



19th

Annual International Conference
on Research in Computational
Molecular Biology

Book of Abstracts



Warsaw, Poland
April 12–15, 2015

R E C  M B
2  1 5

Program at a glance

April 11-15, 2015

Saturday, April 11

18:00-21:00 Welcome reception

Sunday, April 12

08:45-9:00 Opening remarks
 09:00-10:00 **KEYNOTE:** Michael Levitt
 10:00-10:40 **Session 1**
 10:40-11:10 Coffee break
 11:10-12:30 **Session 2 (structure, part 1)**
 12:30-13:30 Lunch
 13:30-14:30 **KEYNOTE:** Bonnie Berger
 14:30-15:10 **Session 3 (alignment free methods)**
 15:10-15:40 Coffee break
 15:40-17:00 **Session 4 (sequence analysis)**
 17:30-19:00 Poster session sponsored by University of Pittsburgh
 20:30-22:00 Social dinner at the new Library of the University of Warsaw

Monday, April 13

09:00-10:00 **KEYNOTE:** M. Madan Babu
 10:00-10:40 **Session 5 (chromatin structure)**
 10:40-11:10 Coffee break
 11:10-12:30 **Session 6 (population)**
 12:30-13:30 Lunch
 13:30-14:10 **Session 7**
 14:10-15:30 **KEYNOTE:** Waław Szybalski
 15:30-19:00 Free time / Sightseeing

Tuesday, April 14

09:00-10:00 **KEYNOTE:** Bas van Steensel
 10:00-10:40 **Session 8 (gene regulation)**
 10:40-11:10 Coffee break
 11:10-12:30 **Session 9 (cancer)**
 12:30-13:30 Lunch
 13:30-14:50 **Session 10 (structure, part 2)**
 14:50-15:20 Coffee break
 15:20-16:40 **Session 11 (networks)**
 16:50-17:20 Business Meeting
 17:30-19:00 Poster session sponsored by National Science Foundation

Wednesday, April 15

09:00-10:00 **KEYNOTE:** Magda Konarska
 10:00-10:40 **Session 12 (splicing)**
 10:40-11:10 Coffee break
 11:10-12:30 **Session 13 (evolutionary trees)**
 12:30-13:30 Lunch
 13:30-14:50 **Session 14 (association analysis)**
 14:30-15:00 Award ceremony and closing remarks

Preface

2015 is already the nineteenth edition of the RECOMB conference, so it has become natural to have high expectations on all aspects of the meeting. Even though this is the first time RECOMB is held in Poland, we hope we will not disappoint our participants this year. As the organizers, we have been lucky to work with Teresa Przytycka as the Program Committee Chair, who has been working with us very tightly on the program as well as organizational matters. Mona Singh and Lenore Cowen, who were responsible for the highlight and poster submissions, were also helpful in timely completion of program of the conference. Putting this conference together would not be possible without the members of the organizing committee as well as the helpful staff of the GlobalWings company who helped us with the logistics and their experience. The same goes for the students from our volunteer team helping during the conference itself. Last, but not least, this conference would not be possible without our sponsors, who have been very generous in supporting RECOMB through direct donations and travel fellowships for 94 students and post-docs. We are looking forward to the conference itself and thank you all for your participation.

Jerzy Tiuryn and Bartek Wilczyński
Conference Chairs

Steering committee

Vineet Bafna
Serafim Batzoglou
Bonnie Berger
Sorin Istrail
Michal Linial

Martin Vingron (Chair)

University of California, San Diego, USA
Stanford University, USA
Massachusetts Institute of Technology, USA
Brown University, USA
The Hebrew University of Jerusalem, Israel
Max Planck Institute for Molecular Genetics, Germany

Previous RECOMB meetings

January 20-23, 1997	Santa Fe, USA
March 22-25, 1998	New York, USA
April 22-25, 1999	Lyon, France
April 8-11, 2000	Tokyo, Japan
April 22-25, 2001	Montréal, Canada
April 18-21, 2002	Washington, USA
April 10-13, 2003	Berlin, Germany
March 27-31, 2004	San Diego, USA
May 14-18, 2005	Boston, USA
April 2-5, 2006	Venice, Italy
April 21-25, 2007	San Francisco, USA
March 30-April 2, 2008	Singapore
May 18-21, 2009	Tucson, USA
August 12-15, 2010	Lisbon, Portugal
March 28-31, 2011	Vancouver, Canada
April 21-24, 2012	Barcelona, Spain
April 7-10, 2013	Beijing, China
April 2-5, 2014	Pittsburgh, USA
April 12-15, 2015	Warsaw, Poland

Organizing committee

Paweł Bednarz	University of Warsaw, Poland
Agata Charzyńska	Polish Academy of Sciences, Poland
Norbert Dojer	University of Warsaw, Poland
Anna Gambin	University of Warsaw, Poland
Ilona Grabowicz	University of Warsaw, Poland
Paweł Górecki	University of Warsaw, Poland
Julia Herman-Iżycka	University of Warsaw, Poland
Aleksander Jankowski	University of Warsaw, Poland
Agnieszka Mykowiecka	University of Warsaw, Poland
Ewa Szczurek	ETH Zurich, Switzerland
Jerzy Tiuryn (Co-chair)	University of Warsaw, Poland
Irina Tuszyńska	University of Warsaw, Poland
Bartek Wilczyński (Co-chair)	University of Warsaw, Poland
Damian Wójtowicz	National Institutes of Health, USA
Rafał Zaborowski	University of Warsaw, Poland

Student volunteers

Konrad Bednarek	Anna Kilian	Karolina Sienkiewicz
Katarzyna Bednarska	Marcin Kostecki	Grzegorz Skoraczyński
Karolina Dawid	Hanna Kranas	Maryna Wiśniewska
Michał Gąsior	Michał Krassowski	Katarzyna Życieńska
Agnieszka Hajduk	Karolina Kuczkowska	
Paweł Jankowski	Jagoda Kuśnierz	
Katarzyna Kędzierska	Mateusz Legięcki	

Program committee

Tatsuya Akutsu	Kyoto University, Japan
Peter Arndt	Max Planck Institute for Molecular Genetics, Germany
Rolf Backofen	University of Freiburg, Germany
Joel Bader	Johns Hopkins University, USA
Vineet Bafna	University of California, San Diego, USA
Nuno Bandeira	University of California, San Diego, USA
Ziv Bar-Joseph	Carnegie Mellon University, USA
Serafim Batzoglou	Stanford University, USA
Jan Baumbach	University of Southern Denmark, Denmark
Niko Beerenwinkel	ETH Zurich, Switzerland
Panayiotis Takis Benos	University of Pittsburgh, USA
Bonnie Berger	Massachusetts Institute of Technology, USA
Jadwiga Bienkowska	Biogen Idec, USA
Mathieu Blanchette	McGill University, Canada
Jacek Błażewicz Poznań	University of Technology, Poland
Michael R. Brent	Washington University, St. Louis, USA
Chakra Chennubhotla	University of Pittsburgh, USA
Lenore J. Cowen (Posters Chair)	Tufts University, USA
Colin Dewey	University of Wisconsin-Madison, USA
Dannie Durand	Carnegie Mellon University, USA
Nadia El-Mabrouk	Université de Montréal, Canada
Julien Gagneur	Ludwig-Maximilians-Universität München, Germany
Anna Gambin	University of Warsaw, Poland
Irit Gat-Viks	Tel Aviv University, Israel
Mikhail Gelfand	Russian Academy of Sciences, Russia
David Gifford	Massachusetts Institute of Technology, USA
Anna Goldenberg	University of Toronto, Canada
Eran Halperin	International Computer Science Institute, USA
Alexander Hartemink	Duke University, USA
Michael Hoffman	University of Toronto, Canada
Trey Ideker	University of California, San Diego, USA
Sorin Istrail	Brown University, USA
Tao Jiang	University of California, Riverside, USA
Igor Jurisica	Ontario Cancer Institute, Canada
Tamer Kahveci	University of Florida, USA
Simon Kasif	Boston University, USA
Carl Kingsford	Carnegie Mellon University, USA
Mehmet Koyuturk	Case Western Reserve University, USA
Rui Kuang	University of Minnesota Twin Cities, USA
Jens Lagergren	KTH Royal Institute of Technology, Sweden
Thomas Lengauer	Max Planck Institute for Informatics, Germany
Christina Leslie	Memorial Sloan Kettering Cancer Center, USA
Michal Linial	The Hebrew University of Jerusalem, Israel
Stefano Lonardi	University of California, Riverside, USA
Paul Medvedev	Pennsylvania State University, USA
Tijana Milenkovic	University of Notre Dame, USA
Satoru Miyano	University of Tokyo, Japan
Bernard Moret	École Polytechnique Fédérale de Lausanne, Switzerland

Chad Myers	University of Minnesota, Twin Cities, USA
William Stafford Noble	University of Washington, USA
Laxmi Parida	IBM T.J. Watson Research Center, USA
Dana Pe'er	Columbia University, USA
Jian Peng	University of Illinois at Urbana-Champaign, USA
Mihai Pop	University of Maryland, USA
Teresa Przytycka (Chair)	National Institutes of Health, USA
Ben Raphael	Brown University, USA
Knut Reinert	Freie Universität Berlin, Germany
Maga Rowicka	University of Texas Medical Branch, USA
Marie-France Sagot	Inria, France
Cenk Sahinalp	Indiana University, USA
David Sankoff	University of Ottawa, Canada
Russell Schwartz	Carnegie Mellon University, USA
Roded Sharan	Tel Aviv University, Israel
Mona Singh (Highlights Chair)	Princeton University, USA
Donna Slonim	Tufts University, USA
Fengzhu Sun	University of Southern California, USA
Glenn Tesler	University of California, San Diego, USA
Jerzy Tiuryn	University of Warsaw, Poland
Alfonso Valencia	Spanish National Cancer Research Centre, Spain
Fabio Vandin	University of Southern Denmark, Denmark
Martin Vingron	Max Planck Institute for Molecular Genetics, Germany
Jerome Waldispuhl	McGill University, Canada
Esti Yeger-Lotem	Ben-Gurion University of the Negev, Israel
Michal Ziv-Ukelson	Ben-Gurion University of the Negev, Israel

External reviewers

Aguiar, Derek
Ahmed, Bilal
Ahmed, Hazem
Akey, Joshua
Albrecht, Mario
Aleksyev, Max
Alhakami, Hind
Almeida, Mathieu
Antczak, Maciej
Arvestad, Lars
Atias, Nir
Ay, Ferhat
Ayati, Marzieh
Azizi, Elham
Barbosa, Eudes
Batra, Richa
Becerra, David
Behr, Jonas
Bernstein, Laurence
Bhasin, Jeffrey
Bishara, Alex
Blumer, Anselm
Botzman, Maya
Bozdag, Serdar
Can, Tolga
Cao, Mengfei
Carmel, Amir
Carty, Mark
Chen, Brian
Chen, Quan
Chicco, Davide
Chikhi, Rayan
Cho, Dongyeon
Cicek, A. Ercument
Constantinescu, Simona
Daniels, Noah
Dao, Phuong
Darby, Charlotte
Davidson, Philip
Davydov, Iakov
Dimitrakopoulos, Christos
Dojer, Norbert
Donald, Bruce
Donmez, Nilgun
Dutkowski, Janusz
Edwards, Matthew
El-Kebir, Mohammed
Elhesha, Rasha
Filippova, Darya
Flamm, Christoph
Frånberg, Mattias
Frellsen, Jes
Frishberg, Amit
Gabr, Haitham
Gartzman, Dalya
Gendoo, Deena
Gitter, Anthony
Golumbeanu, Monica
Gonzalez, Alvaro
Górecki, Paweł
Guo, Yuchun
Guthals, Adrian
Haas, Brian
Hach, Faraz
Haiminen, Niina
Hajirasouliha, Iman
Halldorsson, Bjarni
Hamel, Sylvie
Harel, Tom
Hasan, Abid
Hasan, Mahmudul
Haspel, Nurit
Hauschild, Anne-Christin
Hauswedell, Hannes
He, Dan
Heinig, Matthias
Heringa, Jaap
Hescott, Benjamin
Hill, Christopher
Hodzic, Ermin
Hoener Zu Siederdisen, Christian
Hoffmann, Steve
Hofree, Matan
Hoinka, Jan
Holtby, Daniel
Hormozdiari, Fereydoun
Hosur, Raghavendra
Huang, Lin
Huska, Matt
Irannia, Zohreh Baharvand
Jain, Siddhartha
Jensen, James
Kalinina, Olga
Kayano, Mitsunori

Keasar, Chen	Numanagic, Ibrahim
Keich, Uri	Oesper, Layla
Kierzynka, Michal	Ofer, Dan
Koch, Elizabeth	Ounit, Rachid
Kohlbacher, Oliver	Pan, Weihua
Kolodny, Rachel	Park, Heewon
Kramer, Michael	Paulson, Joseph
Krishnaswamy, Smita	Persikov, Anton
Kuipers, Jack	Pervouchine, Dmitri
Lafond, Manuel	Pfeifer, Nico
Lai, Han	Pfeuffer, Julianus
Lajoie, Mathieu	Pham, Son
Lang, Brian	Pisanti, Nadia
Lange, Sita	Plat, Daniel
Laukens, Kris	Platts, Adrian
Li, Wenyuan	Polishko, Anton
Liang, Xiao	Pons, Carles
Liao, Chung-Shou	Popic, Victoria
Libbrecht, Max	Prabhakaran, Sandhya
Lin, Yen Yi	Pritykin, Yuri
Lindner, Martin S.	Przytycki, Pawel
Linz, Simone	Radom, Marcin
List, Markus	Rarey, Matthias
Liu, Yuling	Reimand, Juri
Love, Michael	Reinharz, Vladimir
Lu, Yong	Ritz, Anna
Lu, Yuheng	Roe, David
Lukasiak, Piotr	Rubinov, Anatoly
Luo, Qiang	Rudolph, Jan
Łacki, Mateusz Krzysztof	Ruffalo, Matthew
Ma, Wenxiu	Sacomoto, Gustavo
Madej, Tom	Savel, Daniel
Maier, Ezekiel	Schaefer, Rob
Malikic, Salem	Sefer, Emre
Mazza, Arnon	Sennblad, Bengt
Mepheron, Andrew	Setty, Manu
Mezlini, Aziz	Shao, Mingfu
Mi, Huaiyu	Sheinman, Misha
Mirebrahim, Seyed Hamid	Shen, Yang
Mironov, Andrey	Sheridan, Paul
Montazeri, Hesam	Silverbush, Dana
Mosig, Axel	Sindi, Suzanne
Mueller, Jonas	Singh, Irtisha
Muthiah, Senthil	Skola, Dylan
Mysickova, Alena	Smaoui, Mohamed
Na, Seungjin	Solomon, Brad
Nachshon, Aharon	Startek, Michał
Nadimpalli, Shilpa	Steerman, Yael
Navlakha, Saket	Syed, Tahin
Nelson, Justin	Todor, Andrei
Niida, Atsushi	Tofigh, Ali

Tomescu, Alexandru I.
Tremblay-Savard, Olivier
van Iersel, Leo
Vandersluis, Benjamin
Veksler-Lublinsky, Isana
Vert, Jean-Philippe
Viner, Coby
Vinga, Susana
Walevski, Uri
Wan, Lin
Wang, Bo
Wang, Jian
Wang, Jie
Wang, Mingxun
Wang, Sheng
Wang, Wen
Wise, Aaron
Wiwie, Christian
Wojciechowski, Pawel
Wolfson, Haim
Wu, Xuebing
Yang, Li
Yavas, Gokhan
Yoo-Ah, Kim
You, Xintian
Yu, Mike
Yuan, Han
Zakov, Shay
Zamalloa, Jose
Zhang, Huanan
Zhang, Wei
Zhang, Yao-Zhong
Zheng, Jie
Zhong, Shan
Zirbel, Craig

Table of Contents

Program	10
Keynote speakers	15
Highlights	23
Accepted papers	27
Posters	33
Author index	113

Program

Saturday, April 11

18:00-21:00 Welcome reception

Sunday, April 12

08:45-9:00 Opening remarks

09:00-10:00 **KEYNOTE:** Michael Levitt
Birth and future of multiscale modeling of macromolecules
Introduction by: Michael Waterman

10:00-10:40 **Session 1.**
Chair: Rolf Backofen

10:00-10:20 Ilan Ben-Bassat, Benny Chor
CRISPR detection from very short reads using partial overlap graphs

10:20-10:40 Rasmus Fonseca, Henry van den Bedem, Julie Bernauer
KGSrna: Efficient 3D kinematics-based sampling for nucleic acids

10:40-11:10 Coffee break

11:10-12:30 **Session 2 (structure, part 1).**
Chair: Lenore Cowen

11:10-11:30 Mark Hallen, Bruce Donald
COMETS (Constrained Optimization of Multistate Energies by Tree Search): A provable and efficient algorithm to optimize binding affinity and specificity with respect to sequence

11:30-11:50 Naama Amir, Dan Cohen, Haim J. Wolfson
DockStar: A novel ILP based integrative method for structural modelling of multimolecular protein complexes

11:50-12:10 Jianzhu Ma, Sheng Wang, Zhiyong Wang, Jinbo Xu
Protein Contact Prediction by Integrating Joint Evolutionary Coupling Analysis and Supervised Learning

12:10-12:30 Huichao Gong, Sai Zhang, Jiangdian Wang, Haipeng Gong, Jianyang Zeng
Constructing Structure Ensembles of Intrinsically Disordered Proteins from Chemical Shift Data

12:30-13:30 Lunch

13:30-14:30 **KEYNOTE:** Bonnie Berger
Computational biology in the 21st century: Algorithms that scale
Introduction by: Mona Singh

14:30-15:10 **Session 3 (alignment free methods).**
Chair: Knut Reinert

14:30-14:50 Srinivas Aluru, Alberto Apostolico, Sharma V. Thankachan
Efficient Alignment Free Sequence Comparison with Bounded Mismatches

14:50-15:10 **HIGHLIGHT:** Rob Patro, Stephen Mount, Carl Kingsford
Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms

-
- 15:10-15:40 Coffee break
- 15:40-17:00 **Session 4 (sequence analysis).**
Chair: Ron Shamir
- 15:40-16:00 Nam-Phuong Nguyen, Siavash Mirarab, Keerthana Kumar, Tandy Warnow
Ultra-large alignments using ensembles of Hidden Markov Models
- 16:00-16:20 Leena Salmela, Kristoffer Sahlin, Veli Mäkinen, Alexandru Tomescu
Gap Filling as Exact Path Length Problem
- 16:20-16:40 Alex Bishara, Yuling Liu, Dorna Kashef-Haghighi, Ziming Weng, Daniel Newburger, Robert West, Arend Sidow, Serafim Batzoglou
Read Clouds Uncover Variation in Complex Regions of the Human Genome
- 16:40-17:00 Igor Mandric, Alex Zelikovsky
ScaffMatch: Scaffolding Algorithm Based on Maximum Weight Matching
- 17:30-19:00 Poster session sponsored by University of Pittsburgh
- 20:30-22:00 Social dinner at the new Library of the University of Warsaw

Monday, April 13

- 09:00-10:00 **KEYNOTE:** M. Madan Babu
The contribution of intrinsically disordered regions to protein function, cellular complexity and human diseases
Introduction by: Michal Linial
- 10:00-10:40 **Session 5 (chromatin structure).**
Chair: Bartek Wilczyński
- 10:00-10:20 Emre Sefer, Geet Duggal, Carl Kingsford
Deconvolution Of Ensemble Chromatin Interaction Data Reveals The Latent Mixing Structures In Cell Subpopulations
- 10:20-10:40 **HIGHLIGHT:** Alon Diamant, Ron Y. Pinter, Tamir Tuller
Three Dimensional Eukaryotic Genomic Organization is Strongly Correlated with Codon Usage Expression and Function
- 10:40-11:10 Coffee break
- 11:10-12:30 **Session 6 (population).**
Chair: Jens Lagergren
- 11:10-11:30 Emily Berger, Deniz Yorukoglu, Bonnie Berger
HapTree-X: An integrative Bayesian framework for haplotype reconstruction from transcriptome and genome sequencing data
- 11:30-11:50 Shou Young, Shai Carmi, Itsik Pe'er
Rapidly Registering Identity-by-Descent Across Ancestral Recombination Graphs
- 11:50-12:10 Laxmi Parida, Filippo Utro, Deniz Yorukoglu, Anna Paola Carrieri, David Kuhn, Saugata Basu
Topological Signatures for Population Admixture
- 12:10-12:30 Jong Wha Joo, Eun Yong Kang, Elin Org, Nick Furlotte, Brian Parks, Aldons Lysis, Eleazar Eskin
Efficient and accurate multiple-phenotype regression method for high dimensional data considering population structure
- 12:30-13:30 Lunch

- 13:30-14:10 **Session 7.**
Chair: Roded Sharan
- 13:30-13:50 Roy Ronen, Glenn Tesler, Ali Akbari, Shay Zakov, Noah A. Rosenberg, Vineet Bafna
Haplotype Allele Frequency (HAF) Score: Predicting Carriers of Ongoing Selective Sweeps Without Knowledge of the Adaptive Allele
- 13:50-14:10 Maja Temerinac-Ott, Armaghan W. Naik, Robert F. Murphy
Deciding when to stop: Efficient experimentation to learn to predict drug-target interactions
- 14:10-15:30 **KEYNOTE:** Waław Szybalski
DNA - the essence of life (+film)
Introduction by: Teresa Przytycka
- 15:30-19:00 Free time / Sightseeing

Tuesday, April 14

- 09:00-10:00 **KEYNOTE:** Bas van Steensel
Mapping genome – nuclear lamina interactions in single cells
Introduction by: Martin Vingron
- 10:00-10:40 **Session 8 (gene regulation).**
Chair: Fabio Vandin
- 10:00-10:20 Yifeng Li, Chih-Yu Chen, Wyeth Wasserman
Deep Feature Selection: Theory and Application to Identify Enhancers and Promoters
- 10:20-10:40 **HIGHLIGHT:** James Zou, Christoph Lippert, David Heckerman, Martin Aryee, Jennifer Listgarten
Epigenome-wide association studies without the need for cell-type composition
- 10:40-11:10 Coffee break
- 11:10-12:30 **Session 9 (cancer).**
Chair: Cenk Sahinalp
- 11:10-11:30 Andrew McPherson, Andrew Roth, Cedric Chauve, S. Cenk Sahinalp
Joint inference of genome structure and content in heterogeneous tumor samples
- 11:30-11:50 Mark Leiserson, Hsin-Ta Wu, Fabio Vandin, Benjamin Raphael
CoMEt: A Statistical Approach to Identify Combinations of Mutually Exclusive Alterations in Cancer
- 11:50-12:10 Fabio Vandin, Ben Raphael, Eli Upfal
On the Sample Complexity of Cancer Pathways Identification
- 12:10-12:30 **HIGHLIGHT:** Mark Leiserson, Fabio Vandin, Hsin-Ta Wu, Jason Dobson, Jonathan Eldridge, Jacob Thomas, Alexandra Papoutsaki, Younhun Kim, Beifang Niu, Michael McLellan, Michael Lawrence, Abel Gonzalez-Perez, David Tamborero, Yuwei Cheng, Gregory Ryslik, Nuria Lopez-Bigas, Gad Getz, Li Ding, Benjamin Raphael
Pan-Cancer Network Analysis Identifies Combinations of Rare Somatic Mutations across Pathways and Protein Complexes
- 12:30-13:30 Lunch

- 13:30-14:50 **Session 10 (structure, part 2).**
Chair: Sorin Istrail
- 13:30-13:50 Stefano Bonissone, Pavel Pevzner
Immunoglobulin classification using the colored antibody graph
- 13:50-14:10 Jonathan D Jou, Swati Jain, Ivelin Georgiev, Bruce R Donald
BWM: A Novel, Provable, Ensemble-based Dynamic Programming Algorithm for Sparse Approximations of Computational Protein Design*
- 14:10-14:30 Yichao Zhou, Yuexin Wu, Jianyang Zeng
Computational Protein Design Using AND/OR Branch-and-Bound Search
- 14:30-14:50 Sumudu Leelananda, Robert Jernigan, Andrzej Kloczkowski
Exploration of Designability of Proteins Using Graph Features of Contact Maps: Beyond Lattice Models
- 14:50-15:20 Coffee break
- 15:20-16:40 **Session 11 (networks).**
Chair: Ben Raphael
- 15:20-15:40 Arnon Mazza, Allon Wagner, Eytan Ruppin, Roded Sharan
Functional alignment of metabolic networks
- 15:40-16:00 Hyunghoon Cho, Bonnie Berger, Jian Peng
Diffusion Component Analysis: Unraveling Functional Topology in Biological Networks
- 16:00-16:20 Surojit Biswas, Meredith McDonald, Derek Lundberg, Jeffery Dangl, Vladimir Jojic
Learning microbial interaction networks from metagenomic count data
- 16:20-16:40 **HIGHLIGHT:** Bo Wang, Aziz Mezlini, Feyyaz Demir, Marc Fiume, Zhuowen Tu, Michael Brudno, Benjamin Haibe-Kains, Anna Goldenberg
Similarity Network Fusion for aggregating data types on a genomic scale
- 16:50-17:20 Business Meeting
- 17:30-19:00 Poster session sponsored by National Science Foundation

Wednesday, April 15

- 09:00-10:00 **KEYNOTE:** Magda Konarska
Analysis of spliceosome function: from the mechanisms of catalysis to interconnections within global networks of gene expression
Introduction by: Jerzy Tiuryn
- 10:00-10:40 **Session 12 (splicing).**
Chair: Vineet Bafna
- 10:00-10:20 Stefan Canzar, Sandro Andreotti, David Weese, Knut Reinert, Gunnar W. Klau
CIDANE: Comprehensive isoform discovery and abundance estimation
- 10:20-10:40 **HIGHLIGHT:** Hui Yuan Xiong, Babak Alipanahi, Leo Lee, Hannes Bretschneider, Daniele Merico, Ryan Yuen, Yimin Hua, Serge Gueroussov, Hamed Najafabadi, Tim Hughes, Quaid Morris, Yoseph Barash, Adrian Krainer, Nebojsa Jojic, Stephen Scherer, Benjamin Blencowe, Brendan Frey
The human splicing code reveals new insights into the genetic determinants of disease

- 10:40-11:10 Coffee break
- 11:10-12:30 **Session 13 (evolutionary trees).**
Chair: Russell Schwartz
- 11:10-11:30 Kai Dührkop, Sebastian Böcker
Fragmentation trees reloaded
- 11:30-11:50 Mingfu Shao, Bernard Moret
A Fast and Exact Algorithm for the Exemplar Breakpoint Distance
- 11:50-12:10 Philippe Gambette, Andreas Gunawan, Anthony Labarre, Stephane Vialette, Louxin Zhang
Locating a Tree in A Phylogenetic Network in Quadratic Time
- 12:10-12:30 **HIGHLIGHT:** Siavash Mirarab, Md. Shamsuzzoha Bayzid, Bastien Boussau, Tandy Warnow
Statistical binning enables an accurate coalescent-based estimation of the avian tree
- 12:30-13:30 Lunch
- 13:30-14:50 **Session 14 (association analysis).**
Chair: Anna Gambin
- 13:30-13:50 Roni Wilentzik, Irit Gat-Viks
A novel probabilistic methodology for eQTL analysis of signaling networks
- 13:50-14:10 Seunghak Lee, Aurelie Lozano, Prabhanjan Kambadur, Eric Xing
An Efficient Nonlinear Regression Approach for Genome-wide Detection of Marginal and Interacting Genetic Variations
- 14:10-14:30 David Manescu, Uri Keich
A symmetric length-aware enrichment test
- 14:30-15:00 Award ceremony and closing remarks

Michael Levitt

Robert W. and Vivian K. Cahill Endowed Professor
in Cancer Research, Department of Structural Biology,
Stanford School of Medicine, Stanford, CA 94305, USA



Birth and future of multiscale modeling of macromolecule

The development of multiscale models for complex chemical systems began in 1967 with publications by Warshel and Levitt recently recognized by the 2013 Nobel Committee for Chemistry. The simplifications used then at the dawn of the age of computational structural biology were mandated by computers that were almost a billion times less cost-effective than those we use today. These same multiscale models have become increasingly popular in application that range from simulation of atomic protein motion, to protein folding and explanation of enzyme catalysis. In this talk I describe the origins of computational structural biology and then go on to show some of the most exciting current and future applications.



Bonnie Berger

Massachusetts Institute of Technology,
Broad Institute of MIT and Harvard
and Harvard Medical School, USA

Computational biology in the 21st century: Algorithms that scale

The last two decades have seen an exponential increase in genomic and biomedical data, which will soon outstrip advances in computing power. Extracting new science from these massive datasets will require not only faster computers; it will require algorithms that scale sub-linearly in the size of the datasets. We show how a novel class of algorithms that scale with the entropy of the dataset by exploiting its fractal dimension can be used to address large-scale challenges in genomic search, sequencing and small molecule search.

M. Madan Babu

MRC laboratory of Molecular Biology,
Francis Crick Avenue, Cambridge CB2 0QH, UK



The contribution of intrinsically disordered regions to protein function, cellular complexity and human diseases

In the 1960s, Christian Anfinsen postulated that the unique three-dimensional structure of a protein is determined by its amino acid sequence. This work laid the foundation for the sequence-structure-function paradigm, i.e. the sequence of a protein determines its structure, and structure determines function. However, a class of polypeptide segments called Intrinsically Disordered Regions (IDRs) defies this postulate. In this lecture, I will first describe established and emerging ideas about how disordered regions contribute to protein function. I will then discuss molecular principles of how regulatory mechanisms such as alternative splicing and asymmetric mRNA localization of transcripts encoding disordered segments can increase the functional versatility of proteins. Finally, I will discuss how disordered regions contribute to human disease and the emergence of cellular complexity during organismal evolution.



Wacław Szybalski

University of Wisconsin-Madison, USA

DNA - the essence of life

Similarly as Stanislaw Ulam, I was born and educated as Ch.E. in the multinational, but very Polish and dearly loved by us city of Lwów (Leopolis, Lemberg, Lvov and Lviv). Therefore, I will start my Ulam's lecture sketching his life as the Leopolitan mathematician and creator of the Monte Carlo method and hydrogen bomb, through our still lasting tragedy of the pre-WWII population of Lwów, as was imposed by terror, murders and finally by practically total deportations, as carried by Soviets, Germans and their henchmen, with shameless approval by signers of the Teheran, Yalta and Potsdam treaties.

Then, I will say how my exposure to Lwów scientific life has lead to our contributions to the physical genetics of DNA, including the basis of life, followed to creating new fields of gene therapy and synthetic biology. Finally, I will introduce the documentary film, "The essence of life", in which the director, Anna Ferens, takes her artist's look at my life, whereas I will try to add the scientist's perspective.

Bas van Steensel

Division of Gene Regulation, Netherlands Cancer Institute,
Amsterdam, the Netherlands



Mapping genome – nuclear lamina interactions in single cells

Mammalian interphase chromosomes interact with the nuclear lamina (NL) through hundreds of large Lamina Associated Domains (LADs). These interactions are thought to contribute to spatial genome organization. In addition, most genes in LADs are repressed, suggesting a role for NL contacts in gene regulation. The dynamics of LAD-NL interactions are poorly understood.

We previously developed a 'molecular contact memory' approach to track LADs in living cells. In each nucleus, only a subset of all LADs is positioned at the periphery; these LADs are in intermittent molecular contact with the NL but remain constrained to the periphery. Upon mitosis, LAD positioning is - at least in part - stochastically reshuffled.

We now report a modified version of our DamID method that can be used to map LAD-NL interactions genome-wide in single human cells. Analysis of more than 100 of these maps reveals a core architecture of gene-poor LADs that contact the NL with high cell-to-cell consistency, interspersed by LADs with more variable NL interactions. The single-cell maps indicate that NL contacts involve multivalent interactions over hundreds of kilobases. Strikingly, we observe extensive intra-chromosomal coordination of NL contacts, even over tens of megabases. Results of additional analyses of these datasets will be discussed.



Magda Konarska

The Rockefeller University, New York, NY 10065

Analysis of spliceosome function: from the mechanisms of catalysis to interconnections within global networks of gene expression

Pre-mRNA splicing, a process during which introns are removed from eukaryotic pre-mRNAs, can generate multiple forms of mRNA products. Three sites in pre-mRNA substrates, the 5' and 3' splice sites and the branch site, participate directly in two consecutive transesterification reactions of splicing, catalyzed by a single enzyme, the spliceosome. Recognition of these sites and their proper positioning within the spliceosome determine the specificity and efficiency of the reaction; alterations in splice site selection lead to alternative splicing. Using genetic analysis in yeast, *Saccharomyces cerevisiae*, we analyze the mechanism of this process. The spliceosome conformations supportive of the first and the second step catalysis exist in competition; numerous mutant alleles in spliceosomal components can improve one of the splicing steps by stabilizing the corresponding spliceosomal conformation, but inhibit the other step. If the active sites for the two splicing reactions are related, then substrates should be positioned similarly for the two steps. Indeed, I will present evidence for the branch site and the 3' splice site binding to the same site at the catalytic center during the first and second step, respectively. This model of interactions at the spliceosomal catalytic center has important implications for substrate selection for splicing.

Genetic analysis can also help to study other nuclear reactions that affect splicing. I will describe an example of a genetic screen to identify mutants that improve splicing of introns with defective 5' splice site. Among the isolated alleles are not only mutant forms of spliceosome components, but also mutants of factors implicated in a wide range of mRNA biogenesis reactions – from transcription, through processing, to mRNA transport – pointing to a delicate network of interactions between different nuclear processes.

A film about Prof. Waław Szybalski

THE ESSENCE OF LIFE

Director: Anna Ferens

Today, the Professor is 93 years old. He skis, swims in the lake, goes to concerts, dances at balls, keeps up both with international affairs and events in Poland; reads a lot, uses Skype and drives a car...

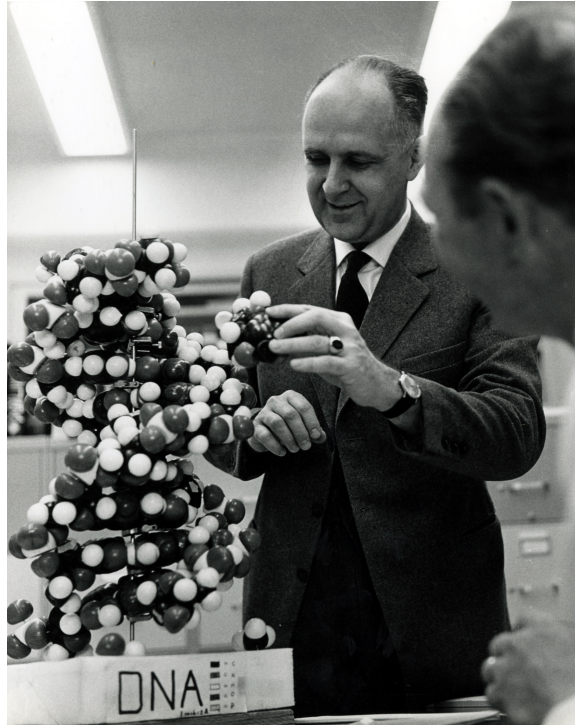


Figure 1: Waław Szybalski 1970

Regarded as "the Father of Synthetic Biology", Prof. Szybalski conducted pioneer research in this field and it was actually him who coined the name for this field of science. Three times he was shortlisted for the Nobel Prize, and a number of scientists who followed his scientific path were actually awarded this prize. Although he has been working for so many years in the United States, he has always emphasized that he comes from Poland and that it is in Poland where he received his education. It is thanks to his activity that so many Polish scientists were granted research fellowships in the United States.

Prof. Szybalski was born in Lwow (now Lviv in Ukraine), where he also launched his scientific career. In Lwow he worked at the laboratory of Prof. Rudolf Weigl, who invented the first effective vaccine against epidemic typhus. There, Szybalski met the anthropologist Jan Czekanowski, the mathematician Stefan Banach, the geographer Eugeniusz Romer, the poet Zbigniew Herbert, the composer Stanislaw Skrowaczewski and many others.

During the war, he was a member of the Polish underground movement. Among others, he supplied medicines to the Warsaw Ghetto. After the war, he worked at Gdansk University of Technology, organized a research unit and its activities. In 1946, he started cooperating with a research center in Copenhagen, where he made research on yeast genetics under the guidance of Prof. Winge at the Carlsberg Laboratory. Moreover, Prof. Szybalski cooperated with the Nobel laureate in physics, Nils Bohr.

Later, he emigrated to the USA, where he joined the famous Cold Spring Harbor Laboratory (1951-1955). There, he met another Nobel laureate, James Watson. Their friendship lasts until today. He continued his research at the Institute of Microbiology of the Rutgers University in New Brunswick,



Figure 2: From left: prof. Zdenka Hradecna, prof. Waclaw Szybalski, Anna Ferens (film director)

NJ (at the invitation of the eminent Nobel laureate, Prof. Selman Waksman). Between 1960 and 2003 Prof. Szybalski worked as a professor at the University of Wisconsin-Madison, where he works until today as a professor emeritus, pursuing his research at a laboratory, which is at his disposal for life.

His scientific achievements and contribution to world science are good enough reasons to make a film about our fellow compatriot. The whole life of Prof. Szybalski is a perfect starting point for a story about the wonder of life.

