipopolysaccharide and flagellar biosynthesis. B) All Sigma-54 activators were identified and annotated. The relation with phosphotransfer-mediated signaling (TCS and PTS) and the transport and assimilation of carboxylates and nitrogen containing metabolites was substantiated. C) The function annotations, that were represented within the genomic context of all genes encoding Sigma-54, its activators and its promoters, were analyzed for intra-phylum representation and inter-phylum conservation. Promoters were localized using a straightforward scoring strategy that was formulated to identify similar motifs. We found clear highly-represented and conserved genetic associations with genes that concern the transport and biosynthesis of the metabolic intermediates of exopolysaccharides, flagella, lipids, lipopolysaccharides, lipoproteins and peptidoglycan.

**Conclusion:** Our analyses directly implicate Sigma-54 as a central player in the control over the processes that involve the physical interaction of an organism with its environment like in the colonization of a host (virulence) or the formation of biofilm.

**B44: Mutalyzer 2: improved sequence variant descriptions from next generation sequencing data and gene variant databases**

J.F.J. Laros, M. Vermaat\*, G.R. Stouten, J.T. den Dunnen, P.E.M. Taschner

Center for Human and Clinical Genetics, Leiden University Medical Center, The Netherlands

Contact: m.vermaat.hg@lumc.nl

**Background:** Using next generation sequencing technology we can explore not a single gene, but the entire genome for sequence variants. Unambiguous and correct sequence variant descriptions are of utmost importance. Mistakes and uncertainties may lead to undesired errors in clinical diagnosis. On a larger scale, they will also interfere with attempts to link the variants to complete individual phenotypes, disease-related, healthy or even prognostic. To check and correct variant descriptions, we have developed Mutalyzer.

**Methods:** The Mutalyzer sequence variation nomenclature checker ([www.mutalyzer.nl/](http://www.mutalyzer.nl/)) names all sequence variants following the Human Genome Variation Society sequence variant nomenclature recommendations ([www.hgvs.org/mutnomen](http://www.hgvs.org/mutnomen)), using a GenBank or Locus Region Genomic (LRG) accession number, a HGCN gene symbol and the variant description as input. Maintenance of the Mutalyzer 2 nomenclature syntax parser has been simplified: using the Pyparsing parser generator in combination with the syntax of the standard nomenclature in extended Backus-Naur form, we can automatically generate a new parser after a nomenclature update.

**Results:** Mutalyzer generates an output containing a description of the sequence variant at DNA level, the effect on all annotated transcripts, its deduced outcome all annotated protein products and gains or losses of restriction enzyme recognition sites. Clicking on a specific transcript provides more detailed information about the structure of the transcript and its protein product. Mutalyzer facilitates batch-wise conversion from chromosomal position numbering used in next generation sequencing data or dbSNP rsIDs to transcript position numbering, as well as checking of sequence variants before import into these databases. 18 new SOAP web services also support the use of Mutalyzer's functionality from other computer programs. The new Name Generator can be used to train your self to generate correct HGVS descriptions.

**Conclusions:** Mutalyzer is used successfully to check new variant submissions in LOVD databases ([www.LOVD.nl/](http://www.LOVD.nl/)). In the setting of the EU-funded Gen2Phen project a gene variant database (LSDB) for all genes has been established accepting all gene variant data available, including those obtained from complete exome or genome resequencing. In addition, LOVD2 uses Mutalyzer 2 to map variants to genomic positions for visualization in the Ensembl and UCSC genome browsers and the NCBI Sequence Viewer. Mutalyzer will be used by LOVD3 to support the description of variants on multiple transcripts of a specific gene.