2015 Anna Ferens ©

Highlights

H177

Epigenome-wide association studies without the need for cell-type composition

James Zou¹, Christoph Lippert¹, David Heckerman¹, Martin Aryee², Jennifer Listgarten¹

¹Microsoft Research, United States, ²Harvard Medical School, United States

J. Zou <jzou@fas.harvard.edu>

Epigenome-wide association studies (EWAS) seek to identify loci whose epigenetic changes are correlated with phenotype. Such analyses complement GWAS and can potentially yield key insights into disease etiology. EWAS face many of the same challenges as GWAS in identifying the needles of true signal in the genomic haystack. Importantly, EWAS faces an added challenge in that the epigenome can vary dramatically across different cell types. When cell-type composition differs between cases and controls, this leads to spurious associations that bury true associations. While the current approach to tackling this problem is to estimate the cell-type composition in each sample using laboriously collected “reference profiles” (measurements from purified cells), we here propose to bypass the need for these reference samples altogether. We propose a method, FaST-LMM-EWASher, that automatically corrects for cell-type heterogeneity, without cell-type compositions knowledge. We validate our method on a bronze standard methylation data set for which we have cell-type composition, and demonstrate that it performs as well as the state-of-art method, which explicitly uses cell-type information. After additional validation through simulations, we apply FaST-LMM-EWASher to breast and colon cancer methylation data from the Cancer Genome Atlas to find disease-relevant associations not obtainable by standard analysis.

Paper: <http://www.nature.com/nmeth/journal/v11/n3/full/nmeth.2815.html>

Presenter: James Zou

H179

Statistical binning enables an accurate coalescent-based estimation of the avian tree

Siavash Mirarab¹, Md. Shamsuzoha Bayzid², Bastien Boussau³, Tandy Warnow⁴

¹University of Texas at Austin, United States, ²BUET, Bangladesh, ³Centre national de la recherche scientifique, France, ⁴the university of illinois at urbana-champaign, United States

S. Mirarab <smirarab@gmail.com>

T. Warnow <warnow@illinois.edu>

Gene tree incongruence arising from incomplete lineage sorting (ILS) complicates species trees estimation. Prevalent ILS, for example resulting from rapid radiations, can reduce the accuracy of the standard concatenation-based approach to species trees estimation. Coalescent-based species tree estimation methods have been developed to address this issue, and have been shown to have good accuracy in the presence of ILS under certain conditions. However, these methods have also been found to be sensitive to gene tree estimation error. We propose a pipeline, called statistical binning, to improve gene tree estimation accuracy. Our pipeline uses bootstrapping to evaluate whether two genes are likely to have the same tree topology, then it groups genes into sets using a graph-theoretic optimization and estimates a tree on each subset using concatenation, and finally produces an estimated species tree from these trees using the preferred coalescent-based method. Statistical binning improves the accuracy of MP-EST, a popular coalescent-based method, and we use it to produce the first genome-scale coalescent-based avian tree of life.

H180

The human splicing code reveals new insights into the genetic determinants of disease

Hui Yuan Xiong¹, Babak Alipanahi¹, Leo Lee¹, Hannes Bretschneider¹, Daniele Merico¹, Ryan Yuen¹, Yimin Hua², Serge Gueroussov¹, Hamed Najafabadi¹, Tim Hughes¹, Quaid Morris¹, Yoseph Barash³, Adrian Krainer², Nebojsa Jovic⁴, Stephen Scherer⁵, Benjamin Blencowe¹, Brendan Frey¹

¹University of Toronto, Canada, ²Cold Spring Harbor Laboratory, United States, ³University of Pennsylvania, United States, ⁴Microsoft Research, United States, ⁵Toronto Hospital for Sick Kids, Canada

B. Frey <frey@psi.toronto.edu>

To facilitate precision medicine and whole-genome annotation, we developed a machine learning technique that scores how strongly genetic variants affect RNA splicing, whose alteration contributes to many diseases. Analysis of more than 650,000 intronic and exonic variants revealed widespread patterns of mutation-driven aberrant splicing. Intronic disease mutations alter splicing 9 times as often as common variants, and missense exonic disease mutations that have the least impact on protein function are 5 times as likely as others to alter splicing. We detected tens of thousands of disease-causing mutations, including those involved in cancers and spinal muscular atrophy. Examination of intronic and exonic variants found using whole-genome sequencing of individuals with autism revealed misspliced genes with neurodevelopmental phenotypes. Our approach provides evidence for causal variants and should enable new discoveries in precision medicine.

Published in Science Express, December 18, 2014, DOI: 10.1126/science.1254806.

Published in print, Science, January 9, 2015.

H191

Three Dimensional Eukaryotic Genomic Organization is Strongly Correlated with Codon Usage Expression and Function

Alon Diament¹, Ron Y. Pinter², Tamir Tuller¹

¹Biomedical Engineering Dept., Tel Aviv University, Israel, ²Computer Science Dept., Technion - Israel Institute of Technology, Israel

A. Diament <dalon@post.tau.ac.il>

T. Tuller <tamirtul@post.tau.ac.il>

Link to paper: <http://www.nature.com/ncomms/2014/141216/ncomms6876/full/ncomms6876.html>

Formerly reported relations between gene function and genomic organization in eukaryotes were relatively weak. Previous studies have demonstrated that codon usage bias is related to all stages of gene expression and to protein function. Here we apply a novel tool for assessing functional relatedness, codon usage frequency similarity (CUFS), which measures similarity between genes in terms of codon and amino acid usage, and may be applied to all of the genes of any sequenced genome. By analyzing chromosome conformation capture data, describing the three-dimensional (3D) conformation of the DNA, we show that the functional similarity between genes captured by CUFS is directly and very strongly correlated with their 3D distance in *S.cerevisiae*, *S.pombe*, *A.thaliana*, *M.musculus* and human. The study provides the first global analysis at single-gene resolution of the spatial organization in multiple eukaryotes. The results emphasize that the importance of 3D genomic localization in eukaryotes is significantly higher than previously thought. The proposed methods can be employed to develop accurate models of genomic evolution and organization; they can facilitate a better understanding of gene function, gene expression and their evolution. We will also describe novel algorithms, which are based on CUFS, for improved inference of 3D genome conformation models.

H212

Pan-Cancer Network Analysis Identifies Combinations of Rare Somatic Mutations across Pathways and Protein Complexes

Mark Leiserson¹, Fabio Vandin², Hsin-Ta Wu¹, Jason Dobson¹, Jonathan Eldridge¹, Jacob Thomas¹, Alexandra Papoutsaki¹, Younhun Kim¹, Beifang Niu³, Michael McLellan³, Michael Lawrence⁴, Abel Gonzalez-Perez⁵, David Tamborero⁵, Yuwei Cheng⁶, Gregory Ryslik⁶, Nuria Lopez-Bigas⁵, Gad Getz⁴, Li Ding³, Benjamin Raphael¹

¹Brown University, United States, ²University of Southern Denmark, Denmark, ³Washington University in St. Louis, United States, ⁴The Broad Institute of MIT and Harvard, United States, ⁵University Pompeu Fabra, Spain, ⁶Yale University, United States

M. Leiserson <mdml@cs.brown.edu>

F. Vandin <vandinf@cs.brown.edu>

B. Raphael <braphael@cs.brown.edu>

Cancers exhibit extensive mutational heterogeneity and the resulting long tail phenomenon complicates the discovery of the genes and pathways that are significantly mutated in cancer. Even recent The Cancer Genome Atlas Pan-Cancer studies have limited power to characterize genes in the long tail leaving an incomplete picture of the functional, somatic mutations in these samples. A prominent explanation for this mutational heterogeneity is the fact that genes act together in various signaling and regulatory pathways and protein complexes. We present HotNet2, a novel algorithm that uses a directed heat diffusion model to simultaneously assess both the significance of mutations in individual proteins and the local topology of interactions among proteins, overcoming limitations of pathway-based enrichment statistics and earlier network approaches. We perform a Pan-Cancer analysis of mutated networks in 3281 samples from 12 cancer types from TCGA using HotNet2. We identify 14 significantly mutated subnetworks that include well-known cancer signaling pathways as well as subnetworks with less characterized roles in cancer including cohesin, condensin, and others. These subnetworks contain dozens of genes with rare somatic mutations across multiple cancers, many with additional evidence supporting a role in cancer. By illuminating these rare combinations of mutations,

Pan-Cancer network analyses provide a roadmap to investigate new diagnostic and therapeutic opportunities across cancer types.

H214

Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms

Rob Patro¹, Stephen Mount², Carl Kingsford³

¹Stony Brook University, United States, ²University of Maryland, College Park, United States, ³Carnegie Mellon University, United States

R. Patro <rob.patro@gmail.com>

S. Mount <smmount@gmail.com>

C. Kingsford <carlk@cs.cmu.edu>

With the rapid growth in both the size and number of experimental samples, the quantification of gene and transcript expression from RNA-seq data has become a significant computational challenge. Traditional approaches for solving this problem rely on the alignment of sequencing reads to the genome or transcriptome — itself a time-consuming task — followed by an analysis of these alignments to resolve the ambiguity introduced by multi-mapping reads. We introduce a novel method, implemented in the software tool Sailfish, for the accurate and rapid quantification of transcript expression from RNA-seq data. By entirely avoiding the mapping of reads, a bottleneck in all previous methods, and adopting an accelerated inference procedure for our k-mer based representation of transcript abundance, Sailfish provides quantification estimates of similar accuracy much faster than existing approaches (typically 20 times faster). We will also present Salmon, a followup approach for accurate transcript quantification that we have recently developed. Salmon maintains the primary benefit of Sailfish — its exceptional speed — while producing even more accurate expression estimates in complex situations and eliminating the need to select an important parameter (the k-mer length) a priori. Sailfish and Salmon are open-source software, and the source code and binaries may be obtained from <http://www.cs.cmu.edu/~ckingsf/software/sailfish/>.

H215

Similarity Network Fusion for aggregating data types on a genomic scale

Bo Wang¹, Aziz Mezlini², Feyyaz Demir², Marc Fiume², Zhuowen Tu³, Michael Brudno², Benjamin Haibe-Kains⁴, Anna Goldenberg¹

¹SickKids Research Institute, Canada, ²University of Toronto, Canada, ³University of California San Diego, United States, ⁴Université de Montréal, Canada

A. Goldenberg <anna.goldenberg@utoronto.ca>

How can we combine multiple types of measurements for a set of patients to create a comprehensive view of a given disease? In this talk I will introduce a novel paradigm for data integration - patient networks. I will present our recently developed Similarity Network Fusion (SNF) to integrate genomic and other types of data for the same set of patients. Briefly, each data type (e.g. mRNA expression, DNA methylation, etc) for a given set of patients is represented as a network of patient similarities. These networks are then combined into one patient similarity network supported by all data. Combining three biological data types, SNF substantially outperformed single data type analysis and other integrative approaches to identify cancer subtypes in five cancers. I will also present an alternative test for differential gene analysis and a novel patient-network based regularization of the Cox survival model. Finally, if the time permits, I will show the new extension of SNF allowing to combine partially overlapping patient cohorts, which makes SNF applicable in situations where some of the data types were collected only for a subset of patients.

Accepted papers

T19

Deep Feature Selection: Theory and Application to Identify Enhancers and Promoters

Yifeng Li¹, Chih-Yu Chen¹, Wyeth Wasserman¹

¹Centre for Molecular Medicine and Therapeutics, University of British Columbia, Canada

Y. Li <yifeng@cmmt.ubc.ca>

W. Wasserman <wyeth@cmmt.ubc.ca>

T30

Efficient and accurate multiple-phenotype regression method for high dimensional data considering population structure

Jong Wha Joo¹, Eun Yong Kang¹, Elin Org¹, Nick Furlotte¹, Brian Parks¹, Aldons Lusis¹, Eleazar Eskin¹

¹UCLA, United States

E. Eskin <eeskin@cs.ucla.edu>

T31

Exploration of Designability of Proteins Using Graph Features of Contact Maps: Beyond Lattice Models

Sumudu Leelananda¹, Robert Jernigan², Andrzej Kloczkowski^{1,3}

¹Nationwide Children's Hospital, United States, ²Iowa State University, United States, ³The Ohio State University, United States

A. Kloczkowski <Andrzej.Kloczkowski@NationwideChildrens.org>

T36

Topological Signatures for Population Admixture

Laxmi Parida¹, Filippo Utro², Deniz Yorukoglu³, Anna Paola Carrieri⁴, David Kuhn⁵, Saugata Basu⁶

¹IBM T J Watson Research Center, United States, ²IBM Research, United States, ³MIT, United States, ⁴University of Milano-Bicocca, Italy, ⁵USDA, United States, ⁶Purdue, United States

L. Parida <parida@us.ibm.com>

T41

COMETS (Constrained Optimization of Multistate Energies by Tree Search): A provable and efficient algorithm to optimize binding affinity and specificity with respect to sequence

Mark Hallen¹, Bruce Donald¹

¹Duke University, United States

B. Donald <brd+recomb15@cs.duke.edu>

T46**Read Clouds Uncover Variation in Complex Regions of the Human Genome****Alex Bishara¹, Yuling Liu¹, Dorna Kashef-Haghighi¹, Ziming Weng¹, Daniel Newburger¹, Robert West¹, Arend Sidow¹, Serafim Batzoglou¹**¹Stanford University, United States

A. Bishara <abishara@cs.stanford.edu>

Y. Liu <yulingl@cs.stanford.edu>

S. Batzoglou <serafim@cs.stanford.edu>

T54**Computational Protein Design Using AND/OR Branch-and-Bound Search****Yichao Zhou¹, Yuexin Wu¹, Jianyang Zeng²**¹Tsinghua University, China, ²Institute for Interdisciplinary Information Sciences, Tsinghua University, China

J. Zeng <zengjy@gmail.com>

T55**A novel probabilistic methodology for eQTL analysis of signaling networks****Roni Wilentzik¹, Irit Gat-Viks¹**¹Tel Aviv University, Israel

R. Wilentzik <roniwile@post.tau.ac.il>

I. Gat-Viks <iritgv@post.tau.ac.il>

T57**Deciding when to stop: Efficient experimentation to learn to predict drug-target interactions****Maja Temerinac-Ott¹, Armaghan W. Naik², Robert F. Murphy²**¹Universität Freiburg, Germany, ²Carnegie Mellon University, United States

M. Temerinac-Ott <temerina@frias.uni-freiburg.de>

R.F. Murphy <murphy@cmu.edu>

T59**An Efficient Nonlinear Regression Approach for Genome-wide Detection of Marginal and Interacting Genetic Variations****Seunghak Lee¹, Aurelie Lozano², Prabhanjan Kambadur³, Eric Xing¹**¹Carnegie Mellon University, United States, ²IBM T. J. Watson Research Center, United States, ³Bloomberg L.P., United States

E. Xing <epxing@cs.cmu.edu>

T63**Gap Filling as Exact Path Length Problem****Leena Salmela¹, Kristoffer Sahlin², Veli Mäkinen¹, Alexandru Tomescu¹**¹Department of Computer Science, University of Helsinki, Finland, ²KTH Royal Institute of Technology, Sweden

L. Salmela <Leena.Salmela@cs.helsinki.fi>

K. Sahlin <kristoffer.sahlin@scilifelab.se>

V. Mäkinen <vmakinen@cs.helsinki.fi>

A. Tomescu <alexandru.tomescu@gmail.com>

T66**A Fast and Exact Algorithm for the Exemplar Breakpoint Distance**Mingfu Shao¹, Bernard Moret¹¹EPFL, Switzerland

M. Shao <mingfu.shao@epfl.ch>

B. Moret <bernard.moret@epfl.ch>

T73**On the Sample Complexity of Cancer Pathways Identification**Fabio Vandin¹, Ben Raphael², Eli Upfal²¹Department of Mathematics and Computer Science, University of Southern Denmark, Denmark, ²Brown University, United States

F. Vandin <vandinfa@imada.sdu.dk>

E. Upfal <eli@cs.brown.edu>

T74**Fragmentation trees reloaded**Kai Dührkop¹, Sebastian Böcker²¹Friedrich-Schiller-University Jena, Germany, ²Friedrich Schiller University Jena, Germany

K. Dührkop <Kai.Duehrkop@uni-jena.de>

S. Böcker <sebastian.boecker@uni-jena.de>

T76**DockStar: A novel ILP based integrative method for structural modelling of multimolecular protein complexes**Naama Amir¹, Dan Cohen¹, Haim J. Wolfson¹¹School of Computer Science, Tel Aviv University, Israel

N. Amir <naamaamir@mail.tau.ac.il>

H.J. Wolfson <wolfson@tau.ac.il>

T89**KGSrna: Efficient 3D kinematics-based sampling for nucleic acids**Rasmus Fonseca¹, Henry van den Bedem², Julie Bernauer³¹Dept. of Computer Science, University of Copenhagen, Denmark, ²Stanford Synchrotron Radiation Lightsource, Stanford University, United States, ³INRIA Saclay-Ile de France, France

R. Fonseca <rfonseca@diku.dk>

T93**CRISPR detection from very short reads using partial overlap graphs**Ilan Ben-Bassat¹, Benny Chor¹¹Tel Aviv University, Israel

I. Ben-Bassat <ilanbb@gmail.com>

T95**HapTree-X: An integrative Bayesian framework for haplotype reconstruction from transcriptome and genome sequencing data**Emily Berger¹, Deniz Yorukoglu¹, Bonnie Berger¹¹Massachusetts Institute of Technology, United States

B. Berger <bab@csail.mit.edu>

T98**Constructing Structure Ensembles of Intrinsically Disordered Proteins from Chemical Shift Data****Huichao Gong¹, Sai Zhang¹, Jiangdian Wang², Haipeng Gong³, Jianyang Zeng¹**¹Institute for Interdisciplinary Information Sciences, Tsinghua University, China, ²Biostatistics and Research Decision Sciences - Asia Pacific, Merck Research Laboratory, China, ³School of Life Sciences, Tsinghua University, China
J. Zeng <zengjy321@tsinghua.edu.cn>**T113****Functional alignment of metabolic networks****Arnon Mazza¹, Allon Wagner², Eytan Ruppin³, Roded Sharan¹**¹Tel Aviv University, Israel, ²University of California, Berkeley, United States, ³University of Maryland, United States

A. Mazza <arnon.ma@yahoo.com>

R. Sharan <roded@post.tau.ac.il>

T114**BWM*: A Novel, Provable, Ensemble-based Dynamic Programming Algorithm for Sparse Approximations of Computational Protein Design****Jonathan D Jou¹, Swati Jain², Ivelin Georgiev¹, Bruce R Donald¹**¹Department of Computer Science, Duke University, United States, ²Computational Biology and Bioinformatics Program, Duke University, United States

B.R. Donald <brd+recomb15@cs.duke.edu>

T119**Ultra-large alignments using ensembles of Hidden Markov Models****Nam-Phuong Nguyen¹, Siavash Mirarab², Keerthana Kumar², Tandy Warnow¹**¹University of Illinois at Urbana-Champaign, United States, ²University of Texas at Austin, United States

N. Nguyen <namphuon@illinois.edu>

S. Mirarab <smirarab@gmail.com>

T. Warnow <warnow@illinois.edu>

T120**Rapidly Registering Identity-by-Descent Across Ancestral Recombination Graphs****Shou Young¹, Shai Carmi¹, Itsik Pe'er¹**¹Columbia University, United States

I. Pe'er <itsik@cs.columbia.edu>

T126**Immunoglobulin classification using the colored antibody graph****Stefano Bonissone¹, Pavel Pevzner²**¹Bioinformatics Program, University of California San Diego, United States, ²Department of Computer Science and Engineering, University of California San Diego, United States

S. Bonissone <sbonisso@ucsd.edu>

T129**Efficient Alignment Free Sequence Comparison with Bounded Mismatches**
Srinivas Aluru¹, Alberto Apostolico¹, Sharma V. Thankachan¹¹Georgia Institute of Technology, United States

S. Aluru <aluru@cc.gatech.edu>

A. Apostolico <axa@cc.gatech.edu>

S.V. Thankachan <sharma.thankachan@gmail.com>

T137**Protein Contact Prediction by Integrating Joint Evolutionary Coupling Analysis and Supervised Learning**Jianzhu Ma¹, Sheng Wang¹, Zhiyong Wang¹, Jinbo Xu¹¹Toyota Technological Institute at Chicago, United States

J. Xu <jinbo.xu@gmail.com>

T138**Haplotype Allele Frequency (HAF) Score: Predicting Carriers of Ongoing Selective Sweeps Without Knowledge of the Adaptive Allele**Roy Ronen¹, Glenn Tesler¹, Ali Akbari¹, Shay Zakov¹, Noah A. Rosenberg², Vineet Bafna¹¹University of California, San Diego, United States, ²Stanford University, United States

V. Bafna <vbafna@cs.ucsd.edu>

T146**CIDANE: Comprehensive isoform discovery and abundance estimation**Stefan Canzar¹, Sandro Andreotti², David Weese², Knut Reinert², Gunnar W. Klau³¹Johns Hopkins University, United States, ²FU Berlin, Germany, ³CWI, Netherlands

S. Canzar <canzar@jhu.edu>

S. Andreotti <Sandro.Andreotti@fu-berlin.de>

T150**Deconvolution Of Ensemble Chromatin Interaction Data Reveals The Latent Mixing Structures In Cell Subpopulations**Emre Sefer¹, Geet Duggal¹, Carl Kingsford¹¹Carnegie Mellon University, United States

E. Sefer <esefer@andrew.cmu.edu>

G. Duggal <geetduggal@gmail.com>

C. Kingsford <carlk@cs.cmu.edu>

T156**CoMEt: A Statistical Approach to Identify Combinations of Mutually Exclusive Alterations in Cancer**Mark Leiserson¹, Hsin-Ta Wu¹, Fabio Vandin¹, Benjamin Raphael¹¹Brown University, United States

B. Raphael <braphael@cs.brown.edu>

T160**Diffusion Component Analysis: Unraveling Functional Topology in Biological Networks****Hyunghoon Cho¹, Bonnie Berger¹, Jian Peng¹**¹Massachusetts Institute of Technology, United States

B. Berger <bab@csail.mit.edu>

J. Peng <jpeng@csail.mit.edu>

T161**ScaffMatch: Scaffolding Algorithm Based on Maximum Weight Matching****Igor Mandric¹, Alex Zelikovsky¹**¹GSU, United States

I. Mandric <mandric.igor@gmail.com>

A. Zelikovsky <alexz@cs.gsu.edu>

T166**A symmetric length-aware enrichment test****David Manescu¹, Uri Keich¹**¹University of Sydney, Australia

U. Keich <uri@maths.usyd.edu.au>

T167**Learning microbial interaction networks from metagenomic count data****Surojit Biswas¹, Meredith McDonald¹, Derek Lundberg¹, Jeffery Dangl², Vladimir Jojic¹**¹University of North Carolina at Chapel Hill, United States, ²HHMI and University of North Carolina at Chapel Hill, United States

S. Biswas <sbiswas@live.unc.edu>

V. Jojic <vjojic@cs.unc.edu>

T170**Joint inference of genome structure and content in heterogeneous tumor samples****Andrew McPherson¹, Andrew Roth², Cedric Chauve¹, S. Cenk Sahinalp¹**¹Simon Fraser University, Canada, ²University of British Columbia, Canada

A. McPherson <andrew.mcpherson@gmail.com>

T173**Locating a Tree in A Phylogenetic Network in Quadratic Time****Philippe Gambette¹, Andreas Gunawan², Anthony Labarre¹, Stephane Vialette¹, Louxin Zhang²**¹Univ. Paris-Est Marne-la-Vallee, France, ²National University of Singapore, Singapore

L. Zhang <matzlx@nus.edu.sg>

Sunday poster session

kindly sponsored by University of Pittsburgh

P178

Target Gene Selection System for Human Disease Related Mouse Models

Soo Young Cho¹, Young Seek Lee², Soojun Park³, Je Kyung Seong¹

¹Seoul National University, The Republic of Korea,

²Hanyang University, The Republic of Korea, ³ETRI, The Republic of Korea

J.K. Seong <snumouse@snu.ac.kr>

Genetically Engineered Mouse (GEM) models are used in high-throughput phenotyping screens for understanding genotype-phenotype associations and its relevance to human disease. However, not all mouse mutant mouse lines with detectable phenotypes are associated with human disease. Here we propose Target gene selection system for Genetically engineered mouse models (TarGo). Using a combination of human disease descriptions, network topology and genotype-phenotype relations, novel genes potentially related to human disease are suggested. We constructed a gene interaction network using Protein-Protein Interaction (PPI), molecular pathway and co-expression data. The information for human disease related genes was obtained from several repositories for human disease signatures. We calculated disease or phenotype specific gene ranking using network topology and disease signature. TarGo provides many novel features for gene function prediction.

P184

Threshold models for genome-enabled prediction of ordinal categorical traits in plant breeding

Osva Antonio Montesinos López¹, Emerico Franco Pérez¹

¹Universidad de Colima, Mexico

O.A.M. López

<oamontes2@hotmail.com>

Categorical scores for disease susceptibility or resistance are often recorded in plant breeding. The aim of this study was to introduce genomic models for analyzing ordinal characters and to assess the predictive ability of genomic predictions for ordered categorical phenotypes using a threshold model counterpart of the Genomic Best Linear Unbiased Predictor (TGBLUP). The threshold model was used to relate a hypothetical underlying scale to the outward categorical response. We present an empirical application where a total of nine models, five without interaction and four with genomic \times environment interaction ($G \times E$) and genomic additive \times additive \times environment interaction ($G \times G \times E$), were used. We assessed the proposed models using data consisting of 278 maize lines genotyped with 46,347 SNPs and evaluated for disease resistance [with ordinal scores from 1 (no disease) to 5 (complete infection)] in three environments (Colombia, Zimbabwe, and Mexico). Models with $G \times E$ captured a sizeable proportion of the total variability, which indicates the importance of introducing interaction to improve prediction accuracy. Relative to models based on main effects only, the models that included $G \times E$ achieved 9-14% gains in prediction accuracy; adding additive \times additive interactions did not increase prediction accuracy consistently across locations.

P190

Alternative definitions of UGA codon regulated by translational regulatory factors and selenium

Chen-Hsiang Yeang¹, Yen-Fu Chen¹, Kai-Neng Chuang¹, Chih-Hsu Lin¹, Shery Yen¹

¹Academia Sinica, Taiwan

C. Yeang <chyeang@stat.sinica.edu.tw>

Translation termination is not always sufficient because of the existence of alternative definitions in stop codons. UGA codon is redefined to incorporate

selenocystein (Sec) through Sec-tRNA^{Sec} in selenoprotein synthesis and regulated by selenium and its translational factors. We developed a quantitative model to depict the dependency of Sec incorporation and termination outcomes under various selenium conditions. The model is based on the following hypotheses: (1) syntheses of the full-length and truncated forms compete for limited mRNA supply, (2) selenoprotein synthesis is limited by Sec-tRNA^{Sec} abundance which is limited by selenium and tRNA^{Sec} precursors, (3) all synthesis reactions follow first-order-kinetics, and (4) truncated peptides possess a considerably shorter half-life than full-length selenoproteins. We demonstrated with both mathematical derivation and simulation that this model captures several prominent characteristics observed from experimental data. First, quantity of full-length selenoprotein increases with elevating selenium concentration and the synthesis level is saturated under high selenium condition. Second, increasing mRNA supply elevates synthesis quantities of both forms yet lowers the ratio of full-length and truncated forms. Furthermore, we developed a conjugate gradient descent algorithm with binary searching strategy to estimate the free parameters of this model from experimental data. The model sheds light on the mechanism of UGA codon redefinition for selenoprotein synthesis and the interplay of regulatory factors.

P217

A Framework for Inferring Fitness Landscapes of Patient-Derived Viruses Using Quasispecies Theory

David Seifert¹, Francesca Di Giallonardo², Karin J. Metzner², Huldrych F. Günthard², Niko Beerenwinkel¹

¹ETH Zurich, Switzerland, ²University Hospital Zurich, Switzerland

D. Seifert <david.seifert@bsse.ethz.ch>

N. Beerenwinkel <niko.beerenwinkel@bsse.ethz.ch>

Fitness is a central quantity in evolutionary models of viruses. However, it remains difficult to determine viral fitness experimentally, and existing in vitro assays can be poor predictors of in vivo fitness of viral populations within their hosts. Next-generation sequencing can nowadays provide snapshots of evolving virus populations, and these data offer new opportunities for inferring viral fitness. Using the equilibrium distribution of the quasispecies

model, an established model of intra-host viral evolution, we linked fitness parameters to the composition of the virus population, which can be estimated by next-generation sequencing. For inference, we developed a Bayesian Markov Chain Monte Carlo method to sample from the posterior distribution of fitness values. The sampler can overcome situations where no maximum likelihood estimator exists, and it can adaptively learn the posterior distribution of highly correlated fitness landscapes without prior knowledge of their shape. We tested our approach on simulated data and applied it to clinical HIV-1 samples to estimate their fitness landscapes in vivo. The posterior fitness distributions allowed for differentiating viral haplotypes from each other, for determining neutral haplotype networks, in which no haplotype is more or less credibly fit than any other, and for detecting epistasis in fitness landscapes. Our implemented approach, called QuasiFit, is available at <http://www.cbg.ethz.ch/software/quasifit>.

P219

Identification of novel therapeutics for complex diseases from genome-wide association data

Mani P. Grover¹

¹Deakin University, Australia

M.P. Grover <manigrover1@gmail.com>

Background

Human genome sequencing has enabled the association of phenotypes with genetic loci, but our ability to effectively translate this data to the clinic has not kept pace. Over the past 60 years, pharmaceutical companies have successfully demonstrated the safety and efficacy of over 1,200 novel therapeutic drugs via costly clinical studies. While this process must continue, better use can be made of the existing valuable data. In silico tools such as candidate gene prediction systems allow rapid identification of disease genes by identifying the most probable candidate genes linked to genetic markers of the disease or phenotype under investigation. Integration of drug-target data with candidate gene prediction systems can identify novel phenotypes which may benefit from current therapeutics. Such a drug repositioning tool can save valuable time and money spent on preclinical studies and phase I clinical trials.

Methods

We previously used Gentrepid (www.gentrepid.org) as a platform to predict 1,497 candidate genes for the seven complex diseases considered in the Wellcome Trust Case-Control Consortium genome-wide association study; namely Type 2 Diabetes, Bipolar Disorder, Crohn's Disease, Hypertension, Type 1 Diabetes, Coronary Artery Disease and Rheumatoid Arthritis. Here, we adopted a simple approach to integrate drug data from three publicly available drug databases: the Therapeutic Target Database, the Pharmacogenomics Knowledgebase and DrugBank; with candidate gene predictions from Gentrepid at the systems level.

Results

Using the publicly available drug databases as sources of drug-target association data, we identified a total of 428 candidate genes as novel therapeutic targets for the seven phenotypes of interest, and 2,130 drugs feasible for repositioning against the predicted novel targets.

Conclusions

By integrating genetic, bioinformatic and drug data, we have demonstrated that currently available drugs may be repositioned as novel therapeutics for the seven diseases studied here, quickly taking advantage of prior work in pharmaceuticals to translate ground-breaking results in genetics to clinical treatments.

P220 miRBP: An architecture to find competitive interaction between microRNA and RNA binding proteins

A Vivekanand¹, G Reshmi¹, M. Radhakrishana Pillai¹

¹Rajiv Gandhi Centre for Biotechnology, India

G. Reshmi <reshmi@rgcb.res.in>

Background: Mature microRNAs (miRNAs) are endogenous small non-coding, about 21- 25 nucleotides in length which are partially complementary to one or more mRNA targets. 'Locking up' of the miRNA target site within a relatively stable RNA secondary structure can hinder miRNA binding by RNA binding proteins (RBP) to the target sites. Overlap of the seed complementary region in the transcript within the binding site for an RBP can result in reduced targeting. Such hindering factors that adversely affect miRNA targeting have not been dealt with in a comprehensive fashion.

Survey of all annotated human transcripts with well-defined 3' UTRs and the coding regions, with respect to the occurrence of miRNA target sites within RNA secondary structure and RNA-protein binding sites, is warranted. miRNAs binding to its target mRNA can be modulated by competitive binding of RBPs. We have proposed a method for finding prevalently occurring motifs and domain of miRNAs binding target sites of all human transcripts including splice variants and highlights in which type of RNA secondary structure these elements were found.

Methods and Results: Curated and downloaded experimentally validated data of RBPs as well as CLASH (Cross Linking and Sequencing of Hybrids) data for miRNA-mRNA interactions. We have developed a pipeline for further identification of interactions between miRNA and RBP in mRNA via machine learning approach to characterize prevalently found miRNA and RBPs which competes. The validation of the results done using available high-throughput datasets.

Conclusion: We have developed a computational platform to identify a broad spectrum of experimentally validated miRNA-mRNA interactions, as well as RBPs overlapping in the miRNA target sites.

P221 Redundans: an assembly pipeline for highly heterozygous genomes

Leszek Pryszcz, Toni Gabaldón¹

¹Comparative Genomics Group. Bioinformatics and Genomics Programme. Centre for Genomic Regulation (CRG), Spain

L. Pryszcz <l.p.pryszcz@gmail.com>

Many genomes display high levels of heterozygosity (i.e. presence of different alleles at the same loci in homologous chromosomes), being those of hybrid organisms an extreme such case. The assembly of highly heterozygous genomes from short sequencing reads is a challenging task because it is difficult to accurately recover the different haplotypes. When confronted with highly heterozygous genomes, the standard assembly process tends to collapse homozygous regions and reports heterozygous regions in alternative contigs. The boundaries between homozygous and heterozygous regions result in multiple paths that are hard to resolve, which leads to highly fragmented assemblies with

a total size larger than expected. This, in turn, causes numerous problems in downstream analyses i.e. fragmented gene models, wrong gene copy number, broken synteny. To circumvent these caveats we have developed a pipeline that specifically deals with the assembly of heterozygous genomes by introducing a step to recognise and selectively remove alternative heterozygous contigs. We tested our pipeline on simulated and naturally-occurring heterozygous genomes and compared its accuracy to other existing tools.

P223

Molecular classification and comparative analysis of *Arthrobacter* genus plasmids

Marius Mihasan¹

¹Alexandru I Cuza University of Iasi, Faculty of Biology, Romania

M. Mihasan <marius.mihasan@uaic.ro>

The genus *Arthrobacter* belong to class Actinobacteria and includes Gram-positive, obligate aerobes bacteria that are often isolated from contaminated or toxic soils. Because of their potential use in detoxification of xenobiotics, these microorganisms have received considerable attention. Albeit that more than 48 *Arthrobacter* genomes have been sequenced so far, little is known about their plasmids and the tools applicable for the genetic manipulation of this bacteria. In the current work, the available sequenced *Arthrobacter* plasmids have been analyzed comparatively in order to identify common genes encoding replication and partition functions. At least one *parA* homolog could be identified on every *Arthrobacter* plasmid. Using the Maximum Likelihood method, the evolutionary relationships of these homologs have been inferred and the *Arthrobacter* plasmids have been classified into 4 clades. A characteristic feature of *Arthrobacter* plasmids is that not all the plasmid presented a *parB* homolog, but when it was present, the *parB* gene was separated by more than 20kb from *parA*. Within the 1 kb region upstream from the putative *parA* gene, directly repeated sequences could be identified on all *Arthrobacter* specific plasmids, with the exception of clade I plasmids. It is highly probable that these sequences function as iterons and thereby mark the replication origin. The pairwise plasmid proteome comparisons showed that the conservation between the *Arthrobacter* spe-

cific plasmid proteoms is low. Beside the *parA* gene, the clade IV plasmids share only one more ORF, namely a putative CHAP -amidase, possibly implicated in plasmid conjugation. The work was supported by the UAIC GI-2014-02 grant.

P224

Trochilus – a platform for evaluation of algorithms that find values of real parameters

Szymon Wasik¹, Jacek Blazewicz¹

¹Poznan University of Technology, Poland

S. Wasik <szymon.wasik@cs.put.poznan.pl>

Finding values of real parameters is an important problem related to the modeling, particularly in bioinformatics. After designing the model the estimation of parameters is required before the full analysis can be conducted regardless of the modeled system nature (viral infection, kinetic network, gene regulatory network, etc.). Because of its importance there have been even created a special session accompanying the IEEE CEC conference that provides evaluation framework that can be used to compare algorithms solving problem of parameters estimation. Nevertheless, the proposed framework has two drawbacks. First, the proposed problems that are used to test algorithms are artificial and weakly related to practical applications. Secondly, the session is organized and evaluates algorithms once every several years, so it is difficult to compare solutions of various researchers that were published between consecutive sessions.

The Trochilus platform is the solution for the lack of standardized, on-line method for evaluation of algorithms that determine values of real parameters in biological systems. It is an on-line Internet application utilizing cloud computing in platform as a service model (PaaS). Using the platform researchers can submit their implementations of algorithms for parameters value estimation. These implementations are then compiled and evaluated automatically in a secure environment based on some bioinformatics models. The researchers can almost instantly see the results of their solutions and comparison with other submitted algorithms. The Trochilus platform has also supplementary functionalities facilitating the research by providing tools for algorithms versioning and dissemination of results.

P225**Heuristic approach for peptide assembly problem****Marcin Borowski¹, Tomasz Głowacki¹, Piotr Formanowicz¹**¹Institute of Computing Science, Poznan University of Technology, Poland

M. Borowski <mborowski@cs.put.poznan.pl>

Determination of amino acid sequences is very important issue of modern molecular biology. Acquisition of knowledge about peptide structure is an important step in discovering their three-dimensional structure and therefore their functionality. Reading sequences of amino acids is called sequencing. There is no analytical method that allows to sequence long peptide chain but methods for determining short peptide sequences are available. The approach to determine long peptide sequence is to cut it into many short pieces, sequence them and assemble these fragments together. There is a need to apply computational methods for assembly process. In our work a peptide assembly problem with errors coming from the digestion phase is considered. The dedicated heuristics to solve this problem is proposed and results of a computational experiment are presented. The results clearly show that the new method strongly outperform other algorithms known from the literature.

P226**Sorting signals targeting mRNA into hepatic extracellular vesicles****Natalia Szóstak¹, Agnieszka Rybarczyk¹, Marta Szachniuk¹, Jacek Błażewicz¹**¹Poznan University of Technology, Poland

N. Szóstak <nszostak@cs.put.poznan.pl>

Intercellular communication mediated by extracellular vesicles (EVs) has proved to play an important role in a growing number of biological processes. mRNA seems to be one of the most interesting content of these vesicles. mRNA localization depends on interactions between the cis-acting elements in the mRNA sequence, known as zipcodes, and trans-acting factors, the RNA-binding proteins.

Here we present results of our research concerning zipcodes targeting mRNA into hepatic EVs. We have combined two approaches: *in silico* and *in vitro*. During the *in silico* phase, we have checked whether mRNA sorting motifs, previously reported

by other groups, may act as cis-acting elements in hepatic cellular system. Negative correlation suggest that the mechanism of mRNA transport in EVs is tissue-specific. More importantly, based on bioinformatics tools, we have found 12 potential motifs, which may act as a zipcode for targeting mRNA into hepatic EVs. Secondary structure of these motifs have been predicted by mfold, showing common folds for most of the candidate motifs. Additionally, miRNA-binding sites scan has been carried out using miRanda, detecting a number of miRNAs that could potentially bind 12 selected motifs. This result supports the potential role of miRNAs in transporting mRNA into EVs. One of the putative zipcode, a 12-nt sequence included in a stem loop-forming region seemed to be particularly interesting. Its ability to target mRNA into EVs has been confirmed by a wet lab experiment. Taking into account that EVs serve as intercellular communicators, our results can have important therapeutics implications.

P227**A model for the effects of ionizing radiation on microRNA-mediated regulation of mRNA levels in cells****Marzena Dolbniak¹, Roman Jaksik¹, Joanna Rzeszowska-Wolny¹, Krzysztof Fujarewicz¹**¹Silesian University of Technology, Poland

M. Dolbniak <marzena.dolbniak@polsl.pl>

Genes are regulated by mechanisms operating at different levels. One of these mechanisms, RNA interference, is based on interactions between RNA-induced silencing complexes (RISCs) and messenger RNAs (mRNAs), which may lead to negative regulation of translation or degradation of mRNA. RISC is addressed to mRNA through complementary interactions between microRNAs and mRNAs, and mutations or modifications of RNA may influence the efficiency of RNA interference.

We hypothesize that oxidation of nucleotides in nucleic acids during exposure of cells to ionizing radiation can influence microRNA-mRNA complementarity and in this way modulate the levels of different mRNAs.

To explore the participation of such a mechanism in the general changes of transcript levels in cells exposed to ionizing radiation, we created a model which predicts changes of mRNA levels on the ba-

sis of the initial levels of miRNAs and mRNAs and compared the changes predicted by this model with those found experimentally in irradiated cultured cells. The model assumes that all mRNA level changes observed a short time after irradiation depend on changes of microRNA-mRNA interactions. Comparison of groups of transcripts which fit or do not fit the model showed that on average these groups differ in transcript length, length of the 3' ends, number of microRNA-targeted sequence motifs, and the structure of their 5' regulatory region, suggesting that our model can be useful in identification of differently regulated transcripts.

Acknowledgement: The work was financially supported by the Polish National Science Center under grants DEC-2012/05/B/ST6/03472

P229

Accurate prediction of nucleotide conformations

Maciej Antczak¹, Tomasz Zok¹, Martin Riedel², David Nebel², Piotr Lukasiak¹, Marta Szachniuk¹, Thomas Villmann², Jacek Blazewicz¹

¹Institute of Computing Science, Poznan University of Technology, Poland, ²University of Applied Sciences Mittweida, Germany

M. Antczak <maciej.antczak@cs.put.poznan.pl>

A knowledge of RNA 3D structure is crucial for better understanding of mechanisms that govern various cellular processes. RNAs 3D structure prediction still remains a difficult challenge. In contrast to proteins, there is no available tool to solve corresponding problem of side-chain conformation prediction for RNAs. Thus, we propose a computational method for reconstruction of high-quality, full-atomic RNA structure using backbone-dependent library of discrete nucleotide conformations onto fixed backbone coordinates of known structure. Proposed solution integrates following components: a backbone-dependent library of conformers used to identify best fitting nucleotide conformations for every input residue and an optimization algorithm for finding of a low-energy conformation in a reasonable time. Conformer libraries were constructed by machine learning approach, which classify nucleotide conformations upon backbone torsion angles, ribose puckering and glycosidic bond of every nucleotide observed in high-resolution

structures. Conformation energy is using non-local steric atom-atom interactions. The optimization algorithm integrates dead-end elimination procedure and graph theory approach to identify global minimum energy conformation. The resultant model is minimized in Cartesian coordinates space using CHARMM force field. We conducted the evaluation test for 40 RNAs of different structural complexity. The average values computed for considered test set was for RMSD: 0.826Å and 0.935Å and for INF: 0.971 and 0.927, for bases only or ribose ring and bases prediction mode respectively. Around 90% of ribose puckering and glycosidic bond angles are predicted correctly within 30° of the X-ray counterparts. We conclude, that proposed approach proves to be useful in the process of structure reconstruction.

P230

Illumina-NextBio Body Atlas now features RNA-seq data.

Eugene Bolotin¹, Chenxi Chen², Apoorva Dhabalia², Jane Su²

¹Illumina.com, United States, ²Illumina Corp., United States

E. Bolotin <ebolotin@illumina.com>

J. Su <jsu@illumina.com>

We have incorporated RNA-seq data into NextBio BodyAtlas (<https://enterprise.nextbio.com>) to allow easy browsing of gene expression signatures for 50 tissues derived from Genotype Tissue Expression Project (GTEx). We achieved this by downloading 729 raw SRA files from GTEx database, assessed the data quality through several software packages such as: FASTQC to access the read quality and RNASEQC to assess levels of RNA-degradation. We then aligned the data using Illumina RNAExpress 1.0 pipeline (STAR aligner followed by manta counter). Count quality was accessed by pairwise correlation and hierarchical clustering. Samples that had between-tissue correlation greater than within-tissue correlation were removed, as well as samples which had 10^6 aligned reads. Overall we excluded 224 samples that failed one or more criteria from a total of 729 samples used. Interestingly, while most tissues displayed high level RNA degradation, the degradation was not correlated to levels of gene expression. We then used edgeR for TMM normalization, followed by differential comparison between each tissue vs all tissues

(excluding comparison tissue). We used ranked differential P-values in NextBio Research Scoring App to rank the tissue specificity of genes between tissues. Using ranked P-values as a signature, 47/50 tissues closely matched the signature of the existing NextBio Affymetrix Array based Body Atlas thus validating our approach. BodyAtlas RNA-seq should be useful for scientific community for its ability to query this dataset by tissue, gene, gene signature (bioset), or pathway (biogroup).

P231

Genes co-localization in topologically associating domains indicates higher co-expression

Rafal Zaborowski¹, Torgeir Hvidsten², Bartek Wilczyński¹

¹Institute of Informatics, University of Warsaw, Poland,

²Department of Chemistry, Biotechnology and Food Sciences, Norwegian University of Life Sciences, Norway

R. Zaborowski <r.zaborowski@mimuw.edu.pl>

B. Wilczyński <bartek@mimuw.edu.pl>

Hi-C is the technique to study genome architecture by analysis of higher order chromatin interactions [1]. Recent studies exploiting Hi-C revealed hierarchical structure of chromatin. In particular a phenomenon of so-called topologically associating domains (TADs) emerged as strongly self-interacting regions of a genome found universally across different species. There are indications that TADs may be responsible for separating different functional loci by influencing the frequency of long range contacts between different parts of chromosomes. It was also reported that TADs are conserved between mice and humans [2].

In this study we investigate whether genes spatial co-localization is preserved across evolution. In our analysis we map genes of two evolutionary related organisms, mouse and human to corresponding TADs. We then calculate average co-expression for each domain based on publicly available microarray experiments data [3]. Finally we look for co-expression patterns depending on whether domain structure were maintained or changed during evolution.

Our results indicate that genes from the same domain are more likely to be co-expressed and that this phenomenon is to some extent conserved

through evolution. Additionally, our results shed new light on the choice of method for topologically associating domains determination from Hi-C data.

[1] Lieberman-Aiden E. et al.: Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* 326, 289 (2009)

[2] Dixon J. R. et al.: Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* 485, 376 (2012)

[3] <http://coexpresdb.jp>

P232

Analytical Formulas for the Potentials of Mean Force of the Interaction of O-phosphorylated and Hydrophobic Amino-Acid Side Chains in Water

Marta Wiśniewska¹, Adam Liwo¹, Mariusz Makowski¹

¹University of Gdansk, Poland

M. Wiśniewska <marta.d.wisniewska@gmail.com>

Phosphorylation is a common post-translational modification of the amino-acid side chains that possess hydroxyl groups (serine, tyrosine, and threonine). The binding the negatively charged group from ATP molecule to such amino-acid side chains causes the changes in local conformations of proteins and the pattern of interactions with other amino-acid side-chains. A convenient characteristic of the side chain – side chain interactions in the context of aqueous environment is the potential of mean force of side-chain pairs in water.

A model of side-chain-side-chain interactions for O-phosphorylated (charged) and hydrophobic side-chains of amino-acids, respectively, to be used in the UNRES force-field, has been proposed. The interaction energy between the nonpolar sites is composed as a sum of a Gay-Berne and a cavity creation terms; while the interaction energy between the nonpolar and charged sites is composed of a Gay-Berne potential, a cavity creation terms and a polarization term. We parametrized the energy function for the models off all twenty one pairs of molecules modeling O-phosphorylated serine (pSer); threonine (pThr); tyrosine, (pTyr), and hydrophobic side-chains of amino acids. All pairs were determined by umbrella-sampling molecular dynamics simulations in explicit water as functions

of distance and orientation, and the analytical expression was fitted to the potentials of mean force. This work was supported by a grant from Polish National Science Centre (UMO-2013/10/E/ST4/00755).

P233

Annotation of genomics data by the bidirectional hidden Markov model unveils variations in Pol II transcription cycle

Benedikt Zacher¹, Michael Lidschreiber², Patrick Cramer², Julien Gagneur¹, Achim Tresch³

¹Gene Center, Germany, ²Max Planck Institute for Biophysical Chemistry, Germany, ³Max Planck Institute for Plant Breeding Research, Germany

B. Zacher <zacher@genzentrum.lmu.de>

J. Gagneur <gagneur@genzentrum.lmu.de>

A. Tresch <tresch@mpipz.mpg.de>

DNA replication, transcription and repair involve the recruitment of protein complexes that change their composition as they progress along the genome in a directed or strand-specific manner. Chromatin immunoprecipitation in conjunction with hidden Markov models (HMMs) has been instrumental in understanding these processes, as they segment the genome into discrete states that can be related to DNA-associated protein complexes. However, current HMM-based approaches are not able to assign forward or reverse direction to states or properly integrate strand-specific (e.g., RNA expression) with non-strand-specific (e.g., ChIP) data, which is indispensable to accurately characterize directed processes. To overcome these limitations, we introduce bidirectional HMMs which infer directed genomic states from occupancy profiles de novo. Application to RNA polymerase II-associated factors in yeast and chromatin modifications in human T cells recovers the majority of transcribed loci, reveals gene-specific variations in the yeast transcription cycle and indicates the existence of directed chromatin state patterns at transcribed, but not at repressed, regions in the human genome. In yeast, we identify 32 new transcribed loci, a regulated initiation-elongation transition, the absence of elongation factors Ctk1 and Paf1 from a class of genes, a distinct transcription mechanism for highly expressed genes and novel DNA sequence motifs associated with transcription termination.

We anticipate bidirectional HMMs to significantly improve the analyses of genome-associated directed processes.

P234

Integrative computational model for tissue specific gene regulation

Paweł Bednarz¹, Bartek Wilczynski²

¹University of Warsaw, Poland, ²Institute of Informatics, University of Warsaw, Poland

P. Bednarz <Pawel.Bednarz@mimuw.edu.pl>

B. Wilczynski <bartek@mimuw.edu.pl>

Understanding the process of cell differentiation, where cells with identical DNA acquire different morphologies requires us to develop accurate models of the mechanisms behind tissue-specific gene expression in multicellular eukaryotic organisms. Unfortunately, our current understanding of these regulatory mechanisms is imperfect and for that reason the methods for systematic identification of regulatory elements and their linkage to genes they regulate are plagued by the problem of inaccurate predictions. Additionally, there are many factors involved in gene regulation, including chromatin accessibility, nucleosome occupancy, transcription factor binding and regulatory domain insulation. To better understand the complex relationships between all these elements, we need computational models integrating available experimental data on different aspects of the process to make more accurate predictions for any tissue-specific gene.

We propose a new model that consists of three layers representing respectively DNA sequential features, nucleosome occupancy and tissue-specific gene expression. To link them together we used machine learning approaches for which the parameters were learned in an iterative expectation-maximization procedure. As a result we obtained a method able to model differences in gene expression between two cell types in adult *Drosophila Melanogaster*. By further investigating the model's structure we were able to find putative sequential features driving differential gene expression and make predictions regarding the locations of true regulatory elements in the vicinity of tissue-specific genes.

P235**Overlapping Genes in Viruses:
Seeking a Unified Theory**Nadav Brandes¹, Michal Linial¹¹The Hebrew University of Jerusalem, Israel

N. Brandes <nadav.brandes@mail.huji.ac.il>

M. Linial <michall@cc.huji.ac.il>

Viruses are the simplest replicating units. Despite a large diversity in shape, size and modes of replication, most viruses (75%) contain overlapping genes (OG). Most theories claim that OG serve for compression purposes. Other theories suggest a role for OG in gene regulation or protein novelty. We seek a unified theory that may explain the extent and patterns of OG throughout the viral world. An unbiased analysis was performed on 100 families from ViralZone that account for all known viruses. We concentrated only on non-trivial OG (overlapping of genes using alternative frames or strands). The OG rate is the sum of OG regions divided by the genome length. Indeed, the genome length and OG rate are in a strong negative correlation. Remarkably, we found that this is mostly a side effect of an overlooked phenomenon – the amount of overlapping is tightly bounded (from 0 to 1500 nt) throughout the entire viral world, regardless to genome length. This overlooked phenomenon applies to all virus sizes, spanning >3 orders of magnitudes in length. Sharpening our observation by limiting the analysis to significant OG (> 300nt) showed that most viruses remain under a bounded threshold, with less than 5-6 significant OG per genome. We conclude that OG has no role in genome compression per-se. Instead, we support the notion of gene novelty. The lack of sequence similarity in OG, the low complexity of many such genes and their short length are consistent with OG as a source for gene novelty and evolution exploration.

P236**Inferring normal and pathological states in cortical neuron circuit model by Sensitivity Analysis algorithms**Jakub Rydzewski¹, Wieslaw Nowak², Giuseppe Nicosia³¹Institute of Physics, Faculty of Physics, Astronomy and Informatics, N. Copernicus University, Poland,²Institute of Physics, N. Copernicus University, Poland,³University of Catania, Italy

J. Rydzewski <jr@doktorant.umk.pl>

The brain activity is to a large extent determined by states of neural cortex microcircuits. The mathematical, nonlinear, 6-parameter model of such a microcircuit was developed in 2008 by Giugliano et al. and applied in the studies of Huntington's disease [1]. Up to now, the link between parameters domains in the mathematical model of Huntington's disease and the pathological states in cortical microcircuits has remained unclear. In this poster, we precisely identify the most crucial input parameters and domains that drive the system into a unhealthy state. Four distinct methods of the sensitivity analysis of this model have been used to indicate parameters crucial in generating rate function R patterns of the microcircuit electrical activity. All the methods have predicted that the mean of synaptic efficacy (J), the size of the simulated neural circuit (N), the background synaptic fluctuating component (s) and the probability of the connection between the neurons (C) are important input factors. We show that the sensitivity analysis together with a constrained sampling of the model's activity and the Jenks natural breaks optimization allow to discriminate computationally between the normal (healthy) and pathological (unhealthy) cortex states and to set up appropriate domains of the parameters. The sensitivity analysis algorithms indicated that s is insensitive to the pathological state while the correlated C and J parameters strongly determine specific pathological regions in the studied model. Thus, the model may be further simplified and used in simulations of other neuronal disorders.

P237 Protein sequence comparison using domain-based approaches

Byungwook Lee¹

¹Korean BioInformation Center, The Republic of Korea
B. Lee <bulee@kribb.re.kr>

The general method used to identify the function of newly discovered proteins is to transfer annotations from well-characterized homologous proteins. The process of selecting homologous proteins can largely be classified into sequence-based and domain-based approaches. Domain-based methods have several advantages for identifying distant homology and homology among proteins with multiple domains, as compared to sequence-based methods. However, these methods are challenged by large protein families defined by ‘promiscuous’ (or ‘mobile’) domains. Here we present a measure of domain architecture similarity, which can be used to identify homologs of multidomain proteins. To distinguish these promiscuous domains from conventional protein domains, we assigned a weight score to each Pfam domain extracted from RefSeq proteins, based on its abundance and versatility. To measure the similarity of two domain architectures, cosine similarity (a similarity measure used in information retrieval) is used. We combined sequence similarity with domain architecture comparisons to identify proteins belonging to the same domain architecture. Using human and mouse proteomes, we compared our method with an unweighted domain architecture method (DAC) to evaluate the effectiveness of domain weight scores. We found that WDAC is better at identifying homology among multidomain proteins.

P238 The Role of Chromosomal Deletions as Prognostic Markers of Drug Sensitivity

Jonathan Crowther¹, Yanyan Cai¹, Layka Abbasi¹, Stein Aerts², Anna Sablina¹

¹VIB, KULeuven, Belgium, ²KULeuven, Belgium
J. Crowther
<jonathan.crowther@cme.vib-kuleuven.be>

The increasing role of pharmacogenomics in oncology compound development and treatment is closely paralleled with the recent advent of high-throughput genomics platforms. Currently, gene

expression signatures are the predominant platform for pharmacogenomic studies, however the stability possessed by fixed genetic chromosomal deletions are appealing in delineating novel biomarkers associated with drug response. The aim of this study is to determine the role of large-scale chromosomal deletions in eliciting response to novel or existing targeted therapies. Loss of the 8p chromosomal arm is one of the most recurring deletions in epithelial cancers. A survey of The Cancer Genome Atlas (TCGA) dataset indicates that approximately 50% of breast cancer patients harbor an 8p deletion. When we have stratified the patients based on 8p status and treatment regimens, we found a worse survival for those harboring an 8p loss-of-heterozygosity (LOH), suggesting that 8p loss triggers resistance to chemotherapies.

To confirm this hypothesis, we have applied the TALEN approach and generated MCF10a isogenic cell lines mimicking the genomic landscape of 8p deletions in TCGA patients. Gene Set Enrichment Analysis (GSEA) has revealed a comparable gene expression signature to that found in TCGA patients with 8p loss-of-heterozygosity. Moreover, we found that 8p LOH cells demonstrated higher survival rate in response to chemotherapeutic drugs such as doxorubicin or vinblastine compared to wild-type 8p cells. These results highlight the potential for the use of chromosomal deletions as prognostic markers of drug response in the clinic.

P239 Combining tree-based and dynamical systems for the inference of gene regulatory networks

Vân Anh Huynh-Thu¹, Guido Sanguinetti¹

¹University of Edinburgh, United Kingdom
V.A. Huynh-Thu <vhuynt@inf.ed.ac.uk>

Reconstructing the topology of gene regulatory networks (GRNs) from time series of gene expression data remains an important open problem in computational systems biology. Existing GRN inference algorithms face one of two limitations: model-free methods are scalable but suffer from a lack of interpretability, and cannot in general be used for out of sample predictions. On the other hand, model-based methods focus on identifying a dynamical model of the system; these are clearly interpretable and can be used for predictions, however they rely

on strong assumptions and are typically very demanding computationally.

Here, we propose a new hybrid approach for GRN inference, called Jump3, exploiting time series of expression data. Jump3 is based on a formal on/off model of gene expression, but uses a non-parametric procedure based on decision trees (called "jump trees") to reconstruct the GRN topology, allowing the inference of networks of hundreds of genes. We show the good performance of Jump3 on in silico and synthetic networks, and applied the approach to identify regulatory interactions activated in the presence of interferon gamma.

Our MATLAB implementation of Jump3 is available at:
homepages.inf.ed.ac.uk/vhuynt/software.html.

P240

An Improved Algorithm for Ancestral Gene Order Reconstruction

Albert Herencsár¹, Krzysztof Piecuch², Broňa Brejová³

¹Department of Computer Science, Faculty of Mathematics, Physics and Informatics, Comenius University in Bratislava, Slovakia, ²Institute of Computer Science, University of Wrocław, Poland, ³Comenius University in Bratislava, Slovakia

A. Herencsár <albertcoder@gmail.com>

B. Brejová <brejova@dcs.fmph.uniba.sk>

Genome rearrangements are large-scale mutations that change the order and orientation of genes in genomes. In the small phylogeny problem, we are given gene orders in several current species and a phylogenetic tree representing their evolutionary history. Our goal is to reconstruct gene orders in the ancestral species, while minimizing the overall number of rearrangement operations that had to occur during the evolution. In the large phylogeny problem, the phylogenetic tree is unknown and it has to be computed, too. The small and large phylogeny problems are NP-hard for most genome rearrangement models.

We have designed a heuristic method for solving the small phylogeny problem, building on ideas from an earlier tool PIVO by Kovac et al 2011. Our tool, PIVO2, contains several improvements, including randomization to select among potentially

many equally good candidates in a hill-climbing search and a more efficient calculation of distances between gene orders. Using PIVO2, we were able to discover better histories for two biological data sets previously discussed in research literature.

We have further extended the PIVO2 algorithm to solve the large phylogeny problem. We build the tree by sequential insertion of the genomes, and we consider several strategies for determining the best location of the inserted genome. After the tree is built, it is further optimized by our solver of the small phylogeny problem and by standard tree rearrangement operations. The software can be found at <http://compbio.fmph.uniba.sk/pivo/>

Preliminary version appeared in Proceedings of Information Technologies - Applications and Theory (ITAT) 2014.

P241

The Dynamics of the N1 Riboswitch Interacting with Aminoglycoside Antibiotics

Marta Kulik¹, Takaharu Mori², Yuji Sugita², Joanna Trylska¹

¹University of Warsaw, Poland, ²RIKEN, Japan

M. Kulik <marta.kulik@wp.pl>

Riboswitches are regulatory elements found in non-coding regions of mRNAs, which bind small ligands and in this way control gene expression in metabolic pathways of bacteria. Therefore, riboswitches could be a promising target for antibacterial compounds. Also, engineered riboswitches are of interest in molecular biology and biotechnology so the methods to design synthetic riboswitches should be developed.

The N1 synthetic riboswitch binds aminoglycoside antibiotics from the neomycin family. Its regulatory activity was shown in yeast (Duchardt-Ferner et al., 2010). Structurally, the N1 riboswitch resembles the mRNA decoding site (A-site), which is also a well-known aminoglycoside binding site in the bacterial ribosomal RNA.

The aim of our project is to understand the ligand recognition mode of this aminoglycoside sensing riboswitch and the structural basis of its switching activity (turning on and off the production of proteins).

We performed all-atom molecular dynamics sim-

ulations of the N1 riboswitch in the free and ligand-bound states. The three ligands were neomycin, ribostamycin and paromomycin, which differently activate the riboswitch *in vivo*. We proposed the structure and equilibrium dynamics of the aminoglycoside-free riboswitch and in the complexes with neomycin and paromomycin. A comparison of the latter two complexes helped explain why this riboswitch is biologically active only upon neomycin binding. Note, that paromomycin and neomycin differ only by one functional group (NH₂ vs OH). To more reliably capture the differences between complexes with active and inactive ligands and enhance the sampling of conformational landscape, we have performed replica-exchange molecular dynamics simulations.

P242

Bioinformatics studies of protein translocons in photosynthetic organelles of the testate amoeba *Paulinella chromatophora*

Przemysław Gagat¹, Andrzej Bodył¹, Paweł Mackiewicz¹

¹University of Wrocław, Poland

P. Gagat <gagat@smorfland.uni.wroc.pl>

About 60 million years ago, the rhizarian amoeba *Paulinella chromatophora* acquired a cyanobacterial endosymbiont in a process resembling ancient plastid primary endosymbiosis that gave rise to glaucophytes, red algae and green plants. Similarly to primary plastids, *Paulinella* endosymbionts (chromatophores), have lost many essential genes, and transferred substantial number of genes to the host nuclear genome via endosymbiotic gene transfer (EGT), including those involved in photosynthesis. This indicates that, similarly to primary plastids, *Paulinella* endosymbionts must have evolved a transport system to import their EGT-derived proteins. We elaborated a model for protein import into *Paulinella* chromatophores based on bioinformatics analyses of the chromatophore genomes (from two *Paulinella* strains: FK01 and CCAC 0185), *Paulinella* EST database and presequences of proteins imported to the chromatophores. For comparative studies, we analysed genomes of primary plastids and all sequenced cyanobacteria. Our model assumes vesicular trafficking to the outer chromatophore membrane and a simplified Tic-like complex at the inner chromatophore mem-

brane. The Tic-like apparatus is composed of Tic21 (protein-conducting channel), Tic32 and Tic62 (calcium and redox-sensing regulatory proteins) and a molecular motor responsible for pulling imported proteins into the endosymbiont stroma; the motor could consist of Hsp93, Hsp70, Hsp40 (Toc12) and GrpE. Because *Paulinella* chromatophores indeed evolved a protein import machinery (part of our model concerning vesicular trafficking to the outer chromatophore has been proved experimentally), the endosymbionts should be acknowledge as true cell organelles, among primary plastids and mitochondria. The participation in the conference was funded by the KNOW Consortium.

P243

RNF: a method and tools to evaluate NGS read mappers

Karel Břinda¹, Valentina Boeva², Gregory Kucherov^{1,3}

¹Université Paris-Est Marne-la-Vallée, France, ²Institut Curie, France, ³CNRS, France

K. Břinda <karel.brinda@univ-mlv.fr>

Aligning reads to a reference sequence is a fundamental step in numerous bioinformatics pipelines. As a consequence, properties of a mapper like sensitivity and precision (with given parameters and on given data) can be critical for the whole pipeline since they influence the overall accuracy of the produced results (e.g., in variant calling). Therefore, there has been an increasing demand of methods for comparing mappers and, especially, measuring effects of their parameters.

Read simulators combined with alignment evaluation tools provide the most straightforward way to evaluate and compare mappers. Every simulated read is accompanied with information about its origin (position in the reference genome). This information is then used for the evaluation of alignments reported by a mapper. In the end, reports with overall statistics about successfulness of the mapper are created.

Because of lack of standards for encoding read origins, every evaluation tool has to be explicitly compatible with the used read simulator. In order to solve this obstacle, we have created a generic format RNF (Read Naming Format) for assigning read names containing encoded read origin.

We have also developed two associated tools: MISHmash for read simulation and LAVender for read alignment evaluation.

MISHmash is a program for simulating samples of NGS reads using several popular tools (DwgSim, Art, Mason, etc.); it provides reads in the RNF format. LAVender is a program for evaluation of mappers using simulated reads in the RNF format. It creates detailed interactive HTML reports such that it is easy to compare given mappers. A great attention is devoted to mapping qualities which are used for parameterization of ROC curves in the diagrams ‘precision – sensitivity’. Adding support for RNF to existing read simulators and alignment evaluation tools is simple and it requires only minor changes in their code.

We believe that RNF will become a widely supported standard which will simplify and innovate the evaluation process of NGS mappers as well as their development.

P244

Multi-level machine learning prediction of protein-protein interactions

Julian Zubek¹, Adam Boniecki², Maciej Mnich³, Marcin Tatjewski¹, Subhadip Basu⁴, Dariusz Plewczynski⁵

¹Institute of Computer Science, Polish Academy of Sciences, Poland, ²Faculty of Mathematics, Informatics and Mechanics, Warsaw University, Poland, ³Faculty of Mathematics and Computer Science, Jagiellonian University, Poland, ⁴Jadavpur University, India, ⁵Centre of New Technologies, University of Warsaw, Poland

J. Zubek <zubekj@gmail.com>

D. Plewczynski <darman@icm.edu.pl>

Accurate identification of protein-protein interactions (PPI) is the key step in understanding proteins’ biological functions, which are typically context-dependent. Many existing PPI predictors rely on aggregated features from protein sequences, however only a few methods exploit local information about specific residue contacts.

In this work we present a two-stage machine learning approach for prediction of protein-protein interactions. We start with the carefully filtered data on protein complexes available for *Saccharomyces cere-*

visiae in the Protein Data Bank (PDB) database. First, we build linear descriptions of interacting and non-interacting sequence segment pairs based on their inter-residue distances. Secondly, we train machine learning classifier to predict binary segment interactions for any two short sequence fragments. The final prediction of the protein-protein interaction is done using the 2D matrix representation of all-against-all possible interacting sequence segments of both analysed proteins.

The level-I predictor achieves 0.88 AUC for micro-scale, i.e. residue-level prediction. The level-II predictor improves the results further by more complex learning paradigm. We perform 30-fold macro-scale, i.e. protein-level cross-validation experiment. The level-II predictor using PSIPRED-predicted secondary structure reaches 0.70 precision, 0.68 recall, and 0.70 AUC, whereas other popular methods provide results below 0.6 threshold (recall, precision, AUC). Our results demonstrate that multi-scale sequence features aggregation procedure is able to improve the machine learning results by more than 10% as compared to other sequence representations.

P245

Predicting eukaryotic signal peptides using hidden Markov models

Michał Burdukiewicz¹, Piotr Sobczyk¹, Paweł Błażej², Paweł Mackiewicz¹

¹University of Wrocław, Poland, ²Wrocław University of Technology, Poland

M. Burdukiewicz <michalburdukiewicz@gmail.com>

N-terminal secretory signal peptides guide proteins to the subcellular endomembrane system and extracellular localization. The signal peptide contains three distinct regions which have different amino acid compositions and usually ends with the cleavage site.

The state-of-art signal peptide predicting software incorporates theoretical knowledge about the structure of signal peptides using hidden Markov models (HMMs). The chosen probabilistic framework enforces some assumptions about structure of signal peptide, e.g., distribution of its length. Here we propose improved prediction method based on a hidden semi-Markov model.

We trained and validated our algorithm on sequences of eukaryotic proteins from UniProt (re-

lease 2014_11). The data set was filtered to remove atypical signal peptides. To minimize the dimensionality of the problem, we aggregated amino acids into several physicochemical groups. We used heuristic algorithm to extract approximate boundaries of the regions. The regional frequencies of physicochemical groups were utilized to train the hidden semi-Markov model which considers variability of signal peptide regions' length. Performance of the novel probabilistic framework called signal.hsmm was assessed in 5-fold cross-validation approach.

The average AUC of signal.hsmm was 0.96. The position of predicted cleavage sites showed the average range of 2.27 amino acids in comparison to real cleavage sites. Having assumptions closer to the real signal peptide structure, signal.hsmm is able to achieve results matching the state-of-art software. The signal.hsmm is accessible as a stand-alone R package and as a web-server (<http://www.smorfland.uni.wroc.pl/signalhsmm>). The participation in the conference was funded by the KNOW Consortium.

P246 **Substrate recognition by Hsp40 chaperones - molecular dynamics studies.**

Maciej Baranowski¹, Stanisław Ołdziej¹

¹Laboratory of Biopolymer Structure, Intercollegiate Faculty of Biotechnology, University of Gdańsk and Medical University of Gdańsk, Poland

M. Baranowski <m.baranowski@biotech.ug.edu.pl>

Hsp40 chaperones are core part of Hsp70/40 system which is crucial for any living organism. In this system, Hsp40 play the role of substrate recognizers, binding unfolded polypeptide chains, and exposing it for Hsp70 for subsequent refolding. However, despite the availability of crystal structures of substrate binding domains of Hsp40, the mechanism of substrate recognition and binding by Hsp40 remains poorly understood. How can a single beta-barrel-like domain bind nearly any of all the proteins present in the cell? How can the chaperone distinguish between folded and unfolded polypeptide?

To gain insight into answers to such questions, we performed a number of all-atom molecular dynamics simulations of two main Hsp40 chaperone from cytosol of yeast *Saccharomyces cerevisiae*: Sis1. The simulations show that, in water solutions, sub-

strate binding domains of Hsp40 chaperones display unusually high degree of plasticity, potentially allowing them to adapt to wide range of different substrates. Moreover, this plasticity allows for significant enlargement of substrate binding cleft when compared to available crystal structures, allowing the chaperone to accommodate larger oligopeptides than those that have been typically tested in the literature. To test this observation, we docked a number of peptides to substrate binding domain of Sis1 and simulated them for few tens of nanoseconds. Based on these simulations, we propose general motif recognized by Sis1-like (class II) Hsp40 chaperones.

Research was funded by Polish National Science Center grant 2013/09/N/NZ2/01979

P247 **Quick Permutation Test: feature filtering of n-gram data** **Piotr Sobczyk¹, Michał Burdukiewicz², Chris Lauber³, Paweł Mackiewicz²**

¹Wrocław University of Technology, Poland, ²University of Wrocław, Poland, ³Dresden University of Technology, Germany

P. Sobczyk <piotr.sobczyk@pwr.edu.pl>

N-grams (k-tuples) are vectors of n characters derived from input sequence(s). They are widely used in bioinformatics as an efficient technique for encoding biological sequence. Unfortunately, converting data to the n-gram format produces cumbersome to analyze high-dimensional spaces.

Relevant n-grams can be selected using permutation test with an objective statistic criterion, which measures relationship between target and feature (n-gram) variables. The permutation test involves computing the criterion statistic for each permutation of feature labels. A fraction of permutations, for which the criterion is more extreme than for the original feature, is p-value. Since the procedure is repeated for each feature, multiple testing enforces the adjustment of obtained p-values.

As permutation tests, when small p-values are required, are computationally expensive and possibly inaccurate, their application for filtering large data sets converted to n-grams is troublesome. However, exact p-values, for this particular permutation test, can be produced significantly faster. Under the null hypothesis, number of times observation has

both positive value of feature and target follows the multinomial distribution. From that, the exact distribution of criterion statistic can be calculated. This approach, called QuiPT (Quick Permutation Test), was compared with results produced by regular permutation test. Information gain was used as the criterion statistic and Benjamini-Hochberg's procedure as p-value adjustment.

Whenever it is possible to calculate the distribution, QuiPT is not only much faster than the permutation test, but also yields more precise results in some scenarios. Our method accurately estimates very small p-values which normally require unfeasibly large number of permutations. QuiPT is a part of the R package biogram (<http://CRAN.R-project.org/package=biogram>).

P248

Monte Carlo simulations of mammalian sex chromosomes evolution.

Piotr Posacki¹

¹University of Wrocław, Poland

P. Posacki <pposacki@wp.pl>

Mammalian X chromosomes are similar to larger autosomes in size and number of genes, whereas Y chromosomes are much smaller and contain only few percent of genes located on X. Y chromosome does not recombine with X, except for a short pseudoautosomal region. X and Y chromosomes descend from one autosome pair that acquired sex determining genes. Therefore it was favorable to switch off recombination in the region containing these genes. One hypothesis suggests that non-recombining region spread due to random accumulation of mutations, and the Muller's ratchet mechanism led to loss of genetic information on Y. On the other hand, adaptive hypothesis explains degeneration of Y as favorable for population by strengthening selection for X, which increased females' fitness and led to higher reproductive potential.

To study the evolution of sex chromosomes, we elaborated Monte Carlo simulations using Penna model. Three modes of XY recombination were tested: no recombination, XY recombination rate (rr) equal to XX rr, and XY rr initially equal to XX rr with possibility of further evolution.

The lack of XY recombination led to fast Y degeneration and X purification in comparison to autosomes. When both XY and XY pairs recombined at equal rate, neither Y degeneration nor X purifi-

cation were observed. In the case of evolving XY rr, this value declined rapidly in population leading to switching off XY recombination, Y degeneration and X purification.

Presented results show that Y degeneration is driven not by random mutation accumulation, but by adaptive mechanism, increasing females' fitness.