---

## Authors Index

---

### A

Aceto  
  Jessica · 105  
Adriaens  
  Michiel · 87  
Aelterman  
  Bart · 86  
Aerts  
  Jan · 12, 41  
  Stein · 12  
Aguilar-Ruiz  
  Jesus · 107  
Algaa  
  Enkhjargal · 25  
Antony  
  Paul · 71  
Aydın Son  
  Yesim · 27, 63  
Azuaje  
  Francisco J. · 2, 4, 5, 8, 35, 72

---

### B

Babu  
  M. Madan · 50  
Baggerman  
  Geert · 67, 77  
Balling  
  Rudi · 71, 98  
Baumuratov  
  Aidos · 71  
Behrmann  
  Iris · 83  
Benabderrahmane  
  Sidahmed · 66  
Bennett  
  Keiryn · 16  
Berthault  
  Pascale · 26  
Beyer  
  Marco · 55  
Billing  
  Anja · 49  
Binder  
  Janos · 34, 96  
Bingqiang

Wang · 35  
Bîrnigen  
  Daniela · 84  
Biryukov  
  Maria · 96  
Bittkowski  
  Meik · 25  
Blackburn  
  J · 64  
Blockeel  
  Hendrik · 94  
Blondel  
  A · 11  
Bonachela-Capdevila  
  Francisco · 84  
Boopathy  
  Rathanam · 74, 78  
Bootsma  
  Hester · 22  
Bottu  
  Guy · 26  
Bowden  
  P · 48, 56, 57  
Boyd  
  Olga · 71  
Breitwieser  
  Florian · 23, 32  
Bresso  
  Emmanuel · 66  
Burghout  
  Peter · 22

---

### C

Calle  
  Luz · 17  
Can  
  Tolga · 79  
Cardinale  
  Francesca · 55  
Çarkacıoğlu  
  Levent · 27  
Castagnetti  
  Fausto · 46  
Cattaert  
  Tom · 17  
Charloteaux

---

Benoat · 17  
Chateaufvieux  
Sébastien · 49  
Cheng  
J · 70  
Cilia  
Elisa · 53  
Cleynen  
Isabelle · 17  
Colinge  
Jacques · 16, 23, 32  
Corte-Real  
Joana P. · 51  
Couceiro  
José · 44  
Crappé  
Jereon · 77  
Cuypers  
Thomas · 43

---

## *D*

Daelemans  
Walter · 30  
Dang  
Thanh Hai · 13  
de Bruin  
Amy · 82  
De Causmaecker  
Patrick · 84  
de Groot  
Philip · 87  
De Knijf  
Jeroen · 30  
de Ligt  
Joep · 33  
De Meyer  
Tim · 24, 47, 68, 73, 75, 92  
De Moor  
B · 60, 91, 100  
Bart · 69, 84, 85  
De Nil  
Simon · 68  
De Paepe  
Ayla · 68  
de Paula Costa  
Eduardo · 94  
De Rijk  
Peter · 30, 86, 93

De Schrijver  
Joachim · 47  
De Smet  
Riet · 38, 59  
de Vos  
Willem M. · 65  
de Vries  
Stefan · 22  
del Sol Mesa  
Antonio · 98  
Del-Favero  
Jurgen · 30, 86, 93  
den Dunnen  
J.T. · 109  
Denil  
Simon · 24, 73, 75  
Deschrijver  
Joachim · 77  
Desmedt  
Christine · 88  
Deus  
Helena · 28  
Devaux  
Yvan · 72  
Devignes  
Marie-Dominique · 14, 66  
Devriendt  
Koenraad · 84  
Dicato  
Mario · 49  
Diederich  
Marc · 49  
Dingli  
David · 46  
Duarte  
Isabel · 31  
Dutta  
Anwesha · 87

---

## *E*

Ederveen  
Thomas H.A. · 82  
Eifes  
Serge · 96  
Eijssen  
Lars · 87  
El Aalamat  
Yousef · 69

---

Emreiahin  
  Mehmet · 79  
Engelen  
  Kristof · 38, 45, 59  
Evelo  
  Chris · 28, 87