P249

Analysis of RNA thermodynamics using Motif Identifier for Nucleic Acid Trajectory

Maciej Jasiński^{1,2}, Anna Górska³, Joanna Trylska¹

¹University of Warsaw, Centre of New Technologies, Poland, ²University of Warsaw, MISMaP College, Poland, ³University of Warsaw, Faculty of Mathematics and Informatics, Poland

M. Jasiński <maciejj@cent.uw.edu.pl>

Molecular dynamics simulations of nucleic acid structures allow monitoring their dynamics on microsecond time scales. However, generated trajectories contain a large number of molecular conformations that have to be post-processed and analyzed. While there are many tools to analyze single RNA or DNA conformations, only few enable analyzing multiple conformations (e.g. thousands trajectory frames).

We have designed Motif Identifier for Nucleic Acid Trajectory (MINT), a software for analyzing 2D and 3D structures of nucleic acid molecules, specifically their full-atom molecular dynamics trajectories or conformation sets derived from other sources (such as X-ray crystallography or nuclear magnetic resonance studies) [1]. For each conformation MINT determines the hydrogen bonding network, identifies secondary structure motifs (helices, triplexes, junctions, loops, bulges etc.) and pseudo-knots. MINT also estimates the electrostatics and energy of stacking interactions. For many conformations, MINT provides average values of the above parameters and their evolution. MINT is a free software released under the terms of the GPLv2.0 license and available at <http://mint.cent.edu.pl>.

Here we show MINT functionality based on all-atom molecular dynamics trajectories of ten base-pair long RNA helix of a mixed sequence. Simulations were performed in NAMD with Amber99 force field in 11 temperatures from 300 K to 500 K. The analysis of the number of hydrogen bonds, the number of

base pairs, the stacking interactions and the average secondary structure in each temperature allows us to estimate the thermodynamic parameters of this helix and compare them with experiments.

Acknowledgements: We acknowledge support from U. Warsaw (CeNT/BST and ICM/KDM/G31-4), National Science Centre (DEC-2012/05/B/NZ1/00035) and European Social Fund (contract number UDA-POKL.04.01.01-00-072/09-00 to MJ).

[1] A. Górska, M. Jasiński, J. Trylska, MINT: software to identify motifs and short-range interactions in trajectories of nucleic acids, 15-Jan-2015 submitted to Nucleic Acids Research journal

P250

Application of Monte Carlo simulations and Metropolis-Hastings algorithm in analysis of mutation accumulation in three codon positions of protein coding sequences

Małgorzata Grabińska¹, Paweł Blazej¹, Paweł Mackiewicz¹

¹University of Wrocław, Poland

P. Mackiewicz <pamac@smorfland.uni.wroc.pl>

Characteristic feature of protein coding sequences is their triplet/codon structure, which is manifested by specific nucleotide composition of three codon positions. This composition is under the influence of selection and mutation processes. In this work, we considered impact of two types of point mutations, i.e. transitions and transversions on protein coding sequences. To study this problem, we elaborated a model of genome evolution based on Monte Carlo simulation approach. The model included also selection against stop codons and a modified version of Metropolis-Hastings algorithm to control specific nucleotide composition of particular codon positions. The simulations were carried out on individuals represented by real bacterial gene sequences. The obtained results showed that excess of transition over transversion is generally profitable for genomes because mean number of eliminated individuals decreased exponentially with the growth of transition/transversion ratio. It is well-known that transversions are more harmful than transitions because they more frequently change coded amino acid. However, the results are not trivial

because the simulations did not consider any selection on coded amino acids. They indicate that nucleotide composition of the first codon positions is optimized for high transition/transversion ratio and these positions appeared most conserved. Interestingly, the second codon positions considered conserved according to the effect on amino acid substitution were more tolerant on mutation accumulation in our simulations. The third codon positions accepted much more mutations than rejected because of very similar composition to the mutational stationary distribution. The participation in the conference was funded by the KNOW Consortium.

P251

Modeling hypothetical amyloid pores

Magdalena Żulpo¹, Malgorzata Kotulska¹

¹Wrocław University of Technology, Department of Biomedical Engineering, Poland

M. Kotulska <malgorzata.kotulska@pwr.edu.pl>

Recently the occurrence of neurodegenerative amyloid-based diseases has increased significantly. These include Alzheimer's disease, Parkinson's disease, amyotrophic lateral sclerosis, and others. Studies indicate that the neurodegenerative processes may also correspond to incorporation of amyloid oligomers into the cell and organelle membranes, creating weakly cation-selective ion channels that allow uncontrolled influx of calcium into nerve cells, which becomes cytotoxic. However, no structure of such channels has been experimentally obtained.

We study the potential of cylindrin structure as a template for modeling amyloid pores, hypothetical transmembrane structures in amyloid diseases. Cylindrin is a six stranded anti-parallel beta barrel from amyloidogenic strands of α B-crystalline, genetically introduced into E. coli cells. It induces cell toxicity of unknown mechanism, which could be related to the amyloid toxicity. Using comparative modeling, we tested stability of cylindrin-based pores obtained with several amyloid forming and non-forming strands, coming from mutated α B-crystalline and prion sup35. We showed that cylindrin could be used as a template for modeling pores from strands of amyloid proteins, although fibril non-formers of the prion could also form a similar structure. Finally, we tested if the cellu-

lar toxicity of cylindrin and related structures may have resulted from its incorporation into the cell membrane and creating conducting ionic channels. The study indicated that cylindrin, its mutants, or amyloid cylindrin-like pores from prion sequences could only locate as peripheral to the membrane, not able to conduct any ions into the cell. This results explains experimental work on large unilamellar vesicles with cylindrin where conductance was not observed.

P252

A Framework For High Throughput Sequencing Analysis Of Childhood Leukemia

Pascal St-Onge¹, Mathieu Lajoie¹, Patrick Beaulieu¹, Simon Drouin¹, Jasmine Healy¹, Daniel Sinnett²

¹Sainte-Justine University Hospital Center, Canada,

²Department of Pediatrics, Faculty of Medicine, University of Montreal, Canada

P. St-Onge <pascal.st-onge@umontreal.ca>

Acute lymphoblastic leukemia (ALL) is the most common pediatric cancer and leading cause of cancer related mortality among children. Because of significant advances in therapeutic regimens, 80% of these children are cured with current therapy. Despite improvements in outcome, many patients (20%) do not respond to current treatments. Moreover, 70% of young adult survivors of childhood cancers will experience long-term effects as a result of their treatments.

New tools that enable better individual tumor characterization and classification are required to improve patient care and outcomes. Key to implementing personalized approaches are the identification of genes and pathways that drive leukemogenesis and modulate drug response, and the development of reliable biomarkers for disease prognosis and treatment. This is now possible with the next generation sequencing technology. However, analysis of high throughput sequencing data represent a significant challenge in terms of computing resources and annotation. We established an automated pipeline to reduce manual interactions that includes the processing of sequencing data from various sources (exome, transcriptome, genome), quality control, rule based filtering and variant annotations. The results are then integrated into a MongoDB database which allow for a rapid detec-

tion of potentially relevant genomic variants and report generation.

With data from over a thousand samples, our pipeline has proven robust and flexible to easily manage important amount of data and different project. It is a first step towards the implementation of a system that would enable integration of multiple types of analysis as well as provide a secure and reliable way of transferring the knowledge from the laboratory to the clinic.

P253

Critical dynamics of gene networks is behind ageing and Gompertz law

Dmitry Podolsky¹, Ivan Molodtsov², Alexander Zenin³, Valeria Kogan², Andrey Tarkhov⁴, Leonid Menshikov⁵, Robert J Shmookler Reis⁶, Peter Fedichev²

¹Massachusetts Institute of Technology, United States,

²Moscow Institute of Physics and Technology, Russian Federation,

³Quantum Pharmaceuticals, Russian Federation,

⁴Moscow State University, Russian Federation,

⁵Northern (Arctic) Federal University, Russian Federation,

⁶University of Arkansas for Medical Sciences, United States

P. Fedichev <peter.fedichev@gmail.com>

Several animal species are considered to exhibit what is called negligible senescence, i.e. they do not show signs of functional decline or any increase of mortality with age. Recent studies in naked mole rat and long-lived sea urchins showed that these species do not alter their gene-expression profiles with age as much as other organisms do. This correlates well with exceptional endurance of naked mole rat tissues to various genotoxic stresses. We quantitatively analyzed the relation between stability of gene regulatory networks (GRNs), mortality and the process of aging, constructed stochastic models of ageing in age-dependent microarray datasets and found that gene networks of most species are inherently unstable. Over a time the instability causes an exponential accumulation of gene-regulation deviations leading to death. However, should the repair systems be sufficiently effective, the gene network can stabilize so that gene damage remains constrained along with mortality of the organism. We applied the suggested model to analyze age-dependent gene expression datasets of model animals and derived a form of the Gompertz law,

relating ageing and mortality with the stochastic genetic network instability. At the same time, this model accounts for the apparently age-independent mortality observed in some exceptionally long-lived animals. The presented analysis provides a new way to analyze effects of aging encoded in the modern -omic data. We suggest a systematic approach to identify biomarkers of aging and develop anti-aging therapeutics.

P254

RNApdbee - secondary structure retrieval from knotted and unknotted RNA structures

Tomasz Zok¹, Maciej Antczak¹, Mariusz Popena², Piotr Lukasiak¹, Ryszard Adamiak^{1,2}, Jacek Blazewicz³, Marta Szachniuk²

¹Institute of Computing Science, Poznan University of Technology, Poland, ²Institute of Bioorganic Chemistry, Polish Academy of Sciences, Poland, ³Poznan University of Technology, Poland

T. Zok <tomasz.zok@cs.put.poznan.pl>

RNA secondary structure carries important information about interactions within the respective molecule. It may serve as input data in many bioinformatics applications, thus, allowing to search for structural motifs, compare interaction networks, predict three dimensional folds, etc. There are distinct approaches to assess the secondary structure of RNA: a wet-lab experiment, in-silico prediction from sequence, a phylogenetic approach or base pair retrieval from PDB files. Here we focus on the latter one, and we present RNApdbee, a fast and highly accurate method aimed to deal with this problem. PDB files often contain inconsistencies in numbering/naming schemes. There are protein-RNA complexes, missing or modified residues and ligands. Furthermore, even a full validation of PDB file does not guarantee a unique secondary structure, because on the atomic level the order of pseudoknots is not revealed directly and needs to be inferred in an optimization scheme. The presented method is designed to handle any PDB input. Its most important advantage lies in ability to precisely analyze high-order pseudoknots which is unavailable in other existing applications for secondary structure processing. RNApdbee has been implemented as a webserver. It allows users to upload their own data or to automatically fetch

the requested structure upon provided PDB id. The input is thoroughly validated and processed by one of available base-pair analyzers. The list of base-pairs is iteratively unknotted to infer the order of pseudoknots. The result is presented in BPSEQ, CT and extended dot-bracket formats as well as visualized graphically. RNApdbee is freely available at <http://rnapdbee.cs.put.poznan.pl>

P255

An introduction to miRseqViewer: Visualization tool for microRNA analysis

Insu Jang¹, Byungwook Lee¹

¹Korean Bioinformation Center, The Republic of Korea
I. Jang <insoo078@kribb.re.kr>

Deep sequencing of small RNAs has become a routine process in recent years, but no dedicated viewer is as yet available to explore the sequence features simultaneously along with secondary structure and gene expression of microRNA (miRNA). We present a highly interactive application that visualizes the sequence alignment, secondary structure and normalized read counts in synchronous multipanel windows. This helps users to easily examine the relationships between the structure of precursor and the sequences and abundance of final products and thereby will facilitate the studies on miRNA biogenesis and regulation. The project manager handles multiple samples of multiple groups. The read alignment is imported in BAM file format. Implemented features comprise sorting, zooming, highlighting, editing, filtering, saving, exporting, etc. Currently, miRseqViewer supports 84 organisms whose annotation is available at miRBase. miRseqViewer is available at [hOp://msv.kobic.re.kr](http://msv.kobic.re.kr).

P256

Optimality of the canonical genetic code under multiobjective and mutually exclusive criteria

Małgorzata Wnętrzak¹, Paweł Błażej¹, Paweł Mackiewicz¹

¹University of Wrocław, Poland

M. Wnętrzak <mw@smorfland.uni.wroc.pl>

There are different theories concerning origin and structure of the canonical genetic code. However, none of them is unambiguously supported. In this

work, we tested the adaptation hypothesis assuming that the genetic code evolved to minimize effects of harmful mutations and translational errors. Therefore, we compared the canonical code with the best and the worst alternatives, which were evaluated by the multiobjective optimization approach. We considered three objectives, which were costs of amino acid substitutions computed for three positions in codon, separately. For each codon position, we calculated the sum of differences in polar requirements of coded amino acids resulting from all possible single point mutations of a given base in all codons. We examined two models of codes, both preserved the characteristic degeneracy and codon block structure of the canonical code. To generate potential codes under these models, we changed assignments of amino acids to blocks of codons with the same level of degeneration (model 1) or without any constraints (model 2). Additional options for these models included variable cost of mutations involving stop codons. We considered codes with the assignment of the stop translation signals to the same codons as in the canonical code and codes with their unrestricted “coding”. The results indicated that the canonical code shows some optimization to minimize effects of deleterious mutations.

The participation in the conference was funded by the KNOW Consortium.

P257

Influence of methods and molecular markers' selection on inferring phylogenetic relationships in the example of parrots from tribe Arini

Aleksandra Krocak¹, Adam Urantówka², Paweł Mackiewicz¹

¹Department of Genomics, Faculty of Biotechnology, University of Wrocław, Poland, ²Department of Genetics, Faculty of Biology, Wrocław University of Environmental and Life Sciences, Poland

A. Krocak <olakrocak@op.pl>

Many phylogenetic studies based on the same taxa set often produce inconsistent tree topologies. It can result from efficiency of methods and molecular markers used in the reconstruction of phylogenetic trees. To systematically study this problem, we considered eight phylogenetic approaches: two Bayesian analyses (BA) in MrBayes and PhyloBayes, two maximum likelihood (ML) studies in TreeFinder and PAUP, as well as maximum parsimony (MP) and distance methods: neighbor joining (NJ), minimum evolution (ME) and weighted least squares (WLS), in PAUP. The methods were tested on 18 separated and concatenated mitochondrial markers from 7 complete genomes of parrots from tribe Arini. In ML, ME, WLS and MP methods all possible (about 10,000) tree topologies were considered. There was very weak consensus in the best trees found for the particular markers. Even for the same marker different methods proposed various topologies. Average distances between trees measured by symmetric difference of Robinson and Foulds ranged from 5.2 (nd2gene) to 0.9 (control region). In the case of the concatenated set, two trees were proposed, one supported by BA and ML methods, whereas the second by MP and distance approaches. Trees for only four markers and some methods were the same as for the concatenated set. The results showed great variation in phylogenetic analyses based on different methods and mitochondrial markers, which carry different phylogenetic signals. The reliable phylogeny of parrots basing on mitochondrial markers can be obtained only for complete genomes. The participation in the conference was funded by the KNOW Consortium.

mony (MP) and distance methods: neighbor joining (NJ), minimum evolution (ME) and weighted least squares (WLS), in PAUP. The methods were tested on 18 separated and concatenated mitochondrial markers from 7 complete genomes of parrots from tribe Arini. In ML, ME, WLS and MP methods all possible (about 10,000) tree topologies were considered. There was very weak consensus in the best trees found for the particular markers. Even for the same marker different methods proposed various topologies. Average distances between trees measured by symmetric difference of Robinson and Foulds ranged from 5.2 (nd2gene) to 0.9 (control region). In the case of the concatenated set, two trees were proposed, one supported by BA and ML methods, whereas the second by MP and distance approaches. Trees for only four markers and some methods were the same as for the concatenated set. The results showed great variation in phylogenetic analyses based on different methods and mitochondrial markers, which carry different phylogenetic signals. The reliable phylogeny of parrots basing on mitochondrial markers can be obtained only for complete genomes. The participation in the conference was funded by the KNOW Consortium.

P258

Outlier sample detection and exclusion improves differential gene expression analysis

Ilona Grabowicz¹, Bartek Wilczyński¹

¹Warsaw University, Poland

I. Grabowicz <ilona.grabowicz@mimuw.edu.pl>

Differential gene expression analysis is a very commonly performed assay. In particular, it is often performed on dissected tissues to measure occurring gene expression changes in response to different treatments. This can lead to large expression changes in some tissues and smaller in others.

In our study, we focus on differential gene expression analysis using cDNA microarrays. Samples come from different brain tissues from several animals, of which half are undergoing a treatment. As often happens, we are facing very low number of samples. The natural variability among samples coupled with such low number of them can lead to skewed final result, in case there is even a single outlier individual.

Such outlier individuals can be detected by clustering the expression profiles coming from different tissues and identifying the samples failing to cluster in the appropriate group (treatment or control). The same analysis can lead to identification of tissues with strong or weak transcriptional response to the stimulus based on the clustering quality. Our results also indicate that removal of outlier individuals from the dataset can significantly improve the differential expression in the tissues showing strong transcriptional response.

P259 **Computational prediction of changes in quantities of immune cell types**

Amit Frishberg¹, Irit Gat-Viks¹

¹Tel Aviv University, Israel

A. Frishberg <frish.amit@gmail.com>

Immune cell quantification is an important tool for the understanding of cell physiology and for the distinction between different disease phenotypes. Standard technologies (such as FACS or CyTOF) are limited in their ability to quantify a large number of individual cell subpopulations due to the high costs and efforts per a single cell subset. A recent statistical method, called DCQ (digital cell quantification) opened the way for a computational prediction of in vivo changes in quantities of immune cell subpopulations based on gene expression profiles in a whole complex tissue (Altboim et al., 2014). In this study we developed a novel approach that builds on the former DCQ algorithm in more accurate and robust estimates. We tested our approach in simulations and on real data, demonstrating its capabilities in mapping immune cell quantities in a systematic unbiased manner.

P260 **AMYload – web service dedicated to amyloidogenic proteins.**

Paweł Woźniak¹, Małgorzata Kotulska¹

¹Department of Biomedical Engineering, Wrocław University of Technology, Poland

P. Woźniak <pawel.p.wozniak@pwr.edu.pl>

A significant growth in the number of patients with neurodegenerative diseases, such as Alzheimer's or Parkinson's disease, has been observed recently.

Studies show that these diseases are related to the occurrence of specific, amyloidogenic protein sequence fragments, which are prone to aggregate. The analysis of these fragments can provide new knowledge about mechanisms of neurodegenerative diseases development. In the Internet, different websites can be found which contain sets of amyloidogenic sequences. However, these sets are typically represented by the plain text, hence difficult to browse or use for more advanced analysis or modelling.

We present the AMYload web portal which gathers amyloidogenic sequence fragments from different sources, such as TANGO, AmylHex, AmylFrag, and others. AMYload provides easy way to filter the data, e.g. according to fragment length, subsequence occurrence, or the protein name. Selected fragments, along with more extended information and references, can be downloaded in one of several supported file formats. Also, AMYload allows users to add their own sequences which can be later available in the database. Finally, our portal provides the tools for analysis of FASTA sequences with regard to the occurrence of amyloidogenic fragments. For this purpose, different methods, such as FoldAmyloid or AGGRESCAN are implemented. The AMYload website provides more advanced and comfortable ways of studying amyloidogenic sequences. The website is available at <http://comprec-lin.iar.pwr.edu.pl/amyload/>.

P261 **Detection of putative regulatory regions and genomic signal interactions**

Klev Diamanti¹, Husen M. Umer¹, Marcin Kruczyk¹, Marco Cavalli², Claes Wadelius², Jan Komorowski^{1,3}

¹Department of Cell and Molecular Biology, Uppsala University, Sweden, ²Department of Immunology, Genetics and Pathology, Uppsala University, Sweden, ³Institute of Computer Science, Polish Academy of Sciences, Poland

K. Diamanti <klev.diamanti@icm.uu.se>

J. Komorowski <jan.komorowski@icm.uu.se>

Transcription factors (TF) bind to DNA regions and regulate transcription. Considering the large amount of data produced by the ENCODE project, TF binding sites (TFBS) can be used to identify regulatory regions in a genome-wide scale. Here we

focused on TF ChIP-seq datasets for seven human cell lines and developed a versatile method to define regulatory regions that clusters TFBSs based on their density. In order to provide an insight of the TFs' synergistic mechanism we modeled three TF-interaction networks based on their physical genomic locations: neighboring, overlapping and co-occurring. As a result, we outline tethering and antagonistic TF-complexes, and protein-protein interactions.

We identified 128k putative regulatory regions on average, with the majority of them being shorter than 300bp. The putative regulatory regions overlap to a great extent with DNaseI hypersensitive sites (DHS) and more than 99% of them intersect with ENCODE annotations. On average we found 20k putative regulatory elements lying in heterochromatic domains, indicating a large regulatory potential in regions presumed to be transcriptionally silent. Among the strongest TF connections identified in the heterochromatic networks are CTCF and the components of the cohesin complex, RAD21 and SMC3, and the strong cooperation between proteins of the same family such as USF1-USF2 and MAFF-MAFK. Finally, we investigated the implication of the obtained regions on genome looping formation. We discovered that the putative regulatory regions were present in more than 90% of the 10k 3D interacting domains and that the heterochromatic regions cover 30% of this poorly explored mechanism.

P262

Dissecting Dynamic Genetic Variation that Controls Temporal Gene Response in Yeast

Avital Brodt¹, Maya Botzman¹, Eyal David¹, Irit Gat-Viks¹

¹Tel Aviv University, Israel

A. Brodt <avital.brodt@gmail.com>

Inter-individual variation in regulatory circuits controlling gene expression is a powerful source of functional information. The study of associations among genetic variants and gene expression provides important insights about cell circuitry but cannot specify whether and when potential variants dynamically alter their genetic effect during the course of response. Here we develop a computational procedure that captures temporal changes in genetic effects, and apply it to analyze tran-

scription during inhibition of the TOR signaling pathway in segregating yeast cells. We found a high-order coordination of gene modules: sets of genes co-associated with the same genetic variant and sharing a common temporal genetic effect pattern. The temporal genetic effects of some modules represented a single state-transitioning pattern; for example, at 10-30 minutes following stimulation, genetic effects in the phosphate utilization module attained a characteristic transition to a new steady state. In contrast, another module showed an impulse pattern of genetic effects; for example, in the poor nitrogen sources utilization module, a spike up of a genetic effect at 10–20 minutes following stimulation reflected inter-individual variation in the timing (rather than magnitude) of response. Our analysis suggests that the same mechanism typically leads to both inter-individual variation and the temporal genetic effect pattern in a module. Our methodology provides a quantitative genetic approach to studying the molecular mechanisms that shape dynamic changes in transcriptional responses.

P263

Drug Response Modeling of Cancer using Network Pharmacology

Uma Shankavaram¹, Orieta Celiku¹, Xiang Deng¹, Shuping Zhao¹, Anita Tandle¹, Kevin Camphausen¹

¹NIH, United States

U. Shankavaram <uma@mail.nih.gov>

Cancer is a complex disease generally caused by multiple factors, which hamper effective drug discovery. Drug combination or multi target agents provide an alternate way to effectively modify disease networks. Synergistic drug pairs have special potential for treatment since they allow a desired effect to be achieved with lower total dose of administered medicine and usually with fewer side effects. However, the major challenge has been the prediction of chemotherapeutic efficacy based on the biological profile of the tumor.

Because of the lack of gene expression data treated with drug combinations, in this study, we present a computational approach to identify effective drug combinations by exploiting high throughput data. We constructed combinatorial effects of drugs based on large-scale screening data on drug treatment efficacies of 130 drugs under clinical and preclinical

investigation and drug-target binding affinities. We evaluated functional interactions of drugs and their targets using model derived target inhibitory network on glioma cell subset.

Cancer cell based target inhibition network analysis in two case studies using glioma cell lines (U87 and U251) identified several cancer specific pathways, including Focal adhesion, cell cycle, and ECM remodeling. We estimated target synergy scores and identified several synergistic pairs with potential clinical relevance. Target inhibition modeling allowed systematic exploration of functional interactions between drugs and their targets to maximally inhibit multiple survival pathways. This is an ongoing study and currently we are validating some drug combinations in GBM cells. We are hopeful that this systematic approach might recognize novel drug combinatorial treatment modalities for GBM.

P264

Combining genetic polymorphisms with inherited variation in gene expression to explain organismal physiological traits:

Tom Harel¹, Irit Gat-Viks¹

¹Tel Aviv University, Israel

T. Harel <harel@mail.tau.ac.il>

Achieving a good quantitative mechanistic understanding of disease phenotypes based on genetic, molecular and cellular datasets is an important goal in the field of system genetics. One of the central aspects of this goal is understanding how genetic and gene expression datasets combine to encode the phenotypic output in a genetically diverse population of individuals. Despite many studies about disease phenotypes, it is surprising how little we know about the combined effect of genetics and gene expression on phenotypes. In particular, most extant methods utilize only a single type of predictor: classification methods use gene expression data to explain traits, whereas quantitative genetics methods utilize genetic variants to infer the levels of phenotypes. Here, we devise an algorithm for combining both molecular (gene expression) data and genetic variation to infer phenotypes using a regression tree model. Our algorithm was applied to study molecular phenotypes, using genotypes and gene expression data that was measured in bone marrow-derived dendritic cells from recombinant inbred mice strains under viral stimulations. Set-

ting CXCL11 expression level as the phenotype we found several variants and the expression of *Ifnb1* to best explain the phenotypic variation, thus validating the known relation between those immune system components.

P265

The ClpB disaggregase: molecular dynamics simulation of a remodelling machine

Ewa Golas¹, Adam Liwo¹, Harold A. Scheraga², Cezary Czaplewski¹, Jaroslaw Marszalek³

¹University of Gdansk, Poland, ²Cornell University, United States, ³Uniwersytet Gdański Gdański Uniwersytet Medyczny Intercollegiate Faculty of Biotechnology UG&MUG, Poland

E. Golas <ewa.golas.chem@gmail.com>

ClpB belongs to the HSP100 family of molecular chaperones, whose major focus is the disaggregation of protein aggregates. As a member of the AAA+ superfamily, the protein assembles into a multimeric complex to produce a motor protein that is fueled by the energy of ATP hydrolysis. Aggregate substrate is threaded through the complex's central canal. The aim of the present study is to delineate a mechanical connection between local conformational changes in the nucleotide-bearing subunits of ClpB and the corollary of protein dynamics in the entire hexameric complex. As each monomer of ClpB contains two nucleotide-binding subdomains, nine occupancy patterns for the binding of the ADP and ATP nucleotides were investigated by canonical molecular dynamics simulation. The primary tool for the extraction of correlated behaviors was Essential Dynamics, a Principal Component Analysis-based technique. Several trends in protein dynamics are highlighted, as a function of bound nucleotide in the monomers of the complex. In light of their contribution to the threading potential, the dynamics of regions in the central canal are also discussed.

P266**Petri Nets as a Novel Representation of Biomolecular Simulations Data**Anna Gogolinska¹, Wieslaw Nowak²¹Faculty of Mathematics and Computer Science, N.Copernicus University, Poland, ²Institute of Physics, N.Copernicus University, Poland

A. Gogolinska <leii@mat.umk.pl>

W. Nowak <wiesiek@fizyka.umk.pl>

Petri nets (PNs) are mathematical modeling language. The PN has a simple form of a bipartite graph with two types of nodes: places and transitions. Those elements can be associated with various objects, according to the modeled system. A marking describes a distribution of tokens over places. Due to changes in marking PNs are dynamical structures: transitions may fire and transfer tokens among places. Recently we successfully applied PN formalism for modeling of the immune system [1].

Molecular Dynamics simulations (MD) are very useful but time consuming calculations of time evolution of proteins' structures. The MD output files are very large and hard to analyze. Here we report our efforts to create a few novel types of PNs based on the MD trajectories. A few distinct algorithms were designed: OPOAv1, OPOAv2, OPOC. In each algorithm elements of PN represent different aspects of the modeled molecule, e.g. places may represent a conformation of the protein or a localization of one aminoacid. The dynamic of our PNs is adapted to mimic effectively selected aspects of MD simulations. This PN approach may facilitate analysis of large sets of MD data, in particular allows for fast estimates of main features of proteins conformational space. We will present results for chemokines and transport proteins. To the best of our knowledge PNs has been never applied in MD biosimulations before.

[1] Gogolinska, A. and W. Nowak, Petri Nets Approach to Modeling of Immune System and Autism Artificial Immune Systems, 2012, Springer Berlin / Heidelberg. p. 86-99.

P267**Mathematical model of the p53-dependent apoptotic pathway abnormalities in the non-small cell lung cancer**Magdalena Ochab¹, Krzysztof Puszynski¹¹Silesian University of Technology, Poland

M. Ochab <magdaochab@gmail.com>

K. Puszynski <krzysztof.puszynski@polsl.pl>

Non-small cell lung cancer (cell line H157) is one of the most lethal cancer because of relatively low sensitivity to chemotherapy compared to small cell carcinoma. Previous biological research shows that cell line H157 have abnormality in p53-dependent apoptotic pathway. It is published that line H157 does not express PTEN protein and have constitutively active AKT/PKB. There are some controversy between different publications about property of protein p53 in H157, but it is well established that p53 have a mutation in codon 308, in nuclear localisation sequence. In our work we created mathematical model of normal and mutated cells. The model consists of several modules, including negative feedback loop with MDM2, and positive feedback loop with PTEN, PIP and AKT, multistep ubiquitylation and apoptotic proteins from Bcl-2 family on mitochondrial membrane. In mutated cell case, our model takes into account lack of protein PTEN, which destroy a positive feedback loop and mutation in protein p53, which disrupts properly p53 activity. We show that ability to activate apoptosis in mutated cells after irradiation is lower than in normal cell. Our result shows high active AKT level in cells, what is consistent with biological results. Reduction of the phosphorylated AKT level by addition of the PIP inhibitor is sufficient to induce apoptosis in a part of cell population. H157 cell are more sensitive to irradiation after reducing active AKT level, for example by addition PIP inhibitor.

This project was funded by the Polish National Centre for Science granted by decision number DEC-2012/05/D/ST7/02072.

P268**Implementation of hydrodynamic interactions in molecular dynamics simulations with the UNRES force field**

Agnieszka G. Lipska¹, Artur Giełdoń¹, Steve R. Seidman², Harold A Scheraga², Adam Liwo¹

¹University of Gdansk, Poland, ²Cornell University, United States

A.G. Lipska <agnieszka.lipska@phdstud.ug.edu.pl>

Simulations of protein-folding pathways and folding kinetics, together with experimental information, are very important in computational biology. Use of coarse-grained force fields with implicit solvent extends the time- and size-scale of simulations tremendously; however, care must be taken to account for all major factors that influence folding. One of these factors is hydrodynamic interactions (HIs) which are manifested as apparent forces that drive two objects moving through liquid towards each other. In this work, to model HIs, we introduced the Rotne-Prager (RP) tensor of friction coefficients into the Langevin equations of motion with the coarse-grained United Residue (UNRES) model of polypeptide chains developed in our laboratory. The effect of HIs was assessed by running UNRES/Langevin MD simulations of staphylococcal protein A (1BDD) and the FBP 28 WW domain (1E0L). It was found that introduction of HIs slows down the folding because of fast formation of non-native contacts, which act as kinetic traps.

Supported by grants: DEC-2012/06/A/ST4/00376 from the National Science Center of Poland, Mistrz 7./2013 from the Foundation for Polish Science (to AL), GM-14312 from the U.S. National Institutes of Health, and by MCB10-19767 from the U.S. National Science Foundation (to HAS). Computer resources were provided by: (a) the National Science Foundation (<http://www.nics.tennessee.edu/>), (b) the supercomputer resources at the Informatics Center of the Metropolitan Academic Network (ICMAN) in Gdańsk, (c) the Beowulf cluster at the Baker Laboratory of Chemistry, Cornell University, and (d) the Beowulf cluster at the Faculty of Chemistry, University of Gdańsk.

P269**A role of inflammation and immunity in essential hypertension and cardiovascular disease – modeled and analyzed using Petri nets**

Dorota Formanowicz¹, Agnieszka Rybarczyk^{2,3}, Marcin Radom², Piotr Formanowicz^{2,3}

¹Department of Clinical Biochemistry and Laboratory Medicine, Poznan University of Medical Sciences, Poland, ²Institute of Computing Science, Poznan University of Technology, Poland, ³Institute of Bioorganic Chemistry, Polish Academy of Sciences, Poland

A. Rybarczyk <arybarczyk@cs.put.poznan.pl>

The most common type of hypertension, i.e. essential hypertension, whose exact cause remains unknown, is a major risk factor for cardiovascular disease in the western world. Despite numerous studies, its heterogeneous etiology remains poorly understood, making this disease difficult to study and treat. Recently, low-grade inflammatory process together with the innate and adaptive immune responses has been proposed to play a key role in the pathogenesis of this phenomenon. In this study, to investigate the participation of the various factors in the development of this type of hypertension, a Petri net based model has been build and then analyzed. The analysis consisted of generating MCT-sets and t-clusters using specifically selected clustering method. The application of systems approach that has been used in this research has enabled for an in-depth analysis of the studied phenomenon and has allowed to draw valuable biological conclusions.

P270**The importance of iron in the process of atherosclerosis modeled by Petri net based approach**

Dorota Formanowicz¹, Marcin Radom², Agnieszka Rybarczyk^{3,4}, Piotr Formanowicz^{3,4}

¹Department of Clinical Biochemistry and Laboratory Medicine, Poznan University of Medical Sciences, Poland, ²Poznań University of Technology, Institute of Computing Science, Poland, ³Institute of Bioorganic Chemistry, Polish Academy of Sciences, Poland, ⁴Institute of Computing Science, Poznan University of Technology, Poland

M. Radom <marcin.radom@cs.put.poznan.pl>

Despite enormous progress in the understanding of the underlying causes of atherosclerosis, our knowledge is still insufficient, so that we may fully effectively treat this disease and prevent its clinical consequences. There are many factors needing to be considered that play an important role in this phenomenon. One of them are iron ions that catalyze the Fenton reaction. As a result highly toxic hydroxyl radical, which is involved in lipid peroxidation, is formed. In this way modified lipids without limitation are trapped by the macrophages and become a substrate for the atherosclerotic plaque. In this work a systems approach to the study of this complex issue is presented. For this purpose, a Petri nets based model of the participation of iron in the development of atherosclerosis has been built. Afterwards, the analysis of the presented model, based on the generation of MCT-sets and t-clusters calculated on the basis of a set of t-invariants using an appropriately matched clustering method, has allowed to draw biologically interesting conclusions.

P271

Efficient Indexing of Read Collections for Likelihood Computation

Vladimír Boža¹, Jakub Jursa¹, Broňa Brejová¹, Tomas Vinar¹

¹Comenius University in Bratislava, Slovakia

V. Boža <usama@ksp.sk>

B. Brejová <brejova@dcs.fmph.uniba.sk>

T. Vinar <vinar@fmph.uniba.sk>

We study the problem of efficient indexing of a collection of sequencing reads. In particular, our goal is to recover all reads in the collection related to a particular sequence. Several previous works proposed data structures for querying reads that contain a specific k -mer (usually, the upper bound on k has to be specified in advance). These works mostly concentrated on compressing the index (Philippe et al., 2011; Valimaki and Rivals, 2013) or indexing repetitive collections of reads (Siren, 2012; Claude et al., 2010).

In our work, we attempt to exploit the fact that the collection of reads is in fact the result of deep sequencing of a genome and as such, the reads themselves should align well to some underlying sequence that is unknown to us. We experimentally compare several strategies based on this assumption and show that they are more

efficient (in speed and memory) than above mentioned approaches to the problem.

This work is related to our new sequence assembler GAML (Boza et al., 2014). GAML attempts to find the sequence assembly that would maximize the probability of seeing a particular collection of sequencing reads given the sequence assembly. GAML uses an iterative optimization routine where the assembly with high likelihood is found by applying many local modifications that improve the likelihood. In each step, we need to update the likelihood by identifying reads that were affected by the updates applied in that particular step.

References

Boza, V., Brejova, B., and Vinar, T. (2014). GAML: Genome assembly by maximum likelihood. In *Algorithms in Bioinformatics (WABI)*, volume 8701 of *Lecture Notes in Bioinformatics*, page 122. Springer.

Claude, F., Farina, A., Martinez-Prieto, M. A., and Navarro, G. (2010). Compressed q-gram indexing for highly repetitive biological sequences. In *Bioinformatics and BioEngineering (BIBE)*, pages 86–91. IEEE.

Philippe, N., Salson, M., Lecroq, T., Leonard, M., Commes, T., and Rivals, E. (2011). Querying large read collections in main memory: a versatile data structure. *BMC Bioinformatics*, 12(1):242.

Siren, J. (2012). Compressed full-text indexes for highly repetitive collections. PhD thesis, University of Helsinki.

Valimaki, N. and Rivals, E. (2013). Scalable and versatile k-mer indexing for high-throughput sequencing data. In *Bioinformatics Research and Applications (ISBRA)*, volume 7875 of *Lecture Notes in Computer Science*, pages 237–248. Springer.

P272**The relationship between duplication and essentiality for mammalian genes**Mitra Kabir¹, Andrew J. Doig¹, Kathryn E. Hentges¹¹University of Manchester, United Kingdom

M. Kabir <mitra.kabir@postgrad.manchester.ac.uk>

A.J. Doig <andrew.doig@manchester.ac.uk>

K.E. Hentges

<kathryn.hentges@manchester.ac.uk>

Essential genes perform crucial tasks for the survival and development of the individual. Non-essential genes may be useful but not critical. We investigated a number of *Mus musculus* genes to study the role of gene duplication on mammalian gene essentiality, where an essential gene is defined to be required for survival to 3 days post birth. Gene duplication is a frequent event in multicellular eukaryotes which potentially generates new genes and/or functions. We labelled mouse essential and non-essential genes as defined by single gene knockout experiments as singletons or duplicates. Previous studies in mouse reported that singletons and duplicates have similar likelihoods of being essential. In contrast, our analysis showed that essential genes originating from duplicates are considerably lower in proportion than those originating from singletons. Also, at all developmental stages a significantly higher proportion of singletons and essential genes are expressed than duplicates and non-essential genes. We observed that essential genes are more ancient than non-essential genes. Furthermore, expressed singletons have an older phyletic age than duplicates expressed at the same stage of development. We found that duplicates with similar patterns of developmental co-expression are more likely to be non-essential; essential genes did not have such a trend. A greater proportion of genes duplicated by SSD were found to be essential than those duplicated by WGD. Overall, our findings suggest that duplicate genes in mouse are less likely to be essential than singletons. This study reveals new insights into the relationship of gene essentiality, developmental expression, and gene duplication.

P273**Assessment of the resolution of the UNRES force field with structure-based restraints.**Agnieszka Karczyńska¹, Paweł Krupa¹, Magdalena Mozolewska¹, Adam Liwo¹¹University of Gdańsk, Poland

A. Karczyńska

<agnieszka.karczynska@phdstud.ug.edu.pl>

UNRES (from UNited RESidue) is a physics-based coarse-grained force field [1], which is used for physics-based prediction of protein structure combined with the Multiplexed Replica Exchange Molecular Dynamics (MREMD) [2] algorithm. To enhance the predictive power of the method, we recently started to develop an approach in which UNRES simulations are carried out subject to the restraints on the $C\alpha...C\alpha$ distances and the $C\alpha...C\alpha...C\alpha...C\alpha$ backbone virtual-bond dihedral angles taken from template-based models of the proteins whose structure is to be predicted. In particular, UNRES is able to reorganize incorrectly packed domains in template-based models. In this work, we have assessed the asymptotic accuracy of this approach by running UNRES/MD and UNRES/MREMD simulations with restraints taken from the actual experimental structures. We used 12 proteins with different size and type of secondary structure (α , β and $\alpha + \beta$). All calculations were started from fully extended structures. The average $C\alpha$ root-mean square deviation was 4.51 Å, which suggests that improvement of the force field should be directed at enhancing the local-interaction potentials.

Supported by grant DEC-2013/10/M/ST4/00640 from the National Science Center of Poland and grant Mistrz 7./2013 from the Foundation for Polish Science. Computational resources were provided by (a) the supercomputer resources at the Informatics Center of the Metropolitan Academic Network (ICMAN) in Gdańsk, (b) the 952-processor Beowulf cluster at the Baker Laboratory of Chemistry, Cornell University, and (c) our 488-processor Beowulf cluster at the Faculty of Chemistry, University of Gdańsk.

[1] Liwo A, Czaplewski C, Ołdziej S, Rojas AV, Kaźmierkiewicz R, Makowski M, Murarka RK, Scheraga, HA. Simulation of protein structure and

dynamics with the coarse-grained UNRES force field. *Coarse-Graining of Condensed Phase and Biomolecular Systems.*, ed. G. Voth, Taylor & Francis, 2008, Chapter 8, pp. 107-122.

[2] Rhee YM, Pande VS. Multiplexed-Replica Exchange Molecular Dynamics Method for Protein Folding Simulation. *Biophysical Journal*, 2003, Volume 84, Issue 2, pp. 775-786.

P274

Predicting the functional impact of genetic sequence variation in human disease

Mark Rogers¹, Hashem Shihab¹, Julian Gough¹, Matthew Mort², David Cooper², Ian Day¹, Tom Gaunt¹, Colin Campbell¹

¹University of Bristol, United Kingdom, ²University of Cardiff, United Kingdom

M. Rogers <Mark.Rogers@bristol.ac.uk>

C. Campbell <C.Campbell@bristol.ac.uk>

We present a new approach for predicting whether single nucleotide variants in the human genome are functional in disease. An integrative binary class predictor was developed. The algorithm weights component data-types according to their relative informativeness and assigns a confidence measure to the predicted class label. This approach was evaluated on pathogenic germ-line mutations from the Human Gene Mutation Database, as positives, with negatives derived from the 1000 Genomes Project. The majority of described disease-associated sequence variants lie in non-coding regions and only two other recently proposed predictors exist for these variants. For prediction in non-coding regions, our method uses 4 constituent sources of data. It significantly outperforms both these competing methods, and it is currently state-of-the-art. The most highly weighted source of data was sequence conservation across species: a variant in a highly conserved region has a higher probability of disease association relative to variants in regions of high variability. Restricting to examples with the confidence measure greater than 90%, the classifier achieves a 96% test accuracy while making predictions on nearly 40% of examples. At 95% confidence cutoff, the accuracy increases to 98%, with nearly 16% of examples having label predictions. For prediction in coding regions, our method used 10 sources of data. In this context, we found that

our predictor was as accurate as the top competing algorithms. We have also devised disease-specific predictors and a web-browser for visualising the results from these studies.

P275

Co-evolution and co-expression based analysis and prediction of G Protein-coupled receptor heterodimerization

Bence Szalai^{1,2}, Susanne Prokop¹, Miklós Cserző^{1,2}, Péter Várnai¹, László Hunyady^{1,2}

¹Semmelweis University, Department of Physiology, Hungary, ²MTA-SE Laboratory of Molecular Physiology, Hungary

B. Szalai <bence.szalai@eok.sote.hu>

Heterodimerization of different G Protein-coupled receptors (GPCRs) can remarkably influence the physiological and pharmacological functions of these receptors. Heterodimerization as a protein-protein interaction can lead to co-evolution of the interacting GPCRs, which phenomenon has not been studied previously. In this study we have investigated the co-evolution of 231 rhodopsin like GPCRs from 59 vertebrate species with the Mirrortree method, based on the similarity of phylogenetic trees. We found that known GPCR heterodimers show significantly higher co-evolution scores than non-interacting receptors. We also investigated the co-expression of rhodopsin like GPCRs in different mouse tissues from a public dataset and found that GPCRs forming heterodimers show higher level of co-expression than non-interacting ones. Based on these co-evolution and co-expression scores we predicted the heterodimerization of several GPCRs, and we also experimentally verified some of these new predictions.

P276

Identification of cis-regulatory elements from chromosome conformation capture data

Robert Schöpflin¹, Martin Vingron¹

¹Max-Planck-Institut für molekulare Genetik, Germany
R. Schöpflin <schoepfl@molgen.mpg.de>

M. Vingron <vingron@molgen.mpg.de>

Gene regulation in eukaryotes involves complex interactions between promoters and regulatory elements that can be located thousands of base pairs away from each other. By the three-dimensional folding of chromatin these distal elements can be brought into spatial proximity to modulate gene expression. However, the specific contacting genomic elements, which are often cell- and stage specific, are widely unknown. The advent of chromosome conformation capture (3C) methodologies in combination with high-throughput sequencing started to shed light on the three-dimensional structure of chromatin in the nucleus. Additionally, an increasing amount of information about chromatin states, the location of enhancer and boundary elements, as well as transcription factor binding sites becomes available by chromatin immunoprecipitation sequencing (ChIP-seq) experiments. Here, we use three-dimensional chromatin contact frequencies from 3C experiments in combination with ChIP-seq data to identify functional chromatin interactions between promoters and their cis-regulatory elements.

P277

Efficient experimental design for systems biology dynamical models

Karol Nieniałowski¹, Michał Komorowski¹

¹Institute of Fundamental Technological Research Polish Academy of Science, Poland

K. Nieniałowski <k.nienaltowski@sysbiosig.org>

Dynamical models in quantitative biology are characterised by much more complex structures and substantially larger sets of parameters than models used in physics and engineering. Moreover available experiments usually provide limited set of data, usually corresponding only to fragments of studied systems. In consequence the reliability of used models is limited and our knowledge of studied processes misrepresented.

Potential remedy is provided by experimental design techniques. Such tools aim at selection of optimal experimental setup to maximise missing information about model parameters or model structure. As models in quantitative biology differ qualitatively from conventional models, experimental design tools need to be adapted to their specificity.

Here we present a method specifically tailored

for multi-parameter models of quantitative biology. Our framework enables to determine which parameters of a given model can be identified in a given experiment and predict which experiment should be performed next to maximise the number of identifiable parameters.

Our tool is different from methods developed so far as it is focused in verifying identifiability of individual parameters in large dynamical models, which contain even hundreds of parameters.

We present applicability of our tool analysing models of NF- κ B and MAPK signalling. Surprisingly, we show that certain parameters are virtually impossible to estimate in intuitively designed experiments. Our method helps to guide experimental design in order to render such parameters identifiable.

P278

The potential role of transposable elements as evolutionary helpers in sexual populations - the computational model.

Krzysztof Gogolewski¹, Michał Startek¹, Dariusz Grzebelus², Arnaud Le Rouzic³, Anna Gambin¹

¹Institute of Informatics, University of Warsaw, Poland,

²Institute of Plant Biology and Biotechnology, University of Agriculture in Krakow, Poland, ³Laboratoire Evolution, Génomes et Spéciation, Centre national de la recherche scientifique, France

K. Gogolewski <k.gogolewski@mimuw.edu.pl>

A. Gambin <aniag@mimuw.edu.pl>

Transposable elements (TE) may have a significant influence on speciation and evolution of species. Thus, understanding their behavior and dynamics seems crucial to deepening our knowledge on the evolutionary mechanisms.

The problem of transposon proliferation has been extensively studied from a theoretical point of view, and many models have already been proposed. However, most of them are based on the concept of transposition-selection equilibrium (TSE) that is, the transposon counts within modeled populations are controlled by the struggle between transposons' selfish drive to duplicate regardless of the effect on their host (transposition), and the evolutionary drive to eliminate hosts with high transposon

counts (selection).

In our approach, we analyze through stochastic simulations a model of TEs dynamics in sexual populations that accounts for environmental pressure on host populations. Our computational, stochastic model dealt away with the concept of TSE, and instead it tracked the organisms' phenotype (which were modified by transposition-induced mutations) to investigate TEs dynamics. The first results of this approach gives us the idea that the proliferation of TEs can have positive effect on both adaptation and survivability of organisms. Namely, we suggest the hypothesis that in sexual populations transposons increase their activity when their host is facing rapid environmental change. Moreover, in case of the constant, directed environmental shift the mean copy of TEs in the whole population is roughly constant and decreases when the shift stops. These observations may imply that an answer to the question if transposable elements can be considered as evolutionary helpers is positive.

P279

An entropy model for ranking structures and measuring flexibility in 3D RNA simulations using SimRNA

Wayne Dawson¹, Michal Boniecki¹, Janusz Bujnicki¹

¹Laboratory of Bioinformatics and Protein Engineering, Poland

W. Dawson <wdawson@genesilico.pl>

SimRNA is a recently developed de novo 3D structure prediction program that uses the Monte Carlo method to search the conformation space of RNA using knowledge-based potentials [1]. In developing the 3D model, we have also been exploring the larger physical question of what generates differences in Kuhn length (effectively the inverse of flexibility) in RNA. For illustration, a typical sequence of folded single-stranded RNA has a Kuhn length ranging 4-12 nts, yet that same sequence with its complement (double-stranded RNA) has a Kuhn length on the order of 200 bps. This was shown to be a consequence of freezing out the number of degrees of freedom in the RNA chain [2]. Our 3D studies using SimRNA with its 3D statistical potentials further support the conclusion that the Kuhn length is a function of the Young's modulus

of a stem. This, combined with our entropy modeling [3], is further used to re-rank 3D RNA folding trajectories from SimRNA, where the re-rankings roughly correspond to the first cluster.

References:

- [1] Boniecki et al. SimRNA: a coarse-grained method for RNA folding simulations and 3D structure prediction. (submitted) <http://genesilico.pl/software/stand-alone/simrna>
- [2] Dawson et al. (2012) A new entropy model for RNA: part II. Persistence-related entropic contributions to RNA secondary structure free energy calculations. *Journal of Nucleic Acids Investigation* 3:e2
- [3] Dawson et al. (2014) A new entropy model for RNA: part III. Is the folding free energy landscape of RNA funnel shaped? *Journal of Nucleic Acids Investigation* 5:2652.

P280

Inferring disease mechanisms from multiple gene expression datasets using PPI

Sahar Ansari¹, Michele Donato¹, Sorin Draghici¹

¹Wayne State University, United States

S. Ansari <saharansari@wayne.edu>

Understanding the mechanisms that cause changes in a phenotype requires the identification of the genes that are disrupted in that phenotype, and the relationships between them.

The networks that explain the interactions between genes can be used to predict the mechanism of action of a drug or the responses of the system to a specific impact (e.g. a disease).

The currently available methods fail to discover the condition-specific relationships between genes with high accuracy. Many existing methods find gene regulatory networks without focusing on one specific phenotype. These networks are not precise, because genes interact with each other differently in different conditions.

We use the union of the differentially expressed (DE) genes from each dataset as a unique list of DE genes. We build a "neighbor" network for each gene with the edges from this gene and all genes immediately downstream of it in the PPI network. In the next step, we calculate the enrichment of each neighbor network based on the number of DE genes

they contain. We applied this approach on multiple datasets that come from experiments studying type II diabetes, lung cancer, and colorectal cancer. We assessed the result by comparing the constructed network with the pathways that are associated to these diseases in the KEGG database. The results show that the proposed mechanism includes genes known to have important functions in understudying conditions. Also, many of the interactions included in the putative mechanism are present in pathways that are known to be associated to them.

P281

Discovering interdependencies in genomic data and the case of immune system genes

Michal J. Dabrowski¹, Michal Draminski¹, Klev Diamanti², Jacek Koronacki¹, Jan Komorowski^{1,2}

¹Institute of Computer Science, Polish Academy of Sciences, Poland, ²Department of Cell and Molecular Biology, Uppsala University, Sweden

M.J. Dabrowski <m.dabrowski@ipipan.waw.pl>

Computational methods used today in analyzing genomic data offer discovery of single variables, considered one at a time, under the assumption that they would individually characterize such systems. This is in an obvious disagreement with the actual knowledge of living systems in which very large numbers of variables (features), interact in creating non-linear complexity. However, it is only now that the availability of very large data sets in Life Sciences provided by various sequencing technologies has made it possible to develop and apply new methods specifically focused on discovering interdependencies between the features and their values.

Our approach is founded on classification-based feature selection and rule-based modeling. Interactions among features are modeled as a network of interdependencies, extracted from the classification trees, constructed by the MCFS-ID algorithm. The network is given in the form of directed graph. In addition, on a final level, classification rule sets are provided by the ROSETTA system together with Ciruvis.

This approach has been successfully used in a number of applications. Here we show how it helps discover interdependencies in the human immune

system responses to various stimuli of CD4+ T-cells depending on the racial background. Specifically, we learn that gene-responses related to bacteria characterize Afro-Americans, responses related to viruses – Caucasians, and both classes of responses characterize Asians. Furthermore, the refinement to the level of rules showed that the distribution of attribute values (low, medium, high) across the classes suggests that African-Americans and Asians are much more homogeneous than Caucasians, at least for these genes.

P282

Towards discovering key factors in prediction of post-translational modifications

Marcin Tatjewski^{1,2}, Julian Zubek^{1,2}, Marcin Kierczak³, Dariusz Plewczynski²

¹Institute of Computer Science, Polish Academy of Sciences, Poland, ²Centre of New Technologies, University of Warsaw, Poland, ³Science for Life Laboratory, Department of Medical Biochemistry and Microbiology, Uppsala University, Sweden

M. Tatjewski <marcin.tatjewski@gmail.com>

After several years of research, prediction of Post-Translational Modifications (PTM) is still a subject of intensive studies. In spite of these efforts, a number of relatively fundamental questions regarding both predicting and understanding PTM sites remains unanswered. What is the optimal size of an amino acid sequence window to be used for predicting PTM sites? How many features/residues can be used without over-parametrizing the model? Our goal is to address these and similar questions. To this end, based on UniProtKB database, we compiled 275 datasets consisting of positive and negative examples of PTM sites. Each dataset for one type of PTM. Our datasets cover a broad range of modification types, ranging from different types of phosphorylations (differentiated also by kinases) to different types of acetylation or methylation. Using the datasets, we performed several statistical tests on results obtained using previously developed approaches, mainly Auto Motif Servers. This let us draw several interesting conclusions, e.g. increase in sequence window size from 9 to 21 can improve AUC results by 0.02 at maximum, even complex feature extraction based on physicochemical properties achieves at best 0.07 higher AUC result than using raw amino acid sequence encod-

ing. Moreover, we attempted at bringing novel aspects to PTM prediction analysis by employing Monte-Carlo Feature Selection algorithm to investigate which particular sequence positions and which amino acid properties are crucial factors for determining PTM sites. We also clustered the final results in order to summarise the outcomes in a meaningful, easy-to-interpret way.

P283

Approaches for integrative clustering of cancer cell lines for evaluation of therapeutic response

Dennis Wang¹, Sara Dempster¹, Nirmal Keshava¹

¹AstraZeneca, United Kingdom

D. Wang <dennisyq.wang@astrazeneca.com>

New targeted cancer therapies are often specific to a particular type of tumour. Methods, like Similarity Network Fusion, iCluster, and Consensus Clustering, can be used to conduct unsupervised clustering or classification of tumours based on their genomic profiles. Until now, the performance of these methods for finding biologically relevant tumour subgroups had only been evaluated in terms of differentiating patient overall survival. We show that integrative clustering can be used in an unsupervised manner to identify drug-sensitive groups of tumours from the Sanger GDSC cell line resource. Often it is unclear how genes that define the different subgroups are selected by the methods. In order to extract genomic biomarkers for the tumour subgroups that are biologically meaningful, we identified pathways or gene networks affected by the biomarkers through algorithms using diffusion processes. This network based stratification approach finds SNVs or CNVs with weak association to drug response that are connected in the same pathways. We subsequently use these mutational subnetworks as a new feature set to examine association with drug response. This proof of concept aims to show how existing integrative clustering approaches can be adapted to help predict differential response to cancer therapies and identify novel biomarkers.

P284

Signatures of positive selection reveal driver genes across multiple cancer types

Luis Zapata^{1,2}, Hana Susak¹, Oliver Drechsel^{1,2}, Stephan Ossowski¹

¹Centre for Genomic Regulation (CRG), Spain, ²UPF, Spain

S. Ossowski <stephan.ossowski@crgeu>

We present a pan-cancer analysis of the clonal structure of genes participating in tumor evolution. We describe a novel Bayesian approach, cDriver, to accurately rank driver genes in different tumor types, simultaneously using ka/ks estimates, damage score, ploidy, tumor purity and the cancer cell fraction (CCF) of mutations within the tumor cell admixture. We prove that our ranking captures best the landscape of known driver genes in most of the tumors studied. Moreover, we classify genes, based on their CCF profile across multiple patients, into early or late driver genes giving us an insight on the evolutionary history of each tumor. We quantify tumor heterogeneity revealing that blood cancers are the less heterogeneous in terms of genes contributing to tumor development. We also differentiate between cancer-specific and cancer common genes giving us an insight on how each cancer is modulated given an initial set of mutations.

P285

Computational analysis of sequence motifs for discriminating different ChIP-exo profiles of related RbF proteins

Shamba Mondal¹, Yiliang Wei², David Arnosti², Bartek Wilczynski³

¹Nencki Institute of Experimental Biology, Poland,

²Dept of Biochemistry and Molecular Biology, Michigan State University, United States, ³Institute of Informatics, University of Warsaw, Poland

S. Mondal <samiit.hyd@gmail.com>

B. Wilczynski <bartek@mimuw.edu.pl>

Rbf1 and Rbf2 retinoblastoma corepressor proteins in Drosophila do not directly bind to DNA, but bind in a complex with transcription factors. Rbf2 is a recently evolved Retinoblastoma family member in Drosophila, differing from Rbf1 especially in the C-terminus. From the ChIP-seq data of Rbf1 and exo-ChIP-seq data of both Rbf1 and Rbf2, it

was found that Rbf2 targets approximately twice as many genes as Rbf1. We carried out bioinformatic analysis to investigate the basis for differential targeting by these two proteins, first for the following four functional groups- Cytoplasmic Ribosomal Protein (CRP) genes, Mitochondrial Ribosomal Protein (MRP) genes, Cell Cycle Genes (CCG) and Signalling Pathway Genes (SPG); and then for the whole genome.

By grouping the genes containing a ChIP enriched region as determined by MACS into 4 exclusive classes, depending on the presence of MACS peaks for Rbf1 only (208 genes), Rbf2 only (2275), both (1112) or none of the factors (12234), we tested for motif association with ChIP enrichment. STAP program [1] was used to test which TFBS affinity scores correlate with ChIP enrichment for the DNA sequences upstream of the TSS, first for analyzing individual motifs from a list of 127 motifs, and then for testing co-operative interactions between two TFs with motifs selected from the individual runs based on Pearson correlation between predicted binding and observed binding.

Individual and pair-wise motif enrichment analysis were done with MAST [2] for all the 4 exclusive classes each of the above-mentioned functional groups. Suggested by both of our approaches, Beaf-32 was the most promising candidate TF, that seemed to be involved in helping Rbf2 bind DNA. Although a Beaf-32 RNAi experiment showed that this TF was not required for Rbf2 binding, it came up in a parallel study [3] that the dREAM complex functions not only as a repressor, but also appears to recruit insulator proteins like Beaf-32 to block enhancer activity on divergently transcribed genes. Motif strength analysis identified that Rbf2-specific promoters have different preferred motif affinities for multiple factors, suggesting unique targeting mechanisms based on cooperativity of multiple weakly bound TFs as opposed to strong binding of a few TFs in Rbf1 bound promoters.

References:

1. Xin He, Chieh-Chun Chen, Feng Hong, Fang Fang, Saurabh Sinha, Huck-Hui Ng, Sheng Zhong, 2009 A Biophysical Model for Analysis of Transcription Factor Interaction and Binding Site Arrangement from Genome-Wide Binding Data. *PLoS ONE* 01/2009; 4(12):e8155.
2. Timothy L. Bailey and Michael Gribskov, 1998 Combining evidence using p-values: Application to sequence homology searches. *Bioinformatics*,

14(1):48-54.

3. Korenjak, M., E. Kwon, R. T. Morris, E. Andersen, A. Amzallag et al., 2014 dREAM co-operates with insulator-binding proteins and regulates expression at divergently paired genes. *Nucleic Acids Res.* 42: 8939–8953.

P286

Q - A saturation-based peak caller for reproducible analysis of transcription factor binding sites

Peter Hansen¹, Jochen Hecht², Daniel Ibrahim³, Alexander Krannich⁴, Peter N. Robinson¹

¹Charité-Universitätsmedizin Berlin, Germany, ²Berlin Brandenburg Center for Regenerative Therapies (BCRT), Germany, ³Max Planck Institute for Molecular Genetics, Germany, ⁴Berlin Institute of Health, Germany

P. Hansen <peter.hansen@charite.de>

P.N. Robinson <peter.robinson@charite.de>

Q is a fast saturation-based ChIP-seq peak caller that works well in conjunction with the irreproducible discovery rate (IDR) procedure. Q was extensively tested on publicly available datasets from ENCODE and shown to perform well with respect to reproducibility, consistency regarding predicted transcription factor binding motifs, as well as overall run time. Q is implemented in C++ making use of the SeqAn library. There are a number of useful features for the primary analysis of ChIP-seq data. Q can be run with or without data from a control experiment. Duplicate reads are removed on the fly without altering the original BAM file. The average fragment length of the sequencing library, an essential parameter for peak calling and downstream analysis, is estimated automatically from the data.

If performed after removal of duplicates, our procedure yields an equivalent estimation of the average fragment length as the cross-correlation plot of SPP but is approximately three times faster. As a part of this, Q also calculates the relative strand cross-correlation coefficient (RSC), which allows a global quality assessment of the enrichment. In addition Q offers its own quality metrics, which can be used for trouble-shooting and quality control of the results. If desired, Q also generates fragment coverage profiles which can be uploaded to UCSC's genome browser, where they can be displayed in the context

of other related data such as for example ChIP-seq data for histone modifications and cofactors or expression data.

P287

Combining biological networks with ensemble learning for analyzing molecular profiles

Xi Gao¹, Tomasz Arodz¹

¹Virginia Commonwealth University, United States

T. Arodz <tarodz@vcu.edu>

Machine learning methods are often used to help understand complex patterns in molecular profiles obtained using array- or sequencing-based high-throughput technologies. However, the exploration can easily get lost in noisy, high-dimensional data. Our work is aimed at developing machine learning methods that take molecular profiles from different phenotypes as input, and train classification models that serve two goals: prediction of phenotypes for previously unseen samples, as well as discovery of the underlying networks that show phenotypic differences. We propose an algorithm that extends AdaBoost ensemble classifier by incorporating existing knowledge about biological networks into model training, in order to improve classifier accuracy and interpretability. We tested the proposed method on three datasets simulated using GeneNetWeaver tool and compared the results with state-of-art classification methods. Results show that the proposed method is able to produce accurate, sparse classifiers, and can help discover true sources of differences between phenotypes.

P288

Identifying strand-symmetry reversal in the human genome

Michelle Flowers¹, Kristin Helling¹, Jing Zhang¹, John Karro¹

¹Miami University, United States

J. Karro <karroje@miamiOH.edu>

In Mugal et al. [1] the authors proposed a computational method for identifying genes under strand asymmetric substitution pressure: genes in which the substitution rates of the evolutionarily unconstrained bases differ between the 5' and 3' strands. This is a trait that appears to be linked to both the effects of transcription-coupled repair (TCR) and

proximity to genomic origins of replication. However, in that study there was an assumption that any given gene had an associated fixed rate-ratio matrix associated with it (e.g. assuming the A→C substitution rate for a given gene has changed in proportion to that of the A→G substitution rate, to that of C→G substitution rate, etc...). From this it follows that a gene's symmetry type must remain fixed over time: a gene subjected to symmetric substitution pressures over the last 1 Mya must have been subject to this pressure over the last 100 Mya.

We relax the assumption of the fixed rate-ratio matrix to identify genes that have experienced a change in symmetry pressure. By applying a modification of Mugal's method to data sets derived from transposable elements of different age categories, we have identified 188 human genes that appear to have lost their asymmetry status, and 111 human genes that appear to have gained it, since the mammalian radiation. This is presumably due to changes in either the TCR process or the reorganization of the replication origins for reasons uncertain to us at this time.

[1] Mugal, C. F., Grünberg, von, H. H., & Peifer, M. (2008). Transcription-Induced Mutational Strand Bias and Its Effect on Substitution Rates in Human Genes. *Molecular Biology and Evolution*, 26(1), 131–142. doi:10.1093/molbev/msn245

P289

Non-random distribution of recombination hotspot motifs in human genome

Dorota Mackiewicz¹, Stanisław Cebrat¹

¹Department of Genomics, Biotechnology Faculty, University of Wrocław, Poland

D. Mackiewicz <dorota@smorfland.uni.wroc.pl>

Recombination is the main cause of genetic diversity. This phenomenon plays a big role in generating new combinations of alleles, which greatly increases adaptive potential of genomes and promotes efficient selection against deleterious mutations. Thus, errors in the recombination process can lead to chromosomal abnormalities, genome instability or other forms of genomic disorders. Recombination events are confined to narrow chromosome regions called hotspots in which characteristic DNA motifs

are found. Genomic analyses have shown that both recombination hotspots and DNA motifs are distributed unevenly along human chromosomes and are much more frequent in the subtelomeric regions of chromosomes than in their central parts. Detailed analyses of the distribution of DNA motifs presumably related to homologous recombination in the human genome showed that clusters of motifs roughly follow the distribution of recombination hotspots, whereas single motifs show a negative correlation with the hotspot distribution. The consequence of this could be that clusters of DNA motifs rather than single motifs are involved in the positioning of recombination hotspots. This solution can also explain non-conserved hotspot locations in human and chimpanzee genomes and problems with finding homologous recombination DNA motifs for all hotspots in both genomes. The participation in the conference was funded by the KNOW Consortium.

P290

Computational modeling of the effect of amino acid substitutions on the formation of Ubiquitin-based complexes

Mina Maleki¹, Mohammad Haj Dezfulian², Luis Rueda¹

¹University of Windsor, Canada, ²Harvard Medical School, United States

M. Maleki <maleki@uwindsor.ca>

Ubiquitin (Ub), a small albeit evolutionary conserved regulatory protein, has been implicated in many cell processes including cell cycle regulation, DNA repair mechanisms, and gene expression. Due to the critical roles played by ubiquitination in maintaining the homeostatic balance of these cell processes, its deregulation and mutations appear to be a fundamental characteristic in different types of diseases such as breast and ovarian cancers [1]. Therefore, much attention has been focused on understanding and deciphering the role of ubiquitination in an attempt to harness the Ub machinery as potential targets for development of inhibitors. Due to paucity of small molecule inhibitors, much attention has been focused on utilization of ubiquitin scaffolds as a means of blocking the activity of various components of the Ub machinery. While these approaches have been highly successful in development of highly potent inhibitors and activators

of these complexes, they require specialized instruments and are rather experimentally demanding and expensive. In contrast, computational analysis and unsupervised machine learning techniques allow researchers to employ more cost effective methods for this purpose.

We propose a computational method used to discover the least number of amino acid changes required to develop stable Ub complexes. For this, after utilizing a library of ubiquitin variants, FoldX [2] is used to provide a reliable and quantitative estimation of binding free energies and in consequence estimate the stability of protein complexes based of their sequence and structural information. Then, the protein variants with substitution mutations that improve affinity of interactions with less number of mutations are found. Finally, we have verified that the mutated sequence of a protein can be folded properly within a cell. Our results on the dataset used in [3] corroborate that the proposed method not only can find the list of mutations reported in [3], but also can find more mutations with higher stability score.

1. K. Haglund, I. Dikic, "Ubiquitin Signaling and Cancer Pathogenesis," from book titled Protein Degradation: The Ubiquitin-Proteasome System and Disease. Edited by R. J. Mayer, A. Ciechanover, M. Rechsteiner, John Wiley & Sons, 2008.
2. J. Schymkowitz, J. Borg, F. Stricher et al. "The FoldX web server: an online force field," *Nucleic Acids Research*, 2005, vol 33, pW382-8.
3. A. Ernst, G. Avvakumov, J. Tong et al. "A strategy for modulation of enzymes in the ubiquitin system." *Science*. 2013, 339(6119):590-5

P291**A maximum-likelihood approach to force-field training for protein structure prediction and folding simulations**

Bartłomiej Zaborowski¹, Dawid Jagieła¹, Adam Sieradzan¹, Cezary Czaplewski¹, Anna Hałabis², Agnieszka Lewandowska², Wioletta Żmudzińska², Stanisław Oldziej², Adam Liwo³

¹Faculty of Chemistry, University of Gdańsk, Poland,

²Intercollegiate Faculty of Biotechnology, University of Gdańsk and Medical University of Gdańsk, Poland,

³Faculty of Chemistry, University of Gdansk, Poland

A. Liwo <adam.liwo@gmail.com>

A long-sought goal of computational biology is obtaining a force field capable of locating the native protein structures and folding pathways. A common approach is use of a number of training proteins with known native structures and to optimize force-field parameters to locate the native-like structures in the ensemble with the lowest free energy. Various target functions have been used, including the energy gap or Z-score between the native and non-native structures or funnel sculpting. However, all these approaches overemphasize on the conformations which are closest to the native structures, which makes the procedure very unstable unless the set of conformations is fixed (fold recognition). In this work we propose an alternative approach, based on applying the maximum-likelihood principle and use of training-protein conformations determined experimentally at various temperature. Each experimental conformation is an experimental point and the theoretical probability-density function is represented by a sum of Gaussians centred at decoys from the corresponding ensembles generated by simulations. The method was applied to the physics-based coarse-grained UNRES force field developed in our laboratory; it exhibits outstanding numerical stability compared to earlier approaches. The optimized force field has very good predictive power.

Supported by grants DEC-2013/10/M/ST4/00640 from the National Science Center of Poland and Mistrz 7./2013 from the Foundation for Polish Science. Computer resources were provided by: (a) the National Science Foundation (<http://www.nics.tennessee.edu/>), (b) the Aca-

demic Computer Center in Gdańsk (CI TASK), (c) the Interdisciplinary Center of Mathematical and Computer Modeling in Warsaw (ICM).

P292**Analysis of Novel mRNA Transcripts in Prostate Cancer**

Abed Alkhateeb¹, Siva Reddy¹, Iman Rezaeian¹, Urvashi Katiyar¹, Dina Maskoni¹, John Kelly¹, Dora Cavallo-Medved¹, Lisa Porter¹, Luis Rueda¹

¹University of Windsor, Canada

L. Rueda <lrueda@uwindsor.ca>

Detecting different ways of splicing of RNA transcripts at early stages of prostate cancer may reveal novel mechanisms driving disease progression. A novel transcript is a transcript that shares at least one exon with a known transcript in the refSeq, while differing in at least one exon with that transcript. These transcripts are potentially valuable prognostic biomarkers. The factors driving disease progression within each patient population are compounded by variances in genetics and environment. Here, we analyze RNA-seq data gathered from prostate cancer patients from a Chinese population.

Using eight of 28 paired samples (matched tumor and adjacent normal) we aligned mRNA reads to the Human Genome hg19 refSeq, and then constructed a database of all expressed transcripts. By tracking changes in the relative abundance of transcripts with a common transcription start site, we identified changes in splicing for each stage/sub-stage. We then found differentially expressed transcripts in correlation with various prostate cancer stages/sub-stages by estimating abundance of transcripts corresponding to each of these stages.

Our study isolated 4 novel transcripts expressed from the ABCC4, STEAP2, IP6K2 and WWP2 genes. ABCC4, an androgen-regulated gene encoding an ATP binding cassette transporter, is involved in multidrug resistance. STEAP2, another trans-membrane transporter, has isoforms unique to prostate epithelia that are amplified 10-fold in malignant tumors. IP6K2 encodes a protein of the inositol phosphokinase family and suppresses apoptotic activities of cytokine interferon-beta. WWP2 encodes an ubiquitin ligase that regulates levels of PTEN, a well-established tumor suppressor that is lost in approximately 70% of prostate cancers.

The genes all have protein isoforms overexpressed in prostate cancer. Further analysis is warranted to determine the potential of using these transcripts as prognostic markers. It is also important to determine whether targeting the mechanism of their downstream protein products may represent valuable therapeutics to prevent or treat late-stage aggressive prostate cancer.

P293

How do loops form in proteins? A key to protein folding?

Adam Sieradzan¹, Nevena Litova², Adam Liwo³, Antti Niemi^{4,5,6}

¹University of Gdansk, Technical University of Gdansk, Poland, ²Institute for Nuclear Research and Nuclear Energy - Bulgarian Academy of Sciences, Bulgaria, ³University of Gdansk, Poland, ⁴Uppsala University, Sweden, ⁵Beijing Institute of Technology, China, ⁶Université de Tours, France

A. Sieradzan <adams86@wp.pl>

A. Niemi <antti.niemi@physics.uu.se>

The protein folding problem has been studied for decades. Three models of protein folding have been proposed: hydrophobic collapse, diffusion-collision and nucleation-condensation. In our earlier work we found that the correct location of loops is essential for proper protein folding. In this work, we analyze the way how the loop forms and suggest new mechanism of loop formation. As a model protein we used the core structure of gp41 from the HIV envelope glycoprotein (PDB CODE: 1AIK). We studied a loop from soliton-formation perspective and proposed a new mechanism of loop formation. According to this mechanism, the energy is transmitted along the chain until it can no longer be transmitted and becomes accumulated to form a kink.

Acknowledgements

This work was supported by the Foundation for Polish Science (FNP START 100.2014 and Mistrz7./2013), Polish National Science Center (DEC-2012/06/A/ST4/00376), Bulgarian Science Fund (BSF; grant DNTS-CN-01/9/2014), Vetenskaprådet, Carl Trygger Foundation, STINT, Svenska Institutet, and Qian Ren at BIT. Calculations were carried out with the resources at the Academic Computer Center in Gdańsk, CI TASK, and Interdisciplinary Center of Mathematical and Computer Modeling, University of Warsaw (ICM).

P294

Global sensitivity analysis based on knn-estimators for information theoretic measures

Agata Charzyńska¹, Anna Gambin²

¹Institute of Computer Science Polish Academy of Sciences, Poland, ²Institute of Informatics University of Warsaw, Poland

A. Charzyńska <a.charzynska@phd.ipipan.waw.pl>

A. Gambin <aniag@mimuw.edu.pl>

One of the most important element of modern systems biology is modelling of biochemical reaction networks. Mathematical formulation of biological phenomena enables quantitative and qualitative problem capture. Theoretical models usually involve numerous parameters, which are unknown or have only vaguely constrained values. Therefore reliable utilization of a model demands tools to extract relevant information even in the case when parameters are not entirely determined. Classical sensitivity analysis methods enable to quantify the impact of parameters on the model responses, even if parameters are not precisely determined but constrained to a broad range of values. The existing methods, however, are not well adapted to models with multi-variables output with many input parameters, that are typical for models of biochemical dynamics.