---

*F*

Faust  
  Karoline · 39  
Fierro  
  Carolina · 20, 59  
Florentinus  
  AK · 48  
Fostier  
  Jan · 20  
Francke  
  Christof · 7, 8, 10, 65, 81, 108  
Frankild  
  Sune · 34  
Fu  
  Qiang · 38, 59  
Fuxreiter  
  Monika · 50

---

*G*

Gaigneaux  
  Anthoula · 49  
Galhardo  
  Mafalda · 61  
Gedikoğlu  
  Ceyhun · 27  
Geurts  
  Pierre · 18, 105  
Gheysen  
  Godelieve · 75  
Ghoneim  
  Christelle · 54  
Ghoorah  
  Anisah · 14  
Gibson  
  Toby · 37  
Gilissen  
  Christian · 33  
Glaab  
  Enrico · 58  
Goethals

Bart · 15, 30  
Golebiewski  
  Martin · 25  
Goncalves  
  C · 11  
Groot  
  Tom · 6, 7, 8, 9, 10, 65, 81, 108  
Gsponer  
  Jörg · 50  
Gusareva  
  Elena · 17

---

*H*

Haegeman  
  Annelies · 75  
Hagemeijer  
  Yanick · 65  
  Yanik · 7, 10, 108  
Hamed  
  M. · 76  
Hayakawa  
  Eisuke · 67  
Hehir-Kwa  
  Jayne · 33  
Heinäniemi  
  Merja · 61  
Helms  
  Volkhard · 76  
Hermans  
  Peter · 22  
Hermjakob  
  Henning · 37  
Hoegen  
  Brechtja · 82  
Horvatovich  
  Peter · 19  
Huttenhower  
  Curtis · 39  
Huynen  
  Martijn · 31, 43, 80  
Huynh-Thu  
  Vén Anh · 18

---

*I*

Iacucci  
  Ernesto · 60  
Ignatiadis

---

Michail · 88  
Installé  
Arnaud · 100  
Ismael  
Siba · 76

---

*J*

Jaillard  
Magali · 87  
Jankowski  
A · 48, 90  
Jeanray  
Nathalie · 105  
Jensen  
Lars Juhl · 34  
Jong  
Lenneke · 25

---

*K*

Kalender  
Zeynep · 12  
Kania  
Renate · 25  
Kaoma  
Tony · 51, 54  
Kerckhof  
Frederiek-Maarten · 47  
Kerrien  
Samuel · 37  
Kerssemakers  
Jules · 36  
Kistler  
Harold Corby · 55  
Koeglsberger  
Sandra · 71  
Koenders  
Eric · 65  
Koks  
Patrick · 26  
Kolodkin  
Alexey · 98  
Konings  
P · 70, 99  
Kormelink  
Tom Groot · 65, 81, 108  
Kraehenbuhl  
Jean-Pierre · 26

Kreis  
Stephanie · 83  
Kruse  
Kai · 50  
Kulms  
Dagmar · 95  
Kumar  
Amit · 95  
Kutmon  
Martina · 87  
Kvasz  
Alex · 17  
Kyndt  
Tina · 75

---

*L*

Labarre  
Anthony · 62  
Ladant  
D · 11  
Laenen  
Griet · 104  
Lai  
Yuching · 12  
Laine  
E · 11  
Lakshmi  
Venkatachalam · 74, 78  
Lambrechts  
Diether · 93  
Lang  
Benjamin · 50  
Langereis  
Jeroen · 22  
Laros  
J.F.J. · 109  
Laukens  
Kris · 13, 15  
Leach  
Sonia · 85  
Lemmens  
Karen · 59  
Lenaerts  
Tom · 44, 46, 53, 102  
Leunissen  
Jack · 26  
Liekens  
Anthony · 30, 93