We have developed a sensitivity analysis method, based on information theoretic measure, to understand the joint impact of parameters on model dynamics. It seems that the mutual information between a parameters subgroup and the subgroup of output variables of interest can be successfully used as the sensitivity measure. For computation of mutual information we use the differential entropy estimator that employs k-nearest neighbors statistic. Moreover we define and compare the interaction indexes within subgroup of model parameters dependent on model output.

The example of analysis of the p53-Mdm2 model has shown that single sensitivity indexes are coherent with classical local sensitivity analysis, moreover the interactions between parameters captured the negative feedback loop of the model. The advantage of the proposed methodology is its relative simplicity and easy applicability for multi-parameters models.

P295**Computational method for modeling and testing of transcription factor binding sites**

Marcin Pacholczyk¹, Karolina Smolińska¹, Marek Kimmel²

¹Silesian University of Technology, Poland, ²Rice University, United States

M. Pacholczyk <marcin.pacholczyk@polsl.pl>

K. Smolińska <karolina.smolinska@o2.pl>

M. Kimmel <kimmel@rice.edu>

Introduction: Transcription factors are proteins which are able to bind with specific, short DNA sequences - transcription factor binding sites (TFBS). The TFBSs are traditionally modeled by Position Weight Matrices (PWMs) obtained either computationally or from experimental data.

Method: We propose a modification of (Alamanova et al. 2010) computational approach and use DDNA3 energy function to calculate interaction energy between protein and DNA. The algorithm uses crystal structure of TF-DNA complexes. Our data set contains dimers from NF-kB family: p50p50, p50p65 and p50RelB and HSF1 homodimer. We also present a method of improving the PWMs quality by fine tuning of parameters in DDNA3 potential. We selected human promoter sequences with experimentally verified NF-kB and HSF1 binding sites. Sequences were scanned with NucleoSeq 2.0 (Jaksik and Rzeszowska-Wolny 2012) TFBSs detecting algorithm.

Results and discussion: We selected PWMs based on the ROC analysis for each TF. To test presented technique we compared matrices constructed for our method, Alamanova et al. approach and from TRANSFAC database. Matrices detected similar number of experimentally confirmed binding sites.

Conclusion: The comparison shows significant similarity and comparable performance between calculated and experimental matrices (TRANSFAC). The proposed approach can be a promising alternative to experimental techniques of detecting TFBSs.

Alamanova, D., Stegmaier, P., Kel, A. (2010) *BMC Bioinformatics* 492(2), 375-381.

Jaksik, R., Rzeszowska-Wolny, J. (2012) *Gene* 11, 225.

P296**Tree consistent PBWT and their application to reconstructing Ancestral Recombination Graphs and demographic inference**

Vladimir Shchur¹, Richard Durbin¹

¹Wellcome Trust Sanger Institute, United Kingdom

V. Shchur <vlshchur@gmail.com>

The positional Burrows-Wheeler transform PBWT is a representation of a set of haplotypes that supports very efficient data compression and fast haplotype matching. We introduce a modification of PBWT which we call a tree consistent PBWT, or shortly tcPBWT, which has a natural and tractable connection with local coalescent trees. tcPBWT shows some improvement in the compression rate compared to PBWT, which suggests that it has better consistency with the genealogy giving rise to the haplotypes. Building on the natural tree properties of tcPBWT, we show how to use it to build ancestral recombination graphs of a set of haplotypes, which represent possible histories of the set of mutations and recombinations that could have given rise to the data. This process is close to linear in the size of the data set, potentially enabling very rapid genetic inference on large data sets. We will present applications to real and simulated datasets.

P297**Evaluation of de novo transcriptome assemblies from RNA-Seq data**

Nathanael Fillmore¹, Bo Li², Yongsheng Bai³, Mike Collins³, James Thompson³, Ron Stewart³, Colin Dewey¹

¹University of Wisconsin, Madison, United States,

²University of California, Berkeley, United States,

³Morgridge Institute for Research, United States

N. Fillmore <easychair@nate-fillmore.com>

De novo RNA-Seq assembly facilitates the study of transcriptomes for species without sequenced genomes. With the recent development of several de novo transcriptome assemblers, each of which has its own set of user-tunable parameters, there are many options for constructing an assembly. However, selecting the most accurate assembler and parameter settings for a given RNA-Seq data set has remained challenging, especially when the ground truth is unknown. To address this challenge, we

have developed DETONATE, a collection of methods for evaluating de novo transcriptome assemblies with or without a ground truth transcript set. DETONATE consists of two components: RSEM-EVAL, a model-based score for evaluating assemblies when the ground truth is unknown, and REF-EVAL, a refined set of ground-truth-based scores. Our experiments show that RSEM-EVAL correctly reflects assembly accuracy, as measured by REF-EVAL. The RSEM-EVAL score has a broad range of applications, including selecting the best assembler for a particular data set, optimizing parameter settings for an assembler, and guiding new assembler design. With the guidance of RSEM-EVAL, we assembled the transcriptome of the regenerating axolotl limb; this assembly compares favorably to a previously published axolotl assembly, identifying more expressed and differentially expressed genes, many of which are known to play a role in limb regeneration.

P298

MOCCA: Accurate identification of transcription factor binding sites from DNase-seq data

Aleksander Jankowski¹, Jerzy Tiuryn¹, Shyam Prabhakar²

¹University of Warsaw, Poland, ²Genome Institute of Singapore, Singapore

A. Jankowski <ajank@mimuw.edu.pl>

Computational identification of transcription factor (TF) binding sites in the genome, given a particular cell type and conditions, remains a challenging task. One of the approaches relies on proper identification of TF footprints, i.e. regions where TFs are bound to the DNA fragment and protect the DNA from degradation. These footprints can be accurately identified using DNase-seq experimental technique. Here, we present MOCCA, a novel computational method to identify individual TF binding sites from genome sequence information and cell-type-specific experimental data, such as DNase-seq. We combine the strengths of its predecessors, CENTIPEDE and Wellington, while keeping the number of free parameters in the model reasonably low. Our method is unique in allowing for multiple binding modes for a single TF, differing in their cut profile and overall number of DNase I cuts. We show that MOCCA consistently outperforms CENTIPEDE and Wellington across three sources of DNase-seq

data, by assessing the performance of these tools against ChIP-seq profiles. The difference was particularly significant when applied to binding site prediction for low-information-content motifs. Finally, we introduce Closed Chromatin Contribution (CCC) as an indirect measure of TF pioneer factor activity.

P299

Protein Contact Ontology - a tool for annotation of protein residue-residue contacts

Bogumil Konopka¹, Rafal Roszak¹, Malgorzata Kotulska¹

¹Wroclaw University of Technology, Poland

B. Konopka <bogumil.konopka@pwr.edu.pl>

Accuracy and reliability of state-of-the-art contact-site predictors are not good enough to allow contact-based protein structure prediction (Monastyrskyy et al., 2011). This may be caused by the fact that most methods treat all contacts as equal, based on a simple geometrical definition, despite that only some of the contacts are real interaction which truly impact the protein structure. In this work we present the Protein Contacts Ontology (PCO). The ontology provides a standardized vocabulary that allows to formally describe protein residue-residue contacts and their environments.

At the most general level the ontology has three distinct classes: i)'contact_attribute', ii)'residue_attribute', iii)'entity'. The part of the ontology related to 'contact_attribute' defines attributes/properties that can be used to describe protein contact sites e.g. the type of observed physico-chemical interaction. The part related to 'residue_attribute' includes terms that allow to describe amino acid residues. Finally, the terms grouped under 'entity' are used to model objects such as protein structural regions, amino acid residues or contact sites. Following the guidelines provided by OBO consortium (Open Biomedical Ontologies Consortium) (Smith et al., 2007), fragments of other ontologies were reused, where possible.

The presented PCO ontology allows precise annotation of available structural data with the focus on inter-residue contacts. Based on that, a detailed classification of inter-residue contacts can be performed. This will allow to decompose the contact site prediction problem into a set of simpler prob-

lems of predicting certain types of contacts. We hypothesize that this will improve the accuracy of predictions performed by machine learning methods.

P300

Algorithmic approach based on quaternion algebra to analyze causality of structural changes in biomolecular systems

Marcin Sobieraj¹, Marek Kalinowski², Bogdan Lesyng^{1,2}

¹University of Warsaw, Poland, ²Mossakowski Medical Research Center PAS, Poland

M. Sobieraj <msob@biogeo.uw.edu.pl>

B. Lesyng <lesyng@gmail.com>

Structure and function relationship is not straightforward. Although advanced Molecular Dynamics (MD) simulations provide information about intra- and intermolecular motions of complex (bio)molecular systems, which theoretically should provide information about motions of physiological importance, the huge amount of raw MD data without any detailed analysis do not have any significant cognitive value. It is typical problem which refers to functional, dynamical properties of complex systems.

In practical terms we want to answer the question of how changes in one part of the molecular system (cause) influences motions in another part with some time-delay (effect). Formalism detecting and analyzing such relations is called causal analysis (CA), and is widely used for example in econometrics to examine the cause-and-effect relations between multichannel scalar signals. The basic methodology of CA for scalar signals was developed by Clive Granger and Robert Engle, who received the Nobel Prize in 2003 for their research in economics.

An overview of a number of CA models for scalar signals drawn from MD simulations of (bio)molecular systems is given in [1]. One should stress, however, that data obtained from MD simulations, related to structural changes are described by time-dependent 3-dimensional vector signals. That's why we have developed a new algorithmic formalism enable to study causal relations between vector signals. It appears that time-dependent, vector correlation analyses can be

formulated using quaternions algebra. Two quaternion-based algorithmic approaches were designed: the first one applies the Granger causality approach, and the second one utilizes a Directed Transfer Function (DTF) formalism which allows to study causality relations in the frequency domain. Applications of these methods were tested and validated based on MD simulations for a model molecular system located in different environments. Possible applications for complex (bio)molecular structures are indicated.

[1] P. Daniluk et al, From Experimental, Structural Probability Distributions to the Theoretical Causality Analysis of Molecular Changes, Computer Assisted Methods in Engineering and Science, 19,1-20, 2012.

Acknowledgements: This study were supported by BST funds of Faculty of Physics, University of Warsaw and by the Biocentrum-Ochota project (POIG.02.03.00-00-003/09).

Tuesday poster session

kindly sponsored by National Science Foundation

P301

Modeling the landscape of cellular development using the Hopfield network

Atefeh Taherian Fard¹, Sriganesh Srihari², Jessica C Mar³, Mark A Ragan¹

¹Institute for Molecular Bioscience and ARC Centre of Excellence in Bioinformatics, The University of Queensland, Australia, ²Institute for Molecular Bioscience, The University of Queensland, Australia, ³Department of Systems & Computational Biology, Albert Einstein College of Medicine, United States

A.T. Fard <a.taherian@imb.uq.edu.au>

M.A. Ragan <m.ragan@uq.edu.au>

Genetic regulatory networks (GRN) encompass genes and transcription factors (TFs) controlling key biological processes including cellular development and differentiation (e.g. via ‘switches’ in expression levels for FOXA and OCT4 TFs). Cellular differentiation can be thought of as an ‘epigenetic landscape’ in which differentiating cells develop along predetermined trajectories on the landscape under the control of the GRN, eventually ending up in “valleys” representing differentiated cell types.

Here we model this landscape using the Hopfield network (HN), a kind of an artificial neural network constructed using a weight matrix $W[N \times N]$ between N TFs from cells $s \in S$. For a given TF pair (g_i, g_j) , $w[g_i, g_j] \in [-1, 1]$ encodes the agreement or disagreement between g_i and g_j , which here we set to their co-expression. The energy E_{ij} for the pair is given by $E_{ij} = -g_i \cdot (w_{ij}) \cdot g_j$, and the energy of the entire network is $E = -\frac{1}{2} \cdot \sum_i (g_i \cdot \sum_j (w_{ij} \cdot g_j)) = -\frac{1}{2} \cdot s^T W s$, which captures the idea that states with higher overall agreement (i.e. stable states) have lower energy.

We constructed a HN using gene expression data from $|S| = 12$ human embryonic stem cells with $|N| = 3753$ TFs across four stages of differentiation from P7 (pluripotent stage) to P4 (committed stage), as defined in [1]. The energies for these stages were: $E_4 = -3307223.92 \leq E_7 = -1320897.34 \leq E_6 = -755219.61 \leq E_5 = -599724.33$, which gave $\Delta E_{47} = 1986326.58$

($p < 0.001$). This suggested that P4 represents the most stable state in which cells commit to specific cell types, whereas P7, the stage with the highest energy represents the most pluripotent stage. The switches seen in the expression levels for FOXA and OCT4 from P4 to P7 hinted at their roles in this commitment process.

[1] Kolle G, Ho M, Zhou Q, Chy HS et al. Stem Cells (2009); 27(10):2446-56.

P302

Scaling ABySS to longer reads using spaced k-mers and Bloom filters

Shaun D Jackman¹, Karthika Raghavan¹, Benjamin P Vandervalk¹, Daniel Paulino¹, Justin Chu¹, Hamid Mohamadi¹, Anthony Raymond¹, René L Warren¹, Inanç Birol¹

¹BC Cancer Agency Genome Sciences Centre, Canada
S.D. Jackman <sjackman@bcgsc.ca>

I. Birol <ibirol@bcgsc.ca>

Adapting to the continually changing landscape of sequencing technology is a particular challenge when maintaining an assembly software package such as ABySS that spans years of development. It also offers opportunities for better assemblies if new algorithms capitalize on the technology improvements.

Illumina read lengths were 50 nucleotides at the initial release of ABySS, and overlapping MiSeq reads now exceed 500 nucleotides. ABySS and other de Bruijn graph (DBG) assemblers use a hash table to store k-mers, sequences of k nucleotides. A standard hash table requires memory that scales with the value of k. To make better use of longer read lengths without a commensurate increase in memory requires space-efficient data structures. In a new release of ABySS, we use spaced seeds to represent large k-mers while storing a fraction of their nucleotides. For example, two 32-mer separated by a space of 300 nucleotides represents a DBG comparable to a 364-mer DBG, while using the memory

of a 64-mer dBG.

We also introduce an assembly finishing tool to close scaffolding gaps in draft assemblies. The Sealer algorithm fills these gaps by navigating a dBG represented probabilistically by a Bloom filter. Because a Bloom filter is space-efficient, we can employ multiple such filters, using smaller k to span regions of low coverage and larger k to resolve repeats.

We present in this work the performance of ABySS, with a detailed look at the data structures used, and the utility of automated finishing. We demonstrate the scalability of these efficient tools to long reads and large genomes.

P303 Localised Fine Structure in Mass Spectrometry

Mateusz Krzysztof Łącki¹, Anna Gambin¹

¹University of Warsaw, Poland

M.K. Łącki <mateusz.krzysztof.lacki@gmail.com>

A. Gambin <aniag@mimuw.edu.pl>

We present two new developments in the calculation of isotopic fine structure peaks: the probability-prioritised search and the use of Poisson approximations for conditional probabilities. The first one amounts to searching for the more probable peaks before the less important ones. To our best knowledge, this simple idea has not been used in the context of isotopic calculations, exhaustive enumeration and Fourier transform methods being the methods of choice.

Most isotopic calculators aggregate together peaks that have the same nominal mass. The explicit loss of information might be important while recognising the molecules based on their mass in the high resolution mass spectrometers. It gives rise to the algorithmic problem of obtaining less aggregate peak lists. We present a novel way to calculate peaks on varying levels of aggregations. This strategy leads to reduction of the computational complexity, especially when matched with tailored Poisson approximation to the standard model of isotopic variants. One of our methods, a.k.a. DeFiner, presents a new way to approximate the probabilities of the equatransneutronic clusters, i.e. groups of isotopes differing from their element's most abundant isotope by the same number of neutrons. The other method, code-named DeFinest, can extract

individual isotopic variants, reaching the limits of what could be observed in a modern high resolution mass spectrometer.

P304 Feature Selection Server

Witold Rudnicki¹, Szymon Migacz¹, Antoni Rościszewski¹, Andrzej Sułeczki¹, Łukasz Grad¹, Magdalena Zaremba¹, Paweł Tabaszewski¹

¹Interdisciplinary Centre for Mathematical and Computational Modelling, University of Warsaw, Poland

W. Rudnicki <W.Rudnicki@icm.edu.pl>

A. Rościszewski <antoni@icm.edu.pl>

A. Sułeczki <asulecki@icm.edu.pl>

Motivation

Datasets in molecular biology have number of features reaching several thousands (gene expression, proteomics) and even millions (SNPs). Identification of the features that truly contribute to the studied phenomena is a key step, necessary for their understanding.

The number of features is so large that only univariate significance tests for each variable are routinely performed, hence any effects that depend on interactions of several features may be overlooked.

Solution

We have developed very fast CUDA based engine that allows for exhaustive multidimensional searches of all combinations of features that may lead to statistically significant relationships between k -tuple of descriptive variables and a decision variable, with $k = 2,3,4,5$. We present the service based on this engine, that aims at performing feature selection that takes into account joint influence of subsets features in the studied phenomena for data sets described with very large number of features. The first module is devoted to search epistatic interactions due to structural variations in genomes, nevertheless, the algorithm is easily extensible to more general problems and this work is under development.

Results

In the case of 2-dimensional problem with three values per variable, corresponding to the analysis of epistasis in SNP data computations for 1024 objects and 3 276 800 variables take 4018 s on NVIDIA GeForce Titan Black GPU. The GPU engine performance is $1.35 \cdot 10^9$ model evaluations per sec. With such performance it is feasible to compute

2-dimensional GWAS analysis for 10 mln SNPs on a single GPU card.

P305

ResiCon: a web service for the identification of dynamic domains, hinges and interfacial regions in proteins

Maciej Dziubinski¹, Pawel Daniluk¹, Bogdan Lesyng¹

¹University of Warsaw, Poland

M. Dziubinski <ponadto@gmail.com>

P. Daniluk <pawel@bioexploratorium.pl>

With today's simulation methods and computational power it is possible to acquire Molecular Dynamics (MD) trajectories in longer time-scales in which major protein conformational changes can be observed. Because of the large amount of data produced in the course of a simulation, it is necessary to use a fast, automatic procedures to track and describe structural motions.

We propose a new approach called ResiCon, that carries out a data-mining analysis of snapshots acquired for MD trajectories, and finds: dynamic domains (quasi-static parts), interfacial regions and hinges. ResiCon uses a time-generalized concept of contact between amino acid residues to construct a virtual scaffold representing a protein's rigidity (see [1] for the description of local structural properties, including contacts). Stiffness of edges reflects structural variability of corresponding contacts. Next, ResiCon performs a spectral clustering, identifying quasi-static regions as clusters.

Although there are several methods for predicting dynamic domains and/or flexible parts, and/or interfacial regions in proteins, they are not capable of extracting all those characteristics using a single methodology – for review and proposed solutions see e.g. [2]. Furthermore, most of the current approaches are based on the analysis of a single protein configuration. This seems beneficial, however no guarantee is given that the results of such methods are reproducible for different conformational states of the same system.

Preliminary results show that in comparison with other methods ResiCon gives better or similar results. A case study of HIV-1 protease will be presented in which ResiCon accurately determines two dynamic domains and two flaps.

We propose employing ResiCon in the study of flex-

ible (possibly natively unfolded) proteins. With today's simulation methods and computational power it is possible to acquire Molecular Dynamics (MD) trajectories in longer time-scales in which major protein conformational changes can be observed. Because of the large amount of data produced in the course of a simulation, it is necessary to use a fast, automatic procedures to track and describe structural motions.

We propose a new approach called ResiCon, that carries out a data-mining analysis of snapshots acquired for MD trajectories, and finds: dynamic domains (quasi-static parts), interfacial regions and hinges. ResiCon uses a time-generalized concept of contact between amino acid residues to construct a virtual scaffold representing a protein's rigidity (see [1] for the description of local structural properties, including contacts). Stiffness of edges reflects structural variability of corresponding contacts. Next, ResiCon performs a spectral clustering, identifying quasi-static regions as clusters.

Although there are several methods for predicting dynamic domains and/or flexible parts, and/or interfacial regions in proteins, they are not capable of extracting all those characteristics using a single methodology – for review and proposed solutions see e.g. [2]. Furthermore, most of the current approaches are based on the analysis of a single protein configuration. This seems beneficial, however no guarantee is given that the results of such methods are reproducible for different conformational states of the same system.

Preliminary results show that in comparison with other methods ResiCon gives better or similar results. A case study of HIV-1 protease will be presented in which ResiCon accurately determines two dynamic domains and two flaps.

We propose employing ResiCon in the study of flexible (possibly natively unfolded) proteins. ResiCon is publicly available at: <http://dworkowa.imdik.pan.pl/EP>

Acknowledgements

These studies were supported by the Biocentrum-Ochota project (POIG.02.03.00-00-003/09) and by the (DEC-2011/03/D/NZ2/02004) research grant of the National Science Centre.

References

1. P. Daniluk, B. Lesyng. A novel method to compare protein structures using local descriptors. *BMC Bioinformatics*, 12(1):344, 2011
2. Julia Romanowska, Krzysztof S. Nowiński,

Joanna Trylska, Determining Geometrically Stable Domains in Molecular Conformation Sets. *J. Chem. Theory Comput.*, 8 (8), 2588–2599, 2012

P306

The bottleneck of metastasis formation: insights from a stochastic model

Ewa Szczurek¹, Tyll Krüger², Niko Beerenwinkel¹

¹ETH Zurich, Switzerland, ²Wrocław University of Technology, Poland

E. Szczurek <ewa.szczurek@bsse.ethz.ch>

Metastasis formation is a complex process in which cancer cells spread from their primary tumor of origin to distant organs where they initiate new tumors. However, only a tiny fraction of disseminating tumor cells succeeds in establishing a stable colony. There is evidence that while the early steps of this process, including release to the vascular system and infiltration of the secondary organ are efficient, the bottleneck is the initial expansion of the metastatic colony in the new environment. Here, we study this rate-limiting step of metastasis initiation in a quantitative fashion using a size-dependent branching process model. Our model adds to a systematic understanding of metastasis formation and may help defining medical intervention strategies.

We compute the probability of metastatic colony survival and derive critical colony sizes under different plausible initial growth assumptions. One refers to self-stimulation, where tumor cells benefit from each others company, and another to surface hostility, where the tumor cells on the boundary are exposed to detrimental forces of the alien environment. Using established models of primary tumor growth together with our metastasis initiation model, we further obtain the probability of metastatic invasion and expected patient survival given the tumor size. These models fit well to epidemiological data collected for eleven cancers, were validated with independent datasets, and used to predict the impact of treatment delay on metastasis incidence and survival.

P307

Spaced seeds improve metagenomic classification

Karel Břinda¹, Maciej Sykulski^{1,2}, Gregory Kucherov^{1,3}

¹LIGM, Université Paris-Est Marne-la-Vallée, France,

²University of Warsaw, Poland, ³CNRS, France

M. Sykulski <macieksk@mimuw.edu.pl>

G. Kucherov <Gregory.Kucherov@u-pem.fr>

Metagenomics is a powerful approach to study the genetic material contained in environmental samples, which is revolutionized by high-throughput sequencing technologies (Next - Generation Sequencing). Alignment-free comparison methods are based on the analysis of words, usually of fixed size (k-mers), occurring in input sequences. Two recently released tools - LMAT and Kraken - perform metagenomic classification of NGS reads based on counts of shared k-mers between input reads and genomes from a pre-compiled database. In this work we extend the above metagenomics classification approach by using spaced seeds rather than contiguous k-mers.

A spaced seed is a local match triggering an alignment, encoded with a pattern of matches interleaved with spaces - jokers ("don't care" positions). Through a series of computational experiments, we show that spaced seeds significantly improve the accuracy of metagenomic classification of short NGS reads. First, we demonstrate the superiority of spaced seeds in a very simple classification setting, where sequences are classified according to identity rate. Then, we study how well different estimators: coverage/hit-number combined with spaced/contiguous seed, correlate with the alignment quality, by measuring Spearman's rank correlation coefficient, mutual information coefficient, also on real bacterial genomes of *Mycobacterium*. Finally, we modify Kraken software to make it work with spaced seeds: seed-Kraken shows a significant improvement of specificity/sensitivity trade-off. Our experiment with database of RefSeq bacterial genomes provides evidence that spaced seeds can improve the classification accuracy in real-life large-scale metagenomic projects.

P308

Utilizing the Intel Genomics Reference Architecture for Analyzing Genome Sequencing Data

Weronika Sikora-Wohlfeld¹, Abhi K. Basu², Monica Martinez-Canales², Atul J. Butte¹

¹Stanford University, United States, ²Intel Corporation, Big Data Solutions - Data Center Group, United States
W. Sikora-Wohlfeld <wsikora@stanford.edu>

The variant call format (VCF) is a standard file format commonly used to store the genetic variation data resulting from the genome sequencing data analysis. Typically, the VCF file is a starting point of the interpretation analysis, either for clinical or research purposes. Such interpretation involves integrating the variant data of interest with external information, such as other sequencing datasets and orthogonal databases. This operation requires processing the VCF file, which is often achieved using one of the available tools, e.g. VCFtools, or custom scripts. In this study we propose a novel way of representing and manipulating the variant data using the Intel Genomics Reference Architecture.

We compared the performance of the Intel Genomics Reference Architecture (database representation and in-memory queries) and VCFtools using four different benchmark tests: 1) extracting specific information from the VCF file, 2) intersecting two different VCF files, 3) annotating the VCF file and 4) calculating selected statistic using the variant data. We demonstrated that for all tests the Intel Genomics Reference Architecture offered 'near real time' solutions, which greatly outperformed VCFtools.

In this study we show that the Intel Genomics Reference Architecture can be successfully applied to the analysis of VCF files. Due to its superior time- and cost-efficiency as well as scalability, this pipeline emerges as an attractive alternative to the existing solutions, such as VCFtools. Apart from shortening the processing time, the Intel Genomics Reference Architecture offers greater flexibility in the analysis of genome sequencing data.

P309

Entropy-driven partitioning of the hierarchical protein space

Nadav Rappoport¹, Michal Linial¹

¹The Hebrew University of Jerusalem, Israel

N. Rappoport <nadavrap@cs.huji.ac.il>

M. Linial <michall@cc.huji.ac.il>

Modern protein sequencing techniques have led to the determination of >50 million protein sequences. ProtoNet is a clustering system that provides a continuous hierarchical agglomerative clustering tree for all proteins. While ProtoNet performs unsupervised classification of all included proteins, finding an optimal level of granularity for the purpose of focusing on protein functional groups remain elusive. Here, we demonstrate how knowledge-based annotations on protein families can support the automatic unsupervised methods for identifying high-quality protein families. We present a method that yields within the hierarchical clustering an optimal disjoint partition of clusters, relative to manual annotation schemes. The method's principle is to minimize the entropy-derived distance between annotation-based partitions and all available disjoint partitions. We describe the best front (BF) partition of 2478328 proteins from UniRef50. Of 4929553 ProtoNet tree clusters, BF based on Pfam annotations contain 26891 clusters. The high quality of the partition is validated by the close correspondence with the set of clusters that best describe thousands of keywords of InterPro. The BF is shown to be superior to naïve cut in the ProtoNet tree that yields a similar number of clusters. Finally, we used parameters intrinsic to the clustering process to enrich a priori the BF's clusters. We present the entropy-based method's benefit in overcoming the unavoidable limitations of nested clusters in ProtoNet. We suggest that this automatic information-based cluster selection can be useful for other large-scale annotation schemes, as well as for systematically testing and comparing putative families derived from alternative clustering methods.

P310**The statistical and MM/MD molecular modeling of MMP-TIMP system can be helpful in distinction of EDMD forms and healthy controls**

Beata Sokołowska¹, Krystiana Krzyśko^{1,2}, Irena Niebroj-Dobosz¹, Marta Hallay-Suszek², Łukasz Charzewski², Agnieszka Madej-Pilarczyk¹, Michał Marchel³, Irena Hausmanowa-Petrusewicz¹, Bogdan Lesyng²

¹Mossakowski Medical Research Center PAS, Poland,

²University of Warsaw, Faculty of Physics, Poland,

³Medical University of Warsaw, Poland

B. Sokołowska <beta.sokolowska@imdik.pan.pl>

BACKGROUND: Emery-Dreifuss muscular dystrophy (EDMD) is a very rare genetic disease affecting skeletal and heart muscles. The genes are known to be responsible for EDMD encode proteins associated with the nuclear envelope: the emerin (which is transmitted as X-linked trait, X-EDMD form) and the lamins A and C (an autosomal-dominant form, AD-EDMD). In neuromuscular disorders, changes in levels of some matrix metalloproteinases (MMPs) and their tissue inhibitors (TIMPs) are observed.

THE AIM: We present results of the statistical modeling of clinical data, as well as molecular mechanics and molecular dynamics (MM/MD) modeling, focused on interactions between some MMPs and TIMPs in healthy subjects and in patients with EDMD.

SUBJECTS AND METHODS: 1. Subjects: 25 patients with X or AD forms of EDMD and 15 aged-matched healthy controls were studied. 2. Laboratory measurements: The serum levels of MMP-1, -2, -9 and TIMP-1, -2, -3 were quantified using ELISA procedures. 3. MM/MD molecular modeling: The structural models of MMP-TIMP were determined by applying molecular modeling algorithms. 4. Statistical modeling: The pattern recognition, ROC and correlation algorithms were used.

RESULTS: Two molecular models of catalytic domain of MMP-2 or MMP-9 complex with TIMP-1 were obtained with homology-based modeling. Both were stable during the molecular dynamics simulation (10 ns, in CHARMM22 force field without restraints). The comparison of interaction interfaces between the both complexes is reported. According to the statistical analyses X- and AD-EDMD forms can be perfectly distinguished by MMP-1 and/or TIMP-1. TIMP-3 appeared to be

the most efficient in separating EDMD patients and healthy controls and it may be a novel candidate marker for EDMD.

CONCLUSIONS: The statistically significant changes of MMPs and TIMPs levels and their strong correlations are observed in EDMD patients. The presented molecular models explain the interactions in the selected complexes of the MMP-TIMP system. The estimated interactions between some of MMPs and TIMPs may be very helpful in differentiating between both EDMD forms and healthy subjects.

Acknowledgements: This study was partially supported by the Biocentrum-Ochota project (POIG.02.03.00-00-003/09) and carried out using its computational infrastructure.

P311**Multivariate Meta-Analysis of Genome-Wide Association Studies using Univariate Summary Statistics**

Anna Cichonska^{1,2}, Juho Rousu², Pekka Marttinen², Antti J Kangas^{3,4}, Pasi Soininen^{3,4}, Terho Lehtimäki⁵, Olli T Raitakari^{6,7}, Marjo-Riitta Järvelin^{8,9}, Veikko Salomaa¹⁰, Mika Ala-Korpela^{3,4,11}, Samuli Ripatti^{1,12,13}, Matti Pirinen¹

¹Institute for Molecular Medicine Finland FIMM, University of Helsinki, Finland, ²Helsinki Institute for Information Technology HIIT, Department of Information and Computer Science, Aalto University, Finland, ³Computational Medicine, Institute of Health Sciences, University of Oulu and Oulu University Hospital, Finland, ⁴NMR Metabolomics Laboratory, School of Pharmacy, University of Eastern Finland, Finland, ⁵Department of Clinical Chemistry, Fimlab Laboratories, University of Tampere School of Medicine, Finland, ⁶Department of Clinical Physiology and Nuclear Medicine, Research Centre of Applied and Preventive Cardiovascular Medicine, University of Turku, Finland, ⁷Department of Clinical Physiology and Nuclear Medicine, Turku University Hospital, Finland, ⁸Department of Epidemiology and Biostatistics, MRC Health Protection Agency Centre for Environment and Health, School of Public Health, Imperial College London, UK, ⁹Institute of Health Sciences, Biocenter Oulu, University of Oulu, Finland, ¹⁰National Institute for Health and Welfare, Finland, ¹¹Computational Medicine, University of Bristol, UK, ¹²Hjelt Institute, University of Helsinki, Finland, ¹³Department of Human Genetics, Wellcome Trust Sanger Institute, UK

A. Cichonska <anna.cichonska@helsinki.fi>

J. Rousu <juho.rousu@aalto.fi>

S. Ripatti <samuli.ripatti@helsinki.fi>

M. Pirinen <matti.pirinen@helsinki.fi>

A common approach to genome-wide association studies (GWAS) is to perform univariate tests between genotype-phenotype pairs. There are many summary-level results of such analyses publicly available. Using multivariate phenotypes results in increased statistical power and richer GWAS findings. However, low sample sizes in individual studies, and public unavailability of complete multivariate data are the current limitations. Thus, the goal of this work is to establish a computational framework for a multivariate meta-analysis of GWAS using univariate summary statistics.

Our approach is based on canonical correlation analysis, which operates on pooled covariance matrices C_{xy} (of univariate genotype-phenotype association results), C_{xx} (of genotype-genotype correlations) and C_{yy} (of phenotype-phenotype correlations), rather than requiring the original genotypic data X and phenotypic data Y . C_{xx} can be estimated from a reference database representing the study population, such as the 1000Genomes database. Phenotypic correlation structure C_{yy} is computed from C_{xy} . Furthermore, we apply some shrinkage to the full covariance matrix to add robustness to the method.

A multivariate meta-analysis of two studies of nuclear magnetic resonance metabolomics by our approach shows a good agreement with the pooled analysis of the original data sets. Average F1 score (harmonic mean of precision and recall) is equal to 0.84, and root mean squared error between the p-values from our method and the original ones is 5.83 (-log₁₀ scale). Our framework circumvents the unavailability of complete multivariate individual-level data sets, and can provide novel biological insights from already published summary statistics, upon which it is entirely based on.

P312**Searching for Horizontal Gene Transfer events in fungal genomes**

Michał Ciach¹, Anna Muszewska¹

¹Institute of Biochemistry and Biophysics, Polish Academy of Sciences, Poland

M. Ciach <m_ciach@student.uw.edu.pl>

A. Muszewska <musze@ibb.waw.pl>

Horizontal Gene Transfer (HGT) is a transfer of genetic material in a manner other than via inheritance. It is considered to be an important mechanism of evolution of prokaryotes. Many computational methods for finding HGT events in prokaryotes have been developed. These methods can be roughly divided into two groups. The first ones give credible results, but are computationally complex, thus not useful for whole-genomic search. An example is the species and gene trees reconciliation. The methods of the second group are computationally efficient, but give uncertain results. Many of these methods are based on finding an unusual pattern in a DNA sequence, like an atypical GC content and codon usage.

HGT happens also in eukaryotes, including fungi. The transferred genes, and sometimes whole gene clusters, are often of ecological significance. Thus,

even though rare, it is not a negligible phenomenon. However, little attention has been devoted to finding horizontally transferred genes in eukaryotes. Different biology, for example codon usage patterns, of eukaryotes means that some methods developed for prokaryotes cannot be directly applied.

We are developing a new method for finding horizontally transferred genes in whole fungal genomes. Our approach combines the advantages of both groups of methods. We use methods from the second group, adapted to fungal genomes, and a classifier to determine potential HGT events. They are then validated via tree reconciliation. This approach gives results of high precision and recall, at the same time being computationally efficient sufficiently for whole-genomic search.

P313

Genomic analysis of plasmids pJ2P1 and pK8P1 from Antarctic strains of *Pseudomonas* spp.

Marcin Świstak¹, Magdalena Szuplewska¹, Dariusz Bartosik¹

¹Department of Bacterial Genetics, Institute of Microbiology, University of Warsaw, Poland

M. Świstak <m.swistak@students.mimuw.edu.pl>

Bacterial genomes are very variable. Their plasticity depends on the possibility to incorporate exogenous DNA acquired through horizontal gene transfer (HGT). The main agents of HGT are mobile genetic elements (MGE), which are components of nearly all prokaryotic genomes. Among MGE the most abundant class are plasmids - natural vectors which can be transferred between bacterial cells. Plasmid genomes consist of conserved core containing information required for its replication and stabilization in host cell and additional genetic load of adaptive value.

The Department of Bacterial Genetics has gathered a collection of several dozen bacterial strains of genus *Pseudomonas* isolated from aquatic habitats of Antarctica. In order to identify genetic information crucial for adaptation to the endemic environment we performed complex analyzes of a plasmids residing in two strains of these bacteria: (i) pJ2P1 (59 kb) and (ii) pK8P1 (39 kb).

The conserved cores of both replicons contain: (i) replication and (ii) maintenance systems as well as (iii) additional genetic load. Plasmid pJ2P1 has more than 60% of open reading frames (ORFs) as-

sociated with conjugal transfer. The comparative analyses of nucleotide sequence have shown that this plasmid is related to pA506 identified in *Pseudomonas fluorescens* A506. Most ORFs of pK8P1 show similarity to phage genes coding structural proteins of virus as well as many hypothetical proteins of unknown function.

Our analyses revealed the unique data about structure and biology of bacterial mobile genetic elements and range of horizontal gene transfer between bacteria isolated from the endemic environment of Antarctica.

P314

Genome-wide Analysis of DNA Methylation in Human Breast Cancer Cells through Targeted Methyl-Seq

Chang Pyo Hong¹, Dongsung Ryu¹, Sanghoon Song¹, Junsu Ko¹

¹TheragenEteX Bio Institute, TheragenEteX, The Republic of Korea

C.P. Hong <changpyo.hong@therabio.kr>

J. Ko <junsu.ko@therabio.kr>

Dynamics modulated by epigenetic factors such as DNA methylation can regulate gene transcription, thus playing an essential role in determining phenotypic plasticity and their fates in organisms. In recent, the high-throughput sequencing approach has allowed comprehensive mapping of DNA methylation, especially CpG methylation, feasible on genome-wide scale with a high resolution. To identify CpG methylation patterns in different four human breast cancer cell lines, CpG-rich regions of each sample including CpG islands, promoters, regulatory features, cancer-specific differentially methylated regions (DMRs) were captured by using Agilent's Targeted Human Methyl-Seq kit, and then sequenced with an average of 100X depth. Dynamical change of CpG methylation was mostly found in CpG islands and shores within protein-coding genes. In particular, the analysis revealed cell-specific CpG methylation patterns in functional regulatory features such as transcription factor binding sites (TFBSs) and promoters, associated with oncogenes. We have developed an improved algorithm for finding DMRs specific to cell-type with the removal of false-positive DMRs. The DMR-related results were also linked to differentially expressed genes that were analyzed through RNA-Seq, suggesting the the epigenetic modifica-

tions by DNA methylation and the consequent transcription regulation of genes. Our result suggests an approach for targeted Methyl-Seq analysis and the potential of DNA methylation-based biomarker development associated with breast cancer.

P315

Serine protease evolution in fungi with variable lifestyles

Agata Dziedzic¹, Anna Muszewska¹

¹Institute of Biochemistry and Biophysics, Polish Academy of Sciences, Poland

A. Muszewska <ania.muszewska@gmail.com>

Fungi are able to switch between different lifestyles in order to adapt to environmental changes. Their ecological strategy is strongly connected to their secretome as fungi acquire nutrients by secreting hydrolytic enzymes to the surroundings and acquiring the digested molecules. In this study we focus on fungal serine proteases, which distribution is barely described so far. Expanding the repertoire of known proteases in fungal genomes will facilitate evolutionary studies of fungal secretome and will help to understand the relation between secretome and fungal lifestyle.

In order to obtain a complete set of fungal proteases, we performed iterative jackhammer searches against Uniprot protein sequence database and Blast searches against JGI genomes database. Obtained results suggest that serine proteases are more ubiquitous than expected. From 53 serine protease families described in Merops Peptidase Database, 19 are present in fungi. Interestingly, 18 of them are also present in Metazoa - this suggest that, except one (S64), all fungal serine proteases families evolved before animals and fungi diverged. This hypothesis is supported by the presence of most serine proteases in ancestral fungal groups, i. e. Chytridiomycota, Microsporidia, Mucorales. Concerning all fungi species together, the contribution of serine protease families varies. The most abundant are S9 proteases (20.000 species), whereas only 89 species encode proteins from S49 family. Our study shows that S49 is the only one from 19 fungal families not present Ascomycota.

Here, we present a comprehensive evolutionary history of fungal serine protease families in the context of fungal ecology and the fungal tree of life.

P316

Comparison of DNA methylation events in three animal models of temporal lobe epilepsy

Konrad Dębski¹, Katja Kobow², Anna Bot¹, Noora Huusko³, Mark Ziemann⁴, Antony Kaspi⁴, Assam El-Osta⁴, Ingmar Blumcke², Asla Pitkänen^{3,5}, Katarzyna Łukasiuk¹

¹Nencki Institute of Experimental Biology Polish Academy of Sciences, Poland, ²Department of Neuropathology, University Hospital Erlangen, Germany, ³A.I. Virtanen institute for molecular sciences, University of Eastern Finland, Finland, ⁴Epigenetics in Human Health and Disease, Baker IDI Heart and Diabetes Institute, The Alfred Medical Research and Education Precinct, Australia, ⁵Department of Neurology, Kuopio University Hospital, Finland

K. Dębski <k.debski@nencki.gov.pl>

DNA methylation is recently postulated to play a crucial role in gene regulation involved in epileptogenesis. We explore next generation sequencing data obtained from three animal models of temporal lobe epilepsy (TLE) (amygdala stimulation induced SE - aSE, pilocarpine induced SE - pilo, Traumatic Brain Injury - TBI) at 3 mo after induction of epilepsy to find common changes in DNA methylation occurring in epileptic hippocampus at chronic state. We assigned differentially methylated regions ($p < 0.01$) to multiple genomic features (CpG Islands, SNPs, most conserved sequences, promoters, TSS, 5'UTRs, Exons, Introns, Gene Bodies, 3'UTRs, and 5kb downstream regions).

Distributions of differentially methylated regions differs among the genomic features and it seems to be a model specific. When models were compared in pairs there was 20 methylated regions common between aSE and pilo, 24 common regions between aSE and TBI and 29 between pilo and TBI.

This work is supported by Polish Ministry of Science and Higher Education grant DNP/N119/ESF-EuroEPINOMICS/2012 and National Science Centre, Poland, grant 2014/13/N/NZ2/00587.

P317**Identifying batch effects in high-throughput biological data using dynamic programming based approach****Anna Papiez¹, Michal Marczyk¹, Andrzej Polanski¹, Joanna Polanska²**¹Silesian University of Technology, Poland, ²The Silesian University of Technology, Poland

A. Papiez <anna.papiez@polsl.pl>

M. Marczyk <michal.marczyk@polsl.pl>

J. Polanska <joanna.polanska@polsl.pl>

Batch effects are technical sources of variation present in high-throughput data in various experiments in the molecular biology field. These often may cause unexpected bias in the data as they are usually due to a multitude of factors related with the probe preparation step of a study, some of which are impossible to predict. It is important to consider potential batch effects, as there are numerous cases where data normalization does not entirely eliminate the issue. Therefore, diverse algorithms have been developed in order to identify and filter this systematic bias.

Regarding batch effect identification, the current techniques are based on modifications of principal component analysis (guided PCA) and a combined identify-remove method implemented as Surrogate Variable Analysis. We propose a different approach that attempts at establishing the partition into batches using dynamic programming. The dynamic programming method has been tested on a number of publicly available microarray datasets from the GEO database that have an explicit division into batches with regard to the time of sample preparation and scanning. The result were further used as input to the gold-standard batch effect filtration method available in the ComBat software.

Proposed procedures has shown to have significant impact on data integrity, and moreover, on the number of features identified as differentially expressed. The research proved batch effect identification and removal to be a necessary step in biomedical high-dimensional data preprocessing. The presented preliminary results demonstrate that the appropriate choice of batch identifying procedure plays a key role in the experimental outcome.

Acknowledgement: This work was supported by SUT grant BKM/524/RAu1/2014/t.17 (AP), POIG.02.03.01-24-099/13(MM) and BK/265/RAU1/2014/10 (JP). All the calculations were carried out using GeCONiI infrastruc-

ture funded by project number POIG.02.03.01-24-099/13.

P318**Amino Acid Properties Conserved in Molecular Evolution****Witold Rudnicki¹, Teresa Mroczek², Paweł Cudek²**¹Interdisciplinary Centre for Mathematical and Computational Modelling, University of Warsaw, Poland,²University of Information Technology and Management, Poland

W. Rudnicki <W.Rudnicki@icm.edu.pl>

T. Mroczek <tmroczek@wsiz.rzeszow.pl>

That amino acid properties are responsible for the way protein molecules evolve is natural and is also reasonably well supported both by the structure of the genetic code and, to a large extent, by the experimental measures of the amino acid similarity. Nevertheless, there remains a significant gap between observed similarity matrices and their reconstructions from amino acid properties.

Therefore, we introduce a simple theoretical model of amino acid similarity matrices, which allows splitting the matrix into two parts – one that depends only on mutabilities of amino acids and another that depends on pairwise similarities between them. Then the new synthetic amino acid properties are derived from the pairwise similarities and used to reconstruct similarity matrices covering a wide range of information entropies.

Our model allows us to explain up to 94% of the variability in the BLOSUM family of the amino acids similarity matrices in terms of amino acid properties. The new properties derived from amino acid similarity matrices correlate highly with properties known to be important for molecular evolution such as hydrophobicity, size, shape and charge of amino acids.

This result closes the gap in our understanding of the influence of amino acids on evolution at the molecular level. The methods were applied to the single family of similarity matrices used often in general sequence homology searches, but it is general and can be used also for more specific matrices. The new synthetic properties can be used in analyzes of protein sequences in various biological applications.

P319**Joint Analysis of Differential Gene Expression in Multiple Studies using Correlation Motifs**Yingying Wei¹, Toyooki Tenzen², Hongkai Ji³¹The Chinese University of Hong Kong, Hong Kong,²Massachusetts General Hospital, United States, ³The Johns Hopkins University, United States

Y. Wei <ywei@sta.cuhk.edu.hk>

The standard methods for detecting differential gene expression are mostly designed for analyzing a single gene expression experiment. When data from multiple related gene expression studies are available, separately analyzing each study is not ideal as it may fail to detect important genes with consistent but relatively weak differential signals in multiple studies. Jointly modeling all data allows one to borrow information across studies to improve the analysis. However, a simple concordance model, in which each gene is assumed to be differential in either all studies or none of the studies, is incapable of handling genes with study-specific differential expression. In contrast, a model that naively enumerates and analyzes all possible differential patterns across studies can deal with study-specificity and allow information pooling, but the complexity of its parameter space grows exponentially as the number of studies increases. Here, we propose a correlation motif approach to address this dilemma. This approach searches for a small number of latent probability vectors called correlation motifs to capture the major correlation patterns among multiple studies. The motifs provide the basis for sharing information among studies and genes. The approach has flexibility to handle all possible study-specific differential patterns. It improves detection of differential expression and overcomes the barrier of exponential model complexity.

P320**Gene signature optimization based on improved method for enrichment analysis**Wojciech Labaj¹, Andrzej Polanski¹¹Silesian University of Technology, Poland

W. Labaj <wojciech.labaj@polsl.pl>

Modern high-throughput DNA microarrays can measure in parallel the expression of hundreds of thousands of targets in form of genes or their alternative variants. These are often summarized by

so called gene signatures – list of genes exhibiting certain patterns of expression in response to experiment perturbation. Often such gene signatures are used as tumor markers – features for classification algorithms. In such an application, in addition to measuring the classification quality it is crucial to assess the stability of these gene signatures.

Functional analysis based on the gene signatures is a complex task. There are still challenges in efficient applications of gene signatures optimization, which include instability of composition of gene signatures, problems in defining sizes (numbers of genes) of gene signatures and possible unreliability of results of inference based on gene signatures. Therefore there are many efforts towards improving algorithms for optimization of gene signatures.

One of the possible way to optimize the size of gene signatures with help of state-of-the-art biological knowledge is the re-annotation of genes to statistical significant GO terms. While there is number of approaches for such GO terms enrichment analysis the selection of one to be used is hard due to the lack of their comprehensive comparison.

Here we provide an approach for unbiased comparison based on classification quality as well as measuring of gene signature stability. With the use of this framework we also present advantages of our own method for gene signature optimization. It is based on the fusion of information coming from statistical testing of differential expression genes and information resulting from statistical testing of enrichment for GO terms.

Acknowledgement: This work was supported by SUT grants no. BKM/525/RAU-2/2014 (WL). The calculations were carried out using GeCONiI infrastructure (POIG.02.03.01-24-099/13).

P321**Accounting for biases in riboprofiling data indicates a correlation between typical codon decoding rates and tRNA copy numbers**Alon Diamant¹, Tamir Tuller¹¹Biomedical Engineering Dept., Tel Aviv University, Israel

A. Diamant <dalon@post.tau.ac.il>

T. Tuller <tamirtul@post.tau.ac.il>

The possible effect of transfer ribonucleic acid (tRNA) concentrations on codon decoding times is a fundamental biomedical research question, and has been continuously debated in recent years. Pre-

vious studies have utilized riboprofiling (Ribo-Seq), a large-scale protocol for monitoring ribosome occupancy on specific mRNAs at nucleotide resolution, to investigate the relation between codon decoding rates as inferred from riboprofiling and tRNA abundance, arriving at contradictory conclusions. This discrepancy is possibly related to high levels of noise and bias in riboprofiling data that call for robust analysis methods and informed modeling. Recently, a promising new approach to account for systematic biases in the riboprofiling protocol, among others by comparing its output to the corresponding output of an RNA-seq experiment, has been suggested by Artieri and Fraser. However, the authors' attempts to identify determinants of translation elongation speed in the coding sequence indicated a role mainly for a single amino acid (proline). Here we show that the rejection of some of the factors that possibly affect translation elongation significantly was due to inaccurate modeling. By utilizing a model for typical codon decoding rates (TDR) that was recently proposed by Dana and Tuller, we show that TDRs are significantly correlated with tRNA copy numbers also after accounting for systematic riboprofiling biases (as was suggested by Artieri and Fraser). Our results indicate that tRNA levels in the cell directly affect translation efficiency and that these conclusions are not due to known biases in the data.

P322

CABS-dock web server for flexible docking of peptides to proteins without prior knowledge of the binding site

Mateusz Kurcinski¹, Michal Jamroz¹, Maciej Blaszczyk¹, Andrzej Kolinski¹, Sebastian Kmiecik¹

¹University of Warsaw, Poland

A. Kolinski <kolinski@chem.uw.edu.pl>

S. Kmiecik <sekmi@chem.uw.edu.pl>

Protein-peptide interactions play a key role in cell functions. Their structural characterization is challenging and highly desirable in drug discovery. In this field, molecular docking techniques have already proven useful. However, because of the limitations in conformational search efficiency, the current docking algorithms accounting for peptide flexibility have to be driven by the knowledge of the binding site. Here, we present highly effective CABS-dock method for flexible protein-peptide docking without prior knowledge of the binding site

[1, 2]. CABS-dock is a multiscale method merging efficient coarse-grained modeling with all-atom optimization of predicted models. CABS-dock allows for the full flexibility of a peptide during its search for a binding site over a protein receptor surface, while fluctuations of receptor structure are also taken into account. Starting from random peptide conformations and positions, we obtain high or medium resolution models over the largest benchmark dataset available to date (including both, bound and unbound docking cases). Our unique method for coupled binding site search and protein-peptide docking can be easily complemented by other computational tools (e.g. high-resolution docking refinement protocols, binding site predictors methods) or experimental data to improve the results of the docking experiment. CABS-dock is freely available as a web server at <http://biocomp.chem.uw.edu.pl/CABSdock>

References

[1] Kurcinski M, Kolinski A, Kmiecik S. (2014) Mechanism of Folding and Binding of an Intrinsically Disordered Protein As Revealed by *ab Initio* Simulations. *Journal of Chemical Theory and Computation*. 10, 2224-2231.

[2] Kurcinski M, Jamroz M, Blaszczyk M, Kolinski A, Kmiecik S. (2015) CABS-dock web server for flexible docking of peptides to proteins without prior knowledge of the binding site. *Nucleic Acids Research* (submitted).

P323

Optimal mutation rates for adaptation in Fisher's geometric model with environmental stress.

Michał Startek¹, Arnaud Le Rouzic², Anna Gambin¹

¹University of Warsaw, Poland, ²Centre national de la recherche scientifique, France

M. Startek <mist@mimuw.edu.pl>

A. Gambin <aniag@mimuw.edu.pl>

Fisher's geometric model is a well-established (class of) models in population genetics, wherein the organism's phenotype is represented in a geometric setting, as a vector of real-valued parameters. Selection is based on a fitness function (dependant on the phenotype), and random mutations are modelled by random changes to the organism's phenotype. The population is represented as a probability measure

over the space of possible phenotypes, representing the proportion of organisms with given phenotype. The applications of these models range from studying adaptation to an environment, to simulating the evolutionary effects of pleiotropy.

In our study we have focused on populations undergoing selection in an environment with shifting phenotypic optimum, and we model their adaptation as well as response to environmental stimuli. We have derived analytical solutions of the model, and we provide an analysis of the equilibrium states reached by populations in varied conditions, allowing us to derive closed formulas for genetic variance, optimal mutation rate, or average fitness (among others).

We shall compare these results to results obtained from computational model in which the mutation rate of a given organism itself is allowed to evolve (as it would in nature for example due to proliferation of transposable elements, or polymerase mutations), and present the surprising conclusions.

P324

Docking based 3D-QSAR studies applied at the BRAF inhibitors to understand the binding mechanism Uzma Mahmood¹, Zaheer Ul-Haq²

¹Sir Syed University of Engineering & Technology, Pakistan, ²ICCBS, Pakistan

U. Mahmood <mehmoodchemist@gmail.com>

BRAF is a great therapeutic target in a wide variety of human cancers. It is the member of Ras Activating Factor (RAF) family of serine/threonine kinase. The RAF signal transduction cascade is a conserved protein pathway that is involved in cell cycle progression and apoptosis. The ERK regulates phosphorylation of different proteins either in cytosol or in nucleus but disorders in ERK signaling pathway cause mutation in BRAF. This cascade in these cells may provide selection of mutated BRAF in which valine is substituted with glutamic acid at position 600. This mutation occurs in activation loop. A number of inhibitors reported to target different members of RAF, some of them have potential to target the BRAF as well. We report the first time detailed study of 3D-QSAR and molecular docking to find interactions and modify the structure on the basis of predicted results for BRAF inhibitors. This study has been designed for the most reliable techniques of 3D QSAR i.e Comparative Molecular

Field Analysis (CoMFA) and Comparative Molecular Similarity Indices Analysis (CoMSIA) of three different data sets. The data sets selected for better evaluation of BRAF inhibitors belongs to 2, 6-Disubstituted Pyrazine, Pyridoimidazolones and its derivatives. Our models would offer help to better comprehend the structure-activity relationships that exist for these classes of compounds and also facilitate the design of novel inhibitors with good chemical diversity. Combined structure based study of Docking and 3D-QSAR studies have provided the evidence to a better understanding of interaction between the inhibitors and BRAF.

P325

Trend control focused integration in modeling of genotype-phenotype interactions

Joanna Zyla¹, Marzena Dolbniak¹, Christophe Badie², Ghazi Alsbeih³, Joanna Polanska¹

¹Silesian University of Technology, Institute of Automatic Control, Poland, ²Public Health England, United Kingdom, ³King Faisal Specialist Hospital & Research Centre, Saudi Arabia

J. Zyla <joanna.zyla@polsl.pl>

J. Polanska <joanna.polanska@polsl.pl>

Integrative analysis is a very powerful tool allowing for combination of the experimental results from several studies. The aim is to develop the integration algorithm of genotype-phenotype modeling with signal trend control.

Material and methods. The sample consists of 44 unrelated individuals, with two type of experimental data collected: genotyping of 567,095 SNPs, and results of qPCR BBC3 and FDXR expressions in two conditions - before and after the 2Gy irradiation. Both investigated genes have pro-apoptotic abilities and are regulated by p53. Two models of interactions were investigated per each SNP-gene. Applying proper statistical one-side tests allowed to control the trend in response to irradiation. The restricted approach requires that for both analysed genes and particular SNP the obtained models are of the same type and direction and statistically significant at a priori assumed level (in our case $\alpha=0.05$). The integration approach replaces the last condition by requirement on integrated p-value being significant.

Results. The restricted approach gives 19,819 SNPs (up-dominant 4,477; down-dominant 4,863; up-recessive 5,319; down-recessive 5,160) with 230 nsS-

NPs, while the integration brings 113,078 in total (26,966; 26,287; 29,380; 30,445 respectively) with 1351 nsSNPs. The comparative functional analysis revealed four relevant SNPs in gene PTPN13 not observed in restricted approach results. PTPN13 plays role in oncogenic transformation and giving pro-apoptotic signals down regulate BBC3 and FDXR.

Conclusions. Integration combined with controlling after the response type and genotype-phenotype interactions increases the number of relevant SNPs enriching biological information obtained.

Acknowledgement: The work was supported by SUT-BKM/514/Rau1/2014/t.16 and GeCONiI (POIG.02.03.01-24-099/13).

P326

Usage of bootstrap sampling in comparing biological diversity of differently sized data samples

Justyna Kotas¹, Christophe Badie², Serge Candéias³, Joanna Polańska¹

¹Silesian University of Technology, Poland, ²Public Health England, United Kingdom, ³The French Alternative Energies and Atomic Energy Commission (CEA), France

J. Kotas <Justyna.Kotas@polsl.pl>

J. Polańska <Joanna.Polanska@polsl.pl>

AIM: The aim of the study was to develop a procedure for statistical comparison of biodiversity of samples with extremely big number of classes (> 7,000) and differing strongly in their size.

MATERIALS AND METHODS: The data were from high-throughput sequencing of murine T cell receptors. The experiment was performed for 3 irradiation doses (0Gy, 0.1Gy, 1Gy) at 3 timepoints (1, 3, 6 months). A wide range of unique sequences was detected giving a 12.5 fold change between two extreme samples. The count of detected classes comes up to more than 94,000 sequences. To compare the distribution of sequences Pielou's diversity measure (J) was proposed to diminish effects of differently sized data and a modification of Hutcheson t-test was applied to compare equitability indices. To correctly estimate variance a multinomial distribution based bootstrap method was used (2,000 iterations).

RESULTS: Almost all groups differ with respect to sequence diversity ($p < 10^{-7}$). The distribution of sequences changes in time and with dose, becoming

more uniformly distributed with age.

group	1m_0gy	3m_0gy	6m_0gy	1m_01gy
j index	0,94710	0,96096	0,96555	0,92888
-95%	0,94691	0,96085	0,96546	0,92858
+95%	0,94729	0,96107	0,96564	0,92918
3m_01gy	6m_01gy	1m_1gy	3m_1gy	6m_1gy
0,96565	0,96217	0,93930	0,96827	0,96900
0,96557	0,96208	0,93897	0,96819	0,96891
0,96573	0,96226	0,93964	0,96835	0,96909

CONCLUSIONS: Pielou's index allows for comparing diversities of differently sized data while bootstrap sampling allows for correct estimation of confidence intervals, enhancing the reliability of the results.

ACKNOWLEDGMENTS: The work was supported by HARMONIA-4/2013/08/M/ST6/924 and GeCONiI (POIG.02.03.01-24-099/13).

P327

Copy Number Variation Analysis of Diffuse Large B-Cell Lymphoma (DLBCL) Subtypes

Prashanthi Dharanipragada¹, Nita Parekh¹

¹IIT, India

P. Dharanipragada <prashanthi.d@research.iit.ac.in>

Diffuse large B-cell lymphoma (DLBCL) is the most aggressive type of non-Hodgkin lymphoma and is fatal in 50% of patients. It is heterogeneous at the molecular level and divided into three main subtypes: Germinal center B-cell (GCB), Active B-cell (ABC) and Primary mediastinal B cell lymphoma (PMBL). The subtypes originate from different stages of B cell differentiation and are characterized by distinct mechanisms of oncogenic activation. However, the underlying molecular framework of DLBCL pathogenesis is poorly understood, leading to an absence of specific prognosis and targeted therapeutics. Role of copy number variations (CNVs) in cancer is well known. In this study, we carry out an analysis of CNVs to understand the molecular pathways involved in DLBCL progression in two subtypes GCB and ABC. Analysis is carried out on publicly available NGS data and CNVs are detected using an R based tool, ReadDepth. The tool detected 1903 (GCB) and 2188 (ABC) CNVs,

of which 953 CNVs are common between the two subtypes. A pathway enrichment analysis of unique CNV associated genes listed in COSMIC database was performed. CNV-affected genes detected in GCB showed high enrichment of Glypican signaling pathway, involved in cell growth and proliferation through TGF-beta signaling, Wnt and MAPK pathways, and Arf6 trafficking event that play a role in tumor cell invasion. A fraction of ABC subtype CNVs were highly enriched with genes participating in chromatin modification of oncogenes and also found previously unreported TRAIL signaling and Nectin adhesion pathways overrepresented in ABC subtype with high confidence.

P328

In Silico Homology Modeling, Functional annotation and Molecular Docking Studies of hypothetical proteins of *Aspergillus fumigatus* Af293.

Shah Md. Shahik¹, Ismot Ara²

¹Department of Genetic Engineering and Biotechnology, Faculty of Biological Sciences, University of Chittagong, Bangladesh, ²Department of Computer Science and Engineering, Faculty of Science and Technology, Atish Dipankar University of Science and Technology, Bangladesh

S.M. Shahik <sm.shahik@gmail.com>

Background: Genome sequencing projects has led to an explosion of huge amount of gene products in which many are of hypothetical proteins with unidentified function. Exploring and annotating the functions of hypothetical proteins is important in *Aspergillus fumigatus* which is a pathogenic fungus that cause disease in individuals with an immunodeficiency.

Methods: In this study sequence of 5 hypothetical proteins of *Aspergillus fumigatus* Af293 has been annotated from NCBI. Various computational tools such as 3D structure was determined by means of homology modeling through Phyre2 and refined by ModRefiner. Then, the designed structure was evaluated with a structure validation program for instance ERRAT, PROCHECK and QMEAN, for further structural analysis. Proteins secondary structural features were determined through SOPMA and interacting networks by STRING. The receptor was analyzed for the active site and pocket finder tools. Docking studies were done through

Autodock Vina software.

Results: The 3D structure of five proteins were modeled and their ligand binding sites were identified. We have found domains and families of two protein and this are revealed that, these proteins might have ADP- ribosylation activity, molecular chaperone, heat shock protein activity, plasminogen/ hepatocyte growth factor activity etc. Protein-ligand study of our predicted ligand shows the lowest energy of -7.4 kcal/mol for gi|66851587 and -6.7 kcal/mol for gi|66845370. **Conclusion:** Structural prediction of these proteins, detection of binding sites and drug designing of current study will cover the way for further extensive investigation of this proteins in wet lab experiments and in that way assist drug design against Aspergillosis.

P329

Genome-wide mapping and computational analysis of non-B DNA structures in vivo

Damian Wójtowicz¹, Fedor Kouzine¹, Ar-ito Yamane², Craig J. Benham³, Rafael C. Casellas¹, David Levens¹, Teresa M. Przytycka¹

¹National Institutes of Health, United States, ²Gunma University, Japan, ³University of California, Davis, United States

D. Wójtowicz <Damian.Wojtowicz@nih.gov>

The canonical double helical structure of B-DNA may undergo various deformations to adopt alternative conformations, non-B DNA structures, including single-stranded DNA, Z-DNA, G-quadruplex, H-DNA, cruciform. Previous studies confirmed the existence of some non-B DNAs in a few gene promoters (e.g. c-myc, ADAM-12) and implicated their role in gene regulation, but it is not known how abundant the alternative DNA conformations might be at the genomic level. Computer-based studies uncovered a large number of sequences across the mammalian genomes that can potentially form non-B DNA structure and play functional roles in regulating DNA transactions.

We developed a new experimental technique, which combines chemical and enzymatic techniques with high-throughput sequencing, to map non-B DNA conformations at the genomic scale in vivo. The new protocol was applied to identify in vivo formation of non-B DNA structures in the genomic

DNA of mouse and human cells. We performed a genome-wide analysis of occurrences of these alternative DNA conformations and compared the experimental data to genomic regions computationally predicted to have a propensity to form non-B DNA conformations. We showed a significant enrichment of ssDNA signal near computationally predicted non-B DNA motifs. Moreover, each type of predicted non-B DNA structures has a distinctive experimentally derived signature. This study provides the first look at genome-wide landscape of the in vivo formation of alternative structures. We find non-B DNA structures to be a common feature of mammalian genomes and to be associated with gene function and cell state. This study promises to be a useful resource for further studies to explore the role of non-B DNA structures in the regulation of DNA transactions.

P330 **Episode clustering problems for unrooted gene trees**

Jaroslav Paszek¹, Paweł Górecki¹

¹Faculty of Mathematics, Informatics and Mechanics
University of Warsaw, Poland
J. Paszek <jaroslav.paszek@gmail.com>

Discovering the location of gene duplications and multiple gene duplication episodes is a fundamental issue in evolutionary molecular biology. The problem introduced by Guigo et al. in 1996 is to map gene duplication events from a collection of rooted, binary gene family trees onto their corresponding rooted binary species tree in such a way that the total number of multiple gene duplication episodes is minimized. There are several models in the literature that specify how gene duplications from gene families can be interpreted as one duplication episode. However, in all duplication episode problems gene trees are rooted. This restriction limits the applicability, since unrooted gene family trees are frequently inferred by phylogenetic methods.

In this poster we propose to extend episode clustering problems to the case when the input gene family trees are unrooted. In particular, by using theoretical properties of unrooted reconciliation, we show an efficient algorithm that reduces our problems into the episode clustering problems defined for rooted trees. We show theoretical properties of the reduction algorithm and evaluation of an empirical dataset.

P331 **Towards more realistic tree reconciliation**

Agnieszka Mykowiecka¹, Paweł Górecki¹

¹Institute of Informatics, Faculty of Mathematics, Informatics and Mechanics, University of Warsaw, Poland
A. Mykowiecka <agnieszka.mykowiecka@mimuw.edu.pl>
P. Górecki <gorecki@mimuw.edu.pl>

Phylogenetic trees are usually reconstructed from aligned DNA sequences by neighbor joining, maximum likelihood or maximum parsimony. In general, different methods yield different results and no clue on which of them, if any, is correct. Bootstrap method allows for assigning confidence value for a hypothesis that a particular tree is a good approximation of evolution of a given set of DNA sequences. Bootstrapping consists in reconstruction of phylogenetic trees for DNA sequence alignments altered by random modification or sampling. The variability in the obtained set is assessed by comparing its elements with the original tree. The results can be interpreted as an indication of the influence of arbitrary changes which do not resemble evolutionary schema on the structure of the phylogenetic tree.

A reconciliation is a map which relates a gene tree with its species tree by postulating gene duplication events. As in the case of phylogenetic tree reconstruction, the resulting scenario of gene duplications may be incorrect. In our work we propose to use bootstrapping approach to define support for gene duplication events when reconciling a given gene tree with its species tree. By comparing gene trees obtained by bootstrapping to the original gene tree we calculate support for both clusters and gene duplication events. While this approach can be used to annotate orthology and paralogy, we show how this method can be used to verify reliability of tree reconciliation with applications to supertree problem.

P332**On the nature of potential energy surface in autism related synaptic protein NRX-NLG complex**Rafal Jakubowski¹, Lukasz Peplowski¹, Wieslaw Nowak¹¹Faculty of Physics, Astronomy and Informatics, Institute of Physics, Nicolaus Copernicus University, Poland
R. Jakubowski <rjakubowski@fizyka.umk.pl>

W. Nowak <wiesiek@fizyka.umk.pl>

The neurexin1 β -neuroligin1 (NRX-NLG) is a protein complex occurring in human synaptic cleft, playing proven role in synaptic junctions formation, maturation, signal transduction and angiogenesis. Unfortunately, mutations occurring in genes encoding these proteins may lead to neural-related disorders like schizophrenia, autism spectrum disorders or mental retardation.

The formation of the complex is governed by intermolecular interactions between NRX and NLG. It is important to know how close these protein must approach each other to form the junctions effectively. Computational techniques using a classical force field concept may answer this question. Here we propose a new algorithm suitable for such studies. We expand our energy minimization based ligand pose refinement approach to larger biological complexes. We employ steered molecular dynamics for enforced dissociation of the complex, which give us force vs. center of mass position curve. Snapshots of the system in force maxima are subject conjugent gradients minimization. Finally, we score obtained results in terms of interface native contacts reproduction. We also determine radius of coalescence of the minimization procedure, as well as we optimize a number of minimization steps in order to keep the best accuracy/computational cost ratio.

Our research brings new insights on computationally obtained force values and points out a new relationship between systems' state and the effective range of most important interactions leading to the complex formation. Data will serve to estimate biological effects of dangerous mutations in NRX and/or NLG genes.

This work was supported by the "Krok w Przyszlosc, 5 edycja" scholarship from the Marshal of Kuyavian-pomeranian voivodeship.

P333**Accurate inference of thousands of novel translated open reading frames and dually coded regions using ribosome footprinting data.**Anil Raj¹, Sidney Wang², Heejung Shim², Yang Li¹, Matthew Stephens³, Yoav Gilad², Jonathan Pritchard^{4,5}¹Stanford University, Genetics, United States,²University of Chicago, Human Genetics, United States,³University of Chicago, Statistics and Human Genetics, United States, ⁴Stanford University, Genetics and Biology, United States, ⁵Howard Hughes Medical Institute, United States

A. Raj <rajanil@stanford.edu>

Understanding the functional effects of gene expression critically depends on the accurate and comprehensive annotation of sequence elements that are translated in each gene. Ribosome footprinting provides an unbiased, high coverage measurement of translation. Although ribosome footprints have been used to identify translated transcripts, these data have not been used to elucidate the precise set of open reading frames (ORF) that are being translated in each transcript.

In this work, we used hidden Markov models to analyze ribosome footprinting data along with sequence information in the cell-specific transcriptome. Our analyses accurately resolves the ORFs that are being translated in human lymphoblastoid cell lines (LCL). We identified thousands of transcripts with previously unannotated translated ORFs, including hundreds of short translated ORFs and hundreds of transcripts with dually coded translated regions. Nearly 70% of previously undiscovered translated regions predicted to have a peptide match were validated using MS data generated in the same cell line. Approximately 90% of these validated novel translated regions consisted of processed pseudogenes and dually coded transcripts, and substantial numbers of these regions are strongly conserved in vertebrates. Finally, using ribosome footprinting data in LCLs from a panel of 72 Yoruba individuals, we find strong evidence for inter-individual variation in the translation of the pair of ORFs for most dually coded transcripts, suggesting regulatory control of ORF usage. The application of our method to ribosome footprinting data across cell types and organisms may prove useful in teas-

ing apart functional and evolutionary differences in gene expression between tissues and species.

P334 **Sequence-structure-function relationships in the GmrSD family of Type IV restriction enzymes**

Magdalena A. Machnicka¹, Katarzyna H. Kaminska¹, Stanislaw Dunin-Horkawicz¹, Janusz M. Bujnicki¹

¹International Institute of Molecular and Cell Biology in Warsaw, Poland

M.A. Machnicka <mmika@genesilico.pl>

GmrSD is a modification-dependent restriction endonuclease that specifically targets and digests glucosylated hydroxymethylcytosine (glc-HMC) modified DNA. It is encoded either as two separate single-domain GmrS and GmrD proteins or as a single protein carrying both domains. Previous studies suggested that GmrS acts as endonuclease and NTPase whereas GmrD binds DNA. Results of this comparative genomics study show that in fact GmrS exhibits similarity to NTP-hydrolyzing domains, however it also might be responsible for DNA recognition. Also in contrast to the previous studies, we attribute the nuclease activity to GmrD as we found it to contain the HNH endonuclease domain. We revealed residues potentially important for structure and function in both domains. Moreover, we found that GmrSD systems exist predominantly as a fused, double-domain form rather than as a heterodimer and that their homologs are often encoded in regions enriched in defense and gene mobility-related elements. Finally, phylogenetic reconstructions of GmrS and GmrD domains revealed that they coevolved and only few GmrSD systems appear to be assembled from distantly related GmrS and GmrD components.

P335

Building a consensus co-expression network for CNS genes relevant to pharmacogenomics

Samuel Handelman¹, Michal Seweryn², Andrzej Kloczkowski³, Wolfgang Sadee⁴

¹Ohio State University, United States, ²Uniwersytet Łódzki, Poland, ³Nationwide Children's Hospital, United States, ⁴The Ohio State University, Department of Pharmacology, United States

S. Handelman <samuel.handelman@gmail.com>

In studies associating genetic variants with efficacy or side-effects of psychiatric medications (for schizophrenia, depression and bipolar disorder), key variants in less than 40 CNS-well-expressed genes are repeatedly implicated. Using next generation sequencing expression measurements from: the Genotype-Tissue Expression project (GTEx); five human studies deposited in the gene expression omnibus; and, from brain transcriptomes produced internally for the use of Expression Genetics in Drug Therapy (XGEN), co-expression networks including these repeatedly-implicated genes are identified. Several alternative methods are utilized: Weighted Gene Coexpression Network Analysis (WGCNA), Algorithm for the Reconstruction of Accurate Cellular Networks (ARACNE), and variations on ARACNE incorporating alternative information measures. Different methods produce divergent results in terms of reproducibility within and between the available expression experiments and tissues, assessed using random nested resampling of the original transcriptomes and data series. The sparse and parsimonious (not-highly connected) consensus co-expression network has relevance both to biological interpretation and to the identification of potential statistical epistasis between gene variants for CNS pharmacotherapy.

P336**Data driven estimation of cutoffs in searching for differentially expressed proteins**

Agnieszka Blachowicz¹, Soile Tapio², Zarko Barjaktarovic², Rafi Benotmane³, Rosemary Finnon⁴, Christophe Badie⁴, Simon Bouffler⁴, Joanna Polanska¹

¹Silesian University of Technology, Poland, ²Helmholtz Zentrum München, Germany, ³SCK•CEN, Belgium, ⁴Public Health England, United Kingdom

A. Blachowicz <agnieszka.blachowicz@polsl.pl>

J. Polanska <Joanna.Polanska@polsl.pl>

There is a lot of statistical and data mining algorithms devoted to the analysis of high-throughput omics data, but very few of them are design for small number of samples. The aim of this study is to the find data driven cutoffs for experiments with extremely small sample size. The proposed solution use mathematical modeling of signal distribution to obtain required thresholds and to estimate the significance of the results.