---

Lisacek  
Frédérique · 26  
Lucien  
Hoffmann · 55  
Luque  
Irene · 44

Marc · 105  
Müller  
Claude P. · 49  
Wolfgang · 25  
Muth  
T · 64

---

*M*

Magno  
Ramiro · 31  
Mahachie John  
Jestinah · 17  
Malliavin  
TE · 11  
Mampaey  
Evi · 24  
Marchal  
Kathleen · 20, 38, 45, 59, 97  
Marchetti  
Gino · 66  
Marée  
Raphaal · 105  
Marshall  
JG · 48, 56, 57, 90  
Martens  
L · 64  
Martinez  
L · 11  
Martinez-Ballesteros  
Maria · 106  
McDonnel  
M. · 57  
Mensaert  
Klaas · 24, 92  
Menschaert  
Gerben · 67, 77  
Meysman  
Pieter · 45, 59  
Michiels  
Stefan · 88  
Moezelaar  
Roy · 10, 108  
Montis  
Valeria · 55  
Moreau  
Yves · 12, 60, 70, 84, 85, 99, 104  
Muller  
Arnaud · 51, 54

---

*N*

Nabuurs  
Sander · 31  
Napoli  
Amedeo · 66  
Naulaerts  
Stefan · 13  
Nazarov  
Petr · 51, 54, 83, 95  
Nepomuceno  
Juan A. · 106, 107  
Nepomuceno-Chamorro  
Isabel · 106, 107  
Nicot  
Nathalie · 83  
Nijssen  
Siegfried · 97  
Nijtmans  
Leo · 43

---

*O*

O'Donoghue  
Sean · 34  
Obbels  
Dagmar · 47  
Overmars  
Lex · 65, 81, 108

---

*P*

Pacheco  
Jorge · 46  
Pasquali  
Matias · 55  
Pattyn  
Filip · 101  
Pavlopoulos  
Georgios · 12, 41  
Petrenko  
V · 48

---

Philippidou  
Demetra · 83  
Pipelers  
Peter · 24  
Plaisance  
Stephane · 40  
Polsky  
Martina · 76  
Popovic  
D · 60, 69  
Prudence  
Cynthia · 29  
Pruvot  
Benoat · 105

---

*R*

Racolta  
Silvia · 95  
Raes  
Jeroen · 39  
Rault  
S · 11  
Reinsbach  
Susanne · 83  
Renaut  
Jenny · 49  
Renckens  
Bernadet · 82  
Renkens  
Joris · 97  
Reshetnyak  
Yana · 29  
Reumers  
Joke · 93  
Rey  
Maja · 25  
Riquelme  
José C. · 106  
Ritchie  
David · 14  
Rix  
Uwe · 16  
Rojas  
Isabel · 25  
Rossier  
Grégoire · 26  
Rosti  
Gianantonio · 46

Rousseau  
Frederic · 44  
Ruano  
Ana Zafra · 44

---

*S*

Saeyns  
Yvan · 18  
Sánchez de Groot  
Natalia · 50  
Sanchez-Rodriguez  
Aminael · 38, 59  
Sanz  
Javier Ruiz · 44  
Satagopam  
Venkata P · 96  
Sathirapongsasuti  
J. Fah · 39  
Sauter  
Thomas · 52, 61, 95  
Schmitt  
Martina · 83  
Schneider  
Reinhard · 21, 34, 58, 96  
Schymkowitz  
Joost · 44  
Secrier  
Maria · 21  
Selwa  
E · 11  
Shengchang  
Gu · 35  
Shi  
Lei · 25  
Siezen  
Roland J · 65, 81, 82, 103, 108  
Sifrim  
Alejandro · 12, 41  
Sinan Karaboga  
Arnaud · 66  
Singhal Sandeep  
K · 88  
Sluijter  
Vincent · 10, 108  
Smail-Tabbone  
Malika · 14, 66  
SON  
Cagdas D. · 79

---

Sotiriou  
Christos · 88

Souchet  
Michel · 66

Speleman  
Frank · 101

Stern  
Olivier · 105

Steyaert  
Sandra · 68

Stouten  
G.R. · 109

Stukalov  
Alexey · 23, 32

Stunnenberg  
Hendrik · 22

Superti-Furga  
Giulio · 16

Sylvestre  
Jean · 26

Szklarczyk  
Radek · 43

---

## *T*

Taschner  
P.E.M. · 109

Teiten  
Marie-Hélène · 49

Thas  
Olivier · 92

Théâtre  
Emilie · 17

Thiele  
H · 57

Thorrez  
Lieven · 104

Timmerman  
Dirk · 100

Todt  
Tilman · 103

Trairatphisan  
Panuwat · 52

Tranchevent  
LC · 60, 84, 85

Traulsen  
Arne · 46

Trefois  
Christophe · 71

Troncoso  
Alicia · 107

Trooskens  
Geert · 24, 68, 73, 75, 77, 92

---

## *U*

Urrea  
Victor · 17

---

## *V*

Valkenburg  
Dirk · 15

Vallar  
Laurent · 51, 54, 83, 95

Van Criekinge  
Wim · 24, 47, 67, 68, 73, 75, 92

van Dam  
John · 80

Van de Plas  
Raf · 91

Van den Bosch  
Thierry · 100

Van Den Bossche  
Maarten · 93

Van den Broeck  
Guy · 97

Van den Bulcke  
Tim · 89

van der Lee  
Robin · 50

van der Velde  
Arjan · 42

van Helden  
Jacques · 26

van Hijum  
Sacha · 22, 82, 103

Van Lishout  
Francois · 17

van Nuland  
Nico · 102

Van Roey  
Kim · 37

Van Steen  
Kristel · 17

Vancriekinge  
Wim · 77

Vandermarliere

---

E · 64  
Vandesompele  
  Jo · 101  
Vanneste  
  E · 70  
Veltman  
  Joris · 33  
Vens  
  Celine · 94  
Verbeeck  
  Nico · 91  
Verbeke  
  Lieven · 20  
Verleyen  
  Elie · 47  
Vermaat  
  M. · 109  
Vermeesch  
  Joris · 12, 70, 99  
Verschoren  
  Alain · 13  
Verwer  
  Sicco · 62  
Veugelers  
  Mark · 40  
Visentin  
  Ivan · 55  
Vissers  
  Lisenka · 33  
Volders  
  Pieter-Jan · 47  
Volkov  
  Alex · 102  
Vranken  
  Wim · 102  
Vu  
  Trung Nghia · 15  
Vuister  
  Geerten W. · 53  
Vyverman  
  Wim · 47

---

## W

Waagmeester  
  Andra · 28  
Waelkens  
  Etienne · 69, 91  
Wagner  
  Daniel R. · 72  
Wanschers  
  Bas · 43  
Wehenkel  
  Louis · 17, 18, 105  
Weidemann  
  Andreas · 25  
Westerhoff  
  Hans V. · 98  
Westermann  
  Frank · 101  
Wetsch  
  Elina · 25  
Wittig  
  Ulrike · 25  
Wu  
  Yan · 20, 38

---

## Y

Yu  
  Shi · 85  
Yuan  
  Liu · 35

---

## Z

Zakeri  
  Pooya · 85  
Zhang  
  Lu · 35, 72  
Zhu  
  P · 57, 90  
Zomer  
  Aldert · 22



---

Our Sponsors and Supporters



**BGL  
BNP PARIBAS**



**JOURNAL OF  
CLINICAL BIOINFORMATICS**



UNIVERSITÉ DU  
LUXEMBOURG

LSRU

LIFE SCIENCES  
RESEARCH  
UNIT



BIOLOGICAL DATABASES



**Partek®**  
turning data into **discovery®**



**MSD**

**LISA** Luxembourg  
Life Sciences  
Association

**nbic**

netherlands  
bioinformatics  
centre