Materials and methods: An experiment was performed for 6 types of biological material taken from AML mice. The protein expression was related to the control level obtained for a control group of healthy mice. Cells were extracted from mouse spleentissue. The first part of them was examined immediately and the second part was used for cell culture setting up (which was examined later). The expression of 320 proteins was measured. Having only one technical replicate, it is impossible to find deregulated proteins using classical statistical tests. Decomposition of signal distribution into Gaussian Mixture Model combined with maximum probability criterion allows for cutoff estimation.

Results: The optimal number of Gaussian components was 2 in each condition. Estimated cutoffs were very diversified within all conditions:

Cutoffs Tumour cell lines Primary tumour
low -1.6432 -1.0856 -2.0986 -1.2784 -3.0934 -2.4648
high 1.5682 1.1751 2.4181 1.4216 2.8047 2.1775

Conclusions: Because of signal distribution variety, adaptive cutoffs give better identification of differentially expressed proteins than usually used fixed level. Additionally, it allows for estimation of FDR or p-value.

Acknowledgments: The study was supported by GeCONiI (POIG.02.02.01-24-099/13).

P337**Variability of protein coding and non-coding co-expression patterns**

Katherine Hartmann¹, Michał Seweryn²

¹The Ohio State University, United States, ²University of Łódź, Poland

K. Hartmann <katherine.hartmann@osumc.edu>

M. Seweryn <panpit@gmail.com>

Expression of individual RNAs varies with genetic and environmental conditions. Yet gene function depends on coordinated expression of multiple genes. Here, we study robustness of co-expression between different types of RNAs. We consider a model (Boltzmann-Gibbs distribution), in which the joint probability of expression is given by a hamiltonian with bivariate interactions. Based on an approach called ARACNE (Margolin et al., 2006), we develop an algorithm to perform a tissue specific co-expression analysis, evaluating the frequency with which genes are co-expressed across sub-samples of human individuals. To quantify interactions we use generalized family of Renyi divergence measures, allowing to up- or down-weight rare events (Kotz, Wang, & Hung, 1990; Rempala & Seweryn, 2013). We introduce the concept of a 'genetic relay' - a transcript which 'links' two RNAs lacking a robust co-expression pattern. We apply this to artery, lymphocyte, and blood RNA-sequencing data from the Genotype and Tissue Expression Project.

We show that a limited number of co-expressed pairs are present in an abundant number of sub-samples and that these stable pairs are more likely to be conserved between tissues. Among pairs present in multiple tissues, we find genes important to basic functions (RNA processing, translation, etc). We find non-coding/protein-coding pairs more often than expected and pairs involving ncRNAs to be dynamic (variable between tissues and subjects) - suggesting they respond to external pressures. Our results suggest a central role of ncRNAs in regulating gene expression patterns and highlight the dynamic nature of these interactions. Supported by U01GM092655 from NIGMS.

P338**Determination of high-grade brain tumours internal structure based on magnetic resonance diffusion imaging and signal decomposition to Gaussian mixture model.**

Franciszek Binczyk¹, Cristian Weber^{2,3}, Michael Götz^{2,3}, Bram Stieltjes⁴, Klaus Maier-Hein^{2,3}, Hans-Peter Meinzer², Rafal Tarnawski⁵, Barbara Bobek-Billewicz⁵, Joanna Polanska¹

¹Silesian University of Technology, Institute of Automatic Control, Data Mining Group, Poland, ²Medical and Biological Informatics, German Cancer Research Center, Germany, ³Medical Image Computing, German Cancer Research Center, Germany, ⁴Department of Radiology, University Hospital Basel, Switzerland, ⁵Maria Sklodowska-Curie Memorial Cancer Center and Institute of Oncology, Poland

F. Binczyk <franciszek.e.binczyk@polsl.pl>

J. Polanska <joanna.polanska@polsl.pl>

Aim: The problem of in vivo tissue differentiation is crucial for modern oncological therapy, especially in high-grade brain tumour that contains different types of tissue reacting differently to radio- and chemotherapy. In this study a technique of automatic tumour internal structure differentiation based on Gaussian mixture model decomposition is proposed.

Material and methods: The data set contains 17 DTI's collected before operation on patients diagnosed with glioblastoma multiforme grade IV. The following parameters were collected: apparent diffusion coefficient (ADC), radial diffusivity (RD), fractional anisotropy (FA), relative anisotropy (RA) and free water map (FW). The non-tumour voxels were filtered out. Every signal distribution within tumour voxels was decomposed and unsupervised k-means clustering in 3D space (mean, variance and weight) was performed. Using maximum probability rule on envelope of clustered components the cutoff values were calculated per every diffusion factor. The obtained cutoffs were used to discretize the signal and multivariate unsupervised clustering was performed to detect internal structure of tumour.

Results: The number of univariate clusters and obtained cutoff values (given in brackets) were equal to: ADC 3 clusters (291; 1167), RA 3 (0.39; 0.58), RD 3 (1198; 1555), FA 2 (0.28), FW 2 clusters

(0.36). Multivariate unsupervised clustering resulted in identification of 5 main groups of voxels, which may be treated as: edema, necrosis, tissues of high and low malignancy, and cerebrospinalfluid (CSF).

Conclusions: The developed technique allows for identification of tumour internal structure at satisfactory level.

Acknowledgments: This work was supported by SUT-BKM-524/Rau-1/2014/18 and GeCONiI (POIG.02.03.01-24-099/13).

P340**Maximum Renyi Entropy models with higher order interactions in genomics**

Michael Schwemmer¹, Michal Seweryn²

¹Mathematical Biosciences Institute, United States,

²University of Lodz, Poland

M. Seweryn <panpit@gmail.com>

Modeling biological data is a difficult task. Interactions underlying observed behavior are complex and one rarely has access to variables essential for the process. An approach to developing models of limited data is the Maximum Entropy Principle (MEP). This approach has proven to be invaluable in identifying pairwise interactions which have greatest influence on observed behavior (Granot-Atedgi et al PLoS Comp Biol, 2013 and references therein). The MEP has typically been used with Shannon entropy (SE), which leads to closed form solution. Yet, there is no reason why other entropy functions cannot be used with MEP, thus we focus on Renyi entropy (RE). The difference between using SE and RE is that maximizing the latter with first moment constraints does not force the solution to have a product form. We find that using MEP with RE allows to uncover higher order interactions missed by SE.

As such, we develop a framework for fitting maximum RE distribution to binary data with pairwise interaction constraints. We account for higher order interactions by choosing the order parameter of RE which minimizes difference between observed and theoretical distributions. Next, we present a simulation study and cases when optimizing RE gives significantly better results than SE. Also, we interpret the higher order interactions found by our algorithm. We apply this approach to (1) 1000genomes data, demonstrating examples of linkage disequi-

librium that forms long haplotypes marking recent evolutionary events; (2) the Framingham Heart Study, identifying examples of linkage patterns differing between diseased and healthy populations.

P341

Association Rule Mining for Metabolic Pathway Prediction

Imene Boudellioua¹, Rabie Saidi², Maria Martin², Victor Soloviev¹

¹King Abdullah University of Science and Technology, Saudi Arabia, ²UniProt - European Bioinformatics Institute, United Kingdom

I. Boudellioua <Imene.Boudellioua@kaust.edu.sa>

Prediction of chemical reactions and pathways is among the most challenging problems of systems biology. In this work, we are tackling the problem of metabolic pathway prediction in the context of metabolism. We developed ARBA (Association-Rule-Based Annotator), a system that utilizes machine learning methods, specifically rule mining techniques, to predict pathways associated with protein entries available in UniProtKB. Our system can be used to enhance the quality of automatically generated annotations as well as annotating unknown proteins. Moreover, this system will provide an insight into the conservation of pathways across prokaryotes that differ in their taxonomic classification. ARBA was successfully applied to gain knowledge about pathway annotation type in all UniProtKB-SwissProt entries with manual assertion evidence corresponding to a specified prokaryotic taxon. ARBA presents this knowledge in the form of association rules that takes into account the organism taxonomy and the InterPro signature matches of protein sequences. These rules are then filtered efficiently using the Skyline operator in order to select the best representative rules in terms of several interestingness metrics to effectively minimize false positives as well as eliminating rules generated out of pure randomness. The resulting rules could be used as models to infer pathways for poorly annotated TrEMBL entries. We carried out an experimental study of the performance of ARBA on real datasets representing various prokaryotic taxa to demonstrate the robustness of our system. We found that ARBA achieved an average overall accuracy as high as 99.98%, F-measure of 0.948, precision of 0.977278673, and recall of 0.920848785.

P342

Methods of identification and analysis of errors in shotgun sequencing data

Mateusz Garbulowski¹, Andrzej Polański¹

¹Silesian University of Technology, Poland

M. Garbulowski <mateusz.garbulowski@polsl.pl>

A. Polański <andrzej.polanski@polsl.pl>

Next-generation sequencing methods bring a lot of data containing many nucleotide sequences. Such a huge collection of sequences and information may contain various types of errors. This study presents some of methods which allow to detection and analysis of raw sequencing data quality. To compare the empirical results we apply a model of shotgun sequencing according to real rules of this technique. Proposed model is based on random formation of reads (short fragments of DNA) and is used to comparison with experimental data. As an input we loaded some sequences in .FASTA format file and then we set number and length of reads to manipulate the depth of coverage. As a result from the modeling stage the set of reads were obtained. To check the true data we gained some raw sequencing datasets in .FASTQ format from Sequence Read Archive (SRA) database that is part of the European Nucleotide Archive. Further, we created tools in R environment to graphical presentation of those mistakes within the meaning of: not recognized nucleotides (N), GC content in reads and other nucleotide quality statistics.

Model of shotgun sequencing is a good way to analyze quality of different types of artificial data. Performed analysis shows how to reach the errors representation from raw sequencing data. Presented results demonstrate the quality of sequencing datasets and model data which may help to better recognizing of the main types of errors.

P343**Expert-driven validation, interpretation and functional description of gene sets.**Aleksandra Gruca¹, Marek Sikora², Łukasz Stypka¹¹Institute of Informatics, Silesian University of Technology, Poland, ²Institute of Innovative Technologies EMAG, Poland

A. Gruca <aleksandra.gruca@polsl.pl>

Information and data collected from scientific experiments, published literature, high-throughput experiment technology, and computational analyses are publicly available via number of bioinformatics databases. Therefore, in the research community, there is a constant effort in creating tools and methods that are able to use the information stored in databases for automated interpretation and verification results of experiments performed in the laboratory.

However, frequently the expert who uses such tools is interested in particular process of event related to the research. For example, in cancer experiment searching for gene signature that could be potentially useful for diagnosis, one may look for genes involved with particular biological process or network related to transformation of normal cell into cancer cells.

To address this problem we present the method based on RuleGO[1] algorithm for expert-driven and semi-automated description of gene sets using combinations of Gene Ontology (GO)[2] terms. Presented method allows the expert to influence the process of analysis by providing a set of so called “seed terms” – the GO terms of special interest which should be used as a base for generated description. The user provides a list of genes for interpretation, reference set of genes and combinations of seed GO terms. Depending on user’s choice, the method will:

- verify user’s hypothesis by evaluating provided combinations of seed GO terms,
- generate description on basis of seed GO terms and then widen the rules by adding additional GO terms to their premises,
- generate description on basis of seed GO terms, widen the rules and then generate more rules in order to describe as many genes from analysed group

as possible.

The method is implemented as web application and is available at: www.rulego.polsl.pl

Acknowledgements: The work was supported by National Science Centre (DEC-2011/01/D/ST6/07007)

[1] Gruca A., Sikora M., Polanski A. (2011) RuleGO: a logical rules based tool for description of gene groups by means of Gene Ontology. *Nucleic Acids Res.*, 39(suppl 2), W293-W301

[2] Ashburner et al. (200) Gene Ontology: tool for the unification of biology. *Nat Genet.*, 25(1) 25-29.

P344**Modeling of transcriptional accuracy suggests two-step proofreading by RNA polymerase**Harriet Mellenius¹, Måns Ehrenberg¹¹Uppsala University, Sweden

H. Mellenius <harriet.mellenius@icm.uu.se>

M. Ehrenberg <ehrenberg@xray.bmc.uu.se>

The accuracy of an enzymatic reaction is the ratio of flow of product formation using the correct substrate over other substrates. The high accuracy of transcription by RNA polymerase is achieved through initial selection of substrates before elongation and proofreading selection by polymerase backtracking and transcript cleavage.

We have developed the standard model of transcription to make quantitative predictions of the template dependent transcriptional accuracy. In the commonly used model of transcription, the rate constants of the reaction system are calculated from experimentally determined melting energies for double-stranded DNA and RNA/DNA hybrid, including the effect on base pair stability from adjacent base-pairs by base stacking. Base stacking and experimental studies suggest that the nucleotide incorporation following a mismatch is impaired. Moreover, transcript cleavage always releases a product one nucleotide longer than the number of backward translocations, at the very least a di-nucleotide. Hence, we conclude that a mismatch can be removed also by detection of its destabilizing effect on the next incorporation, and the polymerase gains one extra round of proofreading.

We therefore propose a new scheme for transcriptional fidelity control, where each base is controlled by initial selection and two rounds of proofreading. Furthermore, we predict the maximum discrimination of initial selection and the two proofreading rounds, as calculated from the substrate–template interaction energy predicted by nearest-neighbor parameters.

P345

Identifying tissue specific enhancers by sequence and histone modifications - a machine learning approach

Julia Herman-Izycka¹, Bartek Wilczynski¹

¹University of Warsaw, Poland

J. Herman-Izycka <julia.hermanizycka@gmail.com>

Identifying tissue specific enhancers has been an important field of study in biology for a number of years now. It is necessary for understanding the differentiation of tissues in higher organisms. Many experimental techniques have been used to characterize hundreds of functional enhancers, however given the number of different cell types it is difficult to identify all specific enhancers experimentally.

Computational methods usually base on our knowledge of characteristic features of such DNA fragments, including specific sequence motifs or epigenetic markers, but quality of predictions has been so far limited and in many cases difficult to assess due to the limited knowledge of truly functional sequences in different contexts. Recently, it has been shown that enhancers can be identified in a genome-wide CAGE experiments by searching for short, divergent transcripts originating from both ends of an active enhancer. Recent study by the FANTOM5 project identified using enhancer-RNAs over 40.000 of such putative enhancer positions.

We use machine learning approach to predict tissue specific, active human enhancers. In particular, we train random forest classifier on FANTOM5 enhancer-RNA database as well as experimentally validated sequences from VISTA database. We combine information about histone modifications occurrence in genome fragments with sequence features. We show that computational identification of long regions containing enhancers (from VISTA database), as well as short FANTOM5-enhancers is possible with good accuracy (AUC between 0.7

and 0.9). Classifiers obtained by training on one group of enhancers are also a good predictors for the other.

P346

Reconstruction of clonal trees and tumor composition from multi-sample cancer sequencing data

Mohammed El-Kebir¹, Layla Oesper¹, Hannah Acheson-Field¹, Ben Raphael¹

¹Brown University, United States

M. El-Kebir

<mohammed_el-kebir@brown.edu>

L. Oesper <layla@cs.brown.edu>

B. Raphael <braphael@brown.edu>

Cancer is a disease resulting from somatic mutations that accumulate during an individual's lifetime. The clonal theory of cancer posits that a tumor evolves over time as different descendants of the original founding cell acquire new somatic mutations. Since somatic mutations are typically measured in human solid tumors only at a single time point, when the patient undergoes surgery, the clonal evolution is not directly observed. Thus, one is faced with the problem of inferring the ancestral relationships between cells in a tumor from measurements at one time point. Recent studies that sequence multiple samples of a tumor from the same time point provide additional data to analyze the process of clonal evolution in the population of cells that give rise to a tumor.

We formalize the problem of reconstructing the clonal evolution of a tumor using single-nucleotide mutations as the Variant Allele Frequency Factorization Problem (VAFFP). We derive a combinatorial characterization of the solutions to this problem and show that the problem is NP-complete. We derive an integer linear programming solution to the VAFFP in the case of error-free data and extend this solution to real data with a probabilistic model for errors. We apply the resulting AncesTree algorithm to 22 tumors from three different studies. We find that AncesTree is better able to identify ancestral relationships between individual mutations than existing approaches, particularly in ultra-deep sequencing data when high read counts for mutations yield high confidence variant allele frequencies.

P347**PyDesc: a framework for structural analysis of biopolymers****Tymoteusz Oleniecki¹, Maciej Dziubiński², Grzegorz Firlik³, Agnieszka Mykowiecka⁴, Paweł Daniluk²**

¹College of Inter-Faculty Individual Studies in Mathematics and Natural Sciences, Poland, ²Department of Biophysics, Faculty of Physics, University of Warsaw, Poland, ³Bioinformatics Laboratory, Mossakowski Medical Research Centre, Polish Academy of Sciences, Poland, ⁴Institute of Informatics, Faculty of Mathematics, Informatics and Mechanics, University of Warsaw, Poland

T. Oleniecki <cxtimmy@o2.pl>

P. Daniluk <pawel@bioexploratorium.pl>

PyDesc is a software package developed in Python to provide a framework for analysis of tertiary structures of biopolymers. It handles both proteins and ribonucleic acids as well as their complexes. It also supports trajectories comprising several sets of coordinates for the same molecule (e.g. NMR or MD data). It has been built upon Biopython PDB parsing capabilities[1] with an aim of creating a feature rich environment for all sorts of structure related tasks.

PyDesc is designed around a concept of a contact – a presumed interaction between amino-acid residues, nucleotides or ligands (mers). It can compute contact maps using several built-in contact definitions. These include criteria based on distances, angles and dihedral angles between certain atoms. New criteria can easily be added by the users.

PyDesc has an extensive set of operations on structures and substructures. Users may create new structures by performing set algebraic operations, selecting particular mers (several selection criteria are provided), or using other predefined operations. Local descriptors of structure[2], which were designed to encompass a complete physico-chemical environment of a residue, are also natively supported. It should be noted that substructures in PyDesc may contain several disjoint pieces of backbone and are not limited to a single chain. The most advanced features include searching for a structural motif in a given structure and computing alignments between structures[3].

PyDesc has also a plugin for PyMol which enables easy visualisation of structures, alignments and

contacts.

PyDesc is distributed under GPL and is available from the authors upon request.

Acknowledgements

These studies were supported by the research grant (DEC-2011/03/D/NZ2/02004) of the National Science Centre.

Bibliography

1. Hamelryck, Thomas, and Bernard Manderick. PDB file parser and structure class implemented in Python. *Bioinformatics* 19.17 (2003): 2308-2310.
2. Hvidsten, Torgeir R., Andriy Kryshtafovych, and Krzysztof Fidelis. Local descriptors of protein structure: A systematic analysis of the sequence-structure relationship in proteins using short- and long-range interactions. *Proteins: Structure, Function, and Bioinformatics* 75.4 (2009): 870-884.
3. Daniluk, Paweł, and Bogdan Lesyng. *Theoretical and Computational Aspects of Protein Structural Alignment. Computational Methods to Study the Structure and Dynamics of Biomolecules and Biomolecular Processes.* Springer Berlin Heidelberg, 2014. 557-598.

P348**Power and Limitations of RNA-Seq: Findings from the SEQC (MAQC-III) consortium****Paweł P. Łabaj¹, David P. Kreil¹**

¹Chair of Bioinformatics, Boku University Vienna, Austria

P.P. Łabaj <Pawel.Labaj@boku.ac.at>

We present a multi-centre cross-platform study of the US-FDA MAQC/SEQC-consortium, introducing a landmark RNA-Seq reference dataset comprising 30 billion reads. Several next-generation-sequencing, microarray, and qPCR platforms were examined. The study design features known mixtures, high-dynamic range ERCC-spikes, and a nested replication structure to support a large variety of complementary benchmark metrics. None of the examined technologies can provide a ‘gold standard’, making the built-in truths critical for the development and validation of novel or improved algorithms and data processing pipelines. In contrast to absolute expression-levels, for relative expression measures, good inter-site reproducibility and agreement of across platforms could be achieved with

additional filtering steps. Comparisons with microarrays identified complementary strengths, with RNA-Seq at sufficient read-depth detecting differential expression more sensitively, and microarrays achieving higher rank-reproducibility. At the gene level, comparable performance was reached at widely varying read-depths, depending on the application scenario. On the other hand, even at read-depths >100 million, we find thousands of novel junctions, with good agreement between platforms. Remarkably, junctions supported by only 10 reads achieved qPCR validation-rates >80%. Finally, the modelling approaches for inferring alternative transcripts expression-levels from read counts along a gene can be applied to probes along a gene in high-density next-generation microarrays. We show that this has advantages in quantitative transcript-resolved expression profiling.

P349

Aggrescan3D web server for protein aggregation prediction taking into account protein structure and its dynamic fluctuations

Rafael Zambrano¹, Michal Jamroz², Agata Szczasiuk², Jordi Pujols¹, Sebastian Kmiecik², Salvador Ventura¹

¹Institut de Biotecnologia i Biomedicina and Departament de Bioquímica i Biologia Molecular, Universitat Autònoma de Barcelona, Spain, ²University of Warsaw, Faculty of Chemistry, Poland

S. Kmiecik <sekmi@chem.uw.edu.pl>

S. Ventura <salvador.ventura@uab.es>

Protein aggregation underlies an increasing number of disorders and constitutes a major bottleneck in the development of therapeutic proteins. Our present understanding on the molecular determinants of protein aggregation has crystallized in a series of predictive algorithms to identify aggregation-prone sites. A majority of these methods rely only on sequence. Therefore, they find difficulties to predict the aggregation properties of folded globular proteins, where aggregation-prone sites are often not contiguous in sequence or buried inside the native structure. The AGGRESCAN3D (A3D) server [1] overcomes these limitations by taking into account the protein structure and the experimental aggregation propensity scale from the well-established AGGRESCAN method [2]. Using the A3D server, the identified aggregation-prone

residues can be virtually mutated to design variants with increased solubility, or to test the impact of pathogenic mutations. Additionally, A3D server enables to take into account the dynamic fluctuations of protein structure in solution, which may influence aggregation propensity. This is possible in A3D Dynamic Mode that exploits the CABS-flex approach for the fast simulations of flexibility of globular proteins [3]. The A3D server can be accessed at <http://biocomp.chem.uw.edu.pl/A3D/>

References

- [1] Zambrano R, Jamroz M, Szczasiuk A, Pujols J, Kmiecik S, Ventura S. (2015) AGGRESCAN3D (A3D): server for prediction of aggregation properties of protein structures. (submitted).
- [2] Conchillo-Sole O, de Groot NS, Aviles FX, Vendrell J, Daura X, Ventura S. (2007) AGGRESCAN: a server for the prediction and evaluation of "hot spots" of aggregation in polypeptides. *BMC Bioinformatics*. 8, 65.
- [3] Jamroz M, Kolinski A, Kmiecik S. (2013) CABS-flex: Server for fast simulation of protein structure fluctuations. *Nucleic Acids Res.* 41, W427-31.

P350

Describing tertiary RNA structure with local spatial motifs

Agnieszka Mykowiecka¹, Tymoteusz Oleniecki², Maciej Dziubiński³, Paweł Daniluk³

¹Faculty of Mathematics, Informatics and Mechanics, University of Warsaw, Poland, ²College of Inter-Faculty Individual Studies in Mathematics and Natural Sciences, University of Warsaw, Poland, ³Department of Biophysics, Faculty of Physics, University of Warsaw, Poland

A. Mykowiecka <agnieszka.mykowiecka@mimuw.edu.pl>

P. Daniluk <pawel@bioexploratorium.pl>

It is well known that tertiary structures of biomolecules may contain recurring spatial motifs. In case of protein structures this fact has been exploited by several methods for structure prediction, comparison, etc. The primary aim of this work was to establish a protocol for describing self contained structural blocks of RNA structures which, while preserving crucial local physicochemical interactions between nucleotides, would allow to represent known RNA structures using a limited dictionary of

such small structural elements. Presented method is an adaptation of the formalism of local descriptors of protein structure[1,2]. It is based on contacts between residues which can be presumed to interact with each other. For the purpose of describing RNA structures, the method had to be extended to take into account ions, which are frequently embedded in RNA structure and are crucial to its stability. Our tests indicate that distances between centers of alkaline rings and distances between nucleotides and ions are sufficient to reconstruct known interactions (such as pairing and stacking) and are well preserved in structures containing a similar motif.

7914 descriptors constructed using these two criteria allowed for good coverage (approx. 80%) of a set of chosen 55 RNA structures. The descriptors were then compared and clustered (an adaptation of method previously used for protein descriptors was used). A set of 884 cluster centroids resulted with coverage of 72%, which indicates a relatively high repetition rate of local structural motifs in RNA structures.

Acknowledgements

These studies were supported by the research grant (DEC-2011/03/D/NZ2/02004) of the National Science Centre.

References

[1] Hvidsten, Torgeir R., Andriy Kryshchak, and Krzysztof Fidelis. Local descriptors of protein structure: A systematic analysis of the sequence-structure relationship in proteins using short- and long-range interactions. *Proteins: Structure, Function, and Bioinformatics* 75.4 (2009):870-884.

[2] Daniluk P., Lesyng B., A novel method to compare protein structures using local descriptors *BMC Bioinformatics* 12(1):344.

P351

Solvary: a tool for the study of complex intracellular signalling pathways

Roman Jaksik¹, Krzysztof Puszyński¹

¹Silesian University of Technology, Poland

R. Jaksik <roman.jaksik@polsl.pl>

Study of intracellular signalling processes might

provide invaluable insight into the development of new targeted anticancer therapies. However the complexity of signalling mechanisms requires sophisticated mathematical models in order to create a close representation of the processes involved.

In this study we present a new application - Solvary that simplifies the evaluation of deterministic and stochastic models of the intracellular processes. It allows for simple model declaration in the form of ODE which can be automatically transformed into reaction propensities when necessary. Each variable can be declared as deterministic or stochastic through a simple switch mechanism, and treated accordingly. The simulations are carried out using either stochastic (Gillespie algorithm) or deterministic (RK) methods or a new method which combines those two approaches if some of the model variables are marked as deterministic and some as stochastic. The simulations can be conducted both on a local computer or on a remote computer cluster due to a very efficient and flexible parallel computing module implemented in the application.

We present the usefulness of Solvary through a study of viability control processes in cancer cells. The study is based on a complex model of p53 pathway build of over 50 nonlinear, differential equations and 10 reaction propensities for stochastic variables such as genes state. By controlling the level of either Mdm2 or Ikbα through RNA interference we show in silico how cellular viability can be reduced to the level specified by our biological experiments.

This work was supported by
DEC-2012/05/D/ST7/02072
and BKM/524/Rau1/2014.

P352

A multivariate approach predicts genes that change in DNA methylation patterns over age in the human brain

Behrooz Torabi Moghadam¹, Michal Dabrowski², Bozena Kaminska², Manfred Grabherr¹, Jan Komorowski¹

¹Uppsala University, Sweden, ²Nencki Institute of Experimental Biology, Poland

B.T. Moghadam <behrooz.torabi@icm.uu.se>

J. Komorowski <jan.komorowski@icm.uu.se>

DNA methylation plays a key role in developmental processes, multiple diseases and aging. Detecting

the changes in DNA methylation patterns over the lifetime of an individual can thus be used for different applications, such as the identification of genes and their interactions and what processes they are involved in, as well as for forensic genetics and disease prevention. Since the mechanisms controlled by methylation are complex, we aimed at identifying combinations of multiple CpG sites in the brain that are associated to major age categories: fetus, child, adolescent and adult. To this end, we applied machine-learning methods to two datasets that were used to investigate the methylation status of human brain samples in 27 000 CpG sites. Monte Carlo feature selection produced a list of ranked and significant CpG sites, while rule-based models allowed the identification of particular methylation levels in these significant sites. Moreover, our approach reports combinations of CpG sites, together with their changes in methylation in the form of easy-to-read IF-THEN rules. The identified sites and their combinations were statistically significant, and the rule models were validated against each other. The strongest interactions of the features in the rules identified with Ciruviz highlighted genes involved in brain and age related processes

P353

Improved chromatin segmentation with a probabilistic model for read counts

Alessandro Mammana¹, Ho-Ryun Chung¹

¹Max Planck Institute for Molecular Genetics, Germany
A. Mammana <mammana@molgen.mpg.de>

A central question in biology pertains to the establishment and maintenance of different phenotypes adopted by cells of a multicellular organism with a constant genotype. In part this diversification of cells is driven by the adoption of distinct epigenomes, e.g. the localization of histone modifications along the genome. Chromatin immunoprecipitation followed by sequencing (ChIP-seq) is a common experimental approach to generate genome wide maps of histone modifications.

The segmentation of epigenomes into chromatin states collapses the ChIP-seq tracks and provides an abstract view on the multi-dimensional data. A chromatin state is a recurrent pattern in the abundances of a given set of histone modifications, possibly related to a particular biological function.

Chromatin segmentation aims at explaining the observed epigenomic data as a long sequence of a small number of hidden chromatin states. There are a number of computational methods for this task, but they have important shortcomings. Here we propose a novel segmentation algorithm. Owing to an accurate probabilistic model for the read counts, our method provides a useful annotation for a considerably larger portion of the genome, shows a stronger association with validation data, and yields more consistent predictions across replicate experiments compared to existing methods.

P354

Bioinformatic analysis of ChIP-seq data on epigenetic response to salt stress in *Arabidopsis thaliana*

Anna Macioszek¹, Anna Fogtman², Aleksandra Kwiatkowska², Maciej Kotliński², Roksana Iwanicka-Nowicka², Andrzej Jerzmanowski², Marta Kobłowska², Bartek Wilczyński¹

¹Faculty of Mathematics, Informatics and Mechanics at University of Warsaw, Poland, ²Institute of Biochemistry and Biophysics at Polish Academy of Sciences, Poland

A. Macioszek <a.macioszek@students.mimuw.edu.pl>

Epigenetic mechanisms play important role in gene expression regulation, including response to stress and other external factors. Acetylation of lysine 16 in histone H4 is yet poorly understood modification known to play important role in response to salt stress in *Arabidopsis thaliana*. We conducted ChIP-seq experiments to examine how pattern of H4K16ac changes in response to stress and we combined the results with results of microarray experiments to see how this modification is related to gene expression. Despite the existence of multiple software packages for ChIP-seq analysis, this task can be challenging, especially in case of wide and dynamically changing histone modifications such as H4K16ac. Here we present various approaches we used and discuss their advantages and limitations.

P355**Knot position within a protein, a DCA approach****Aleksandra Jarmolińska¹, Joanna I Sulkowska¹**¹Center of New Technologies, Poland

A. Jarmolińska <dzarmola@gmail.com>

J.I. Sulkowska <jsulkows@gmail.com>

A knot within a protein is a well established fold. Still, only about 1% of structures in the Protein Data Bank database have such a topology. While still being a lot of data it is sometimes not enough to shine the light on all of their mysteries. That's where the mathematical methods come in.

Sequence conservation of even the simplest of knots - that is the trefoil - is not very high, so it's hard to directly assign function to any particular position (binding site excluded) or find how the native, non trivial topology is encoded in the sequence. It is known that those proteins can self tie, the mechanics behind this process remain a mystery.

Thanks to the mean field Direct Coupling Analysis we are able to find strongly co-evolving pairs of amino acids, which is especially important in the study of such a non-trivial topology. Through application of DCA to different families which conserved trefoil knot, we study the internal dependencies within the structure. As one step on the way to determine which interactions allow the protein folding to end in a knot, we have analyzed the knot's location (i.e. positions of its ends).

We have found that groups of amino acids adjacent to knot's ends have noticeably higher average Direct Information scores (obtained from DCA) measuring the 'strength of their relationship' to other positions within the protein, when compared with other fragments of similar length. This shows that knot location is not random and, subsequently, gives us hope that it is possible to unravel computationally the mystery of knots surviving the evolutionary pressure.

P356**Identification of potential protelomerases and their target sites in publicly available genomes. Phylogenetics analysis****Lukasz Kozlowski¹, Janusz Bujnicki¹**¹The International Institute of Molecular and Cell Biology, Poland

L. Kozlowski <lukaskoz@genesilico.pl>

Protelomerase is a unique resolvase enzyme that is capable of recognition, cleavage and ligation of linear chromosomes or replicons with covalently closed hairpin termini (telomeres). Most of known protelomerases were detected in pathogens; e.g., *Borellia burgdorferi* or phages. Due to the degeneracy of the recognition site of protelomerase, a high recombinant divergence of mobile elements is possible. To date, only a few protelomerases with their target sites are known.

In this study, we have made a systematic screen of publicly available genomes (bacteria and phages) in order to identify new members of the protelomerase family based on sequence and structure homology of known members. First, out of more than 3,300 genomes stored in NCBI, we filtered out those which are linear (over 75,000 fragments of DNA). Next, all potential open reading frames were extracted (in all six frames). In order to find sequence homology, HHsearch was used. For the most promising candidates, homology models were built and compared to known structures with respect to the conservation of the catalytic site and other functionally important parts. Moreover, to avoid potential false positives, we designed a special algorithm for the recognition of potential protelomerase target sites. Based on the imperfect palindromic target site of TelN from *Escherichia coli* phage N15, a special set of rules was deduced. Only genomes containing both potential protelomerase homologue and the target site were considered in further studies. This resulted in the identification of 1422 potential protelomerases.

P357**KnotProt: a database of proteins with knots and slipknots****Joanna Sulkowska¹**¹University of Warsaw, Faculty of Chemistry, Poland

J. Sulkowska <jsulkowska@chem.uw.edu.pl>

The protein topology database KnotProt, <http://knotprot.cent.uw.edu.pl/> [1], collects information about protein structures with open polypeptide chains forming knots or slipknots. The knotting complexity of the catalogued proteins is presented in the form of a matrix diagram that shows users the knot type of the entire polypeptide chain and of each of its subchains. The pattern visible in the matrix gives the knotting fingerprint of a given protein and permits users to determine, for example, the minimal length of the knotted regions (knots' core size) or the depth of a knot, i.e. how many aminoacids can be removed from either end of the catalogued protein structure before converting it from a knot to a different type of knot. In addition, the database presents extensive information about the biological functions, families and fold types of proteins with non-trivial knotting. As an additional feature, the KnotProt database enables users to submit protein or polymer chains and generate their knotting fingerprints.

1. Jamroz M, Niemyska W, Rawdon EJ, Stasiak A, Millett KC, Sulkowski P, Sulowska JI (2014) KnotProt: a database of proteins with knots and slipknots, *Nucleic Acids Research*, DOI: 10.1093/nar/gku1059

P358

How progressive cis element combinations classify conserved orthogonal plant circadian transcriptional modules

Sandra Smieszek¹, Bartłomiej Przychodzen²

¹Royal Holloway, United Kingdom, ²Cleveland Clinic, United States

S. Smieszek <S.Smieszek@rhul.ac.uk>

We aimed to test the proposal that progressive combinations of multiple promoter elements acting in concert of may be responsible for the full range of phases observed in plant circadian output genes. In order to allow reliable selection of informative phase grouping of genes for our purpose, intrinsic cyclic patterns of expression were identified using a novel, non-biased method for the identification of circadian genes. Our non-biased approach identified two dominant, inherent orthogonal circadian trends underlying publicly-available microarray data from plants maintained in constant conditions. Furthermore, these trends were highly

conserved across several plant species. Four phase-specific modules of circadian genes were generated by projection onto these trends and, in order to identify potential combinatorial promoter elements that might classify genes in to these groups, we used a random forest pipeline which merged data from multiple decision trees to look for presence of element combinations. We identified a number of regulatory motifs which aggregated into coherent clusters capable of predicting the inclusion of genes within each phase module with very high fidelity and these motif combinations changed in a consistent, progressive manner from one phase module group to the next, providing strong support for our hypothesis.

P359

SimRNA: a coarse-grained method for RNA folding simulations and 3D structure prediction.

Michał Boniecki¹, Grzegorz Łach¹, Konrad Tomala¹, Wayne Dawson¹, Paweł Łukasz¹, Tomasz Soltyski², Kristian Rother¹, Janusz Bujnicki¹

¹International Institute of Molecular and Cell Biology, Poland, ²Institute of Molecular and Cell Biology, Poland
M. Boniecki <mboni@genesilico.pl>

J. Bujnicki <iamb@genesilico.pl>

The molecules of the ribonucleic acid (RNA) perform a variety of vital roles in all living cells. Their biological function depends on their structure and dynamics, both of which are difficult to experimentally determine, but can be theoretically inferred based on the RNA sequence. We have developed a computational method for molecular simulations of RNA, named SimRNA.

SimRNA is based on a coarse-grained representation of a nucleotide chain, a statistically derived energy function, and Monte Carlo methods for sampling of the conformational space. The backbone of RNA chain is represented by P and C4' atoms, whereas nucleotide bases are represented by three atoms: N1-C2-C4 for pyrimidines and N9-C2-C6 for purines. Despite the bases being represented by only three atoms, other atoms can be implicitly taken into account in terms of the excluded volume. All base-base interactions were modeled using discrete three-dimensional grids built on local systems of coordinates.

All terms of the energy function used were derived from a manually curated database of crystal RNA structures, as a statistical potential. Sampling of the conformational space was accomplished by the use of the asymmetric Metropolis algorithm coupled with a dedicated set of moves. The algorithm was embedded in either a simulated annealing or replica exchange Monte Carlo method. Recent tests demonstrated that SimRNA is able to predict basic topologies of RNA molecules with sizes up to about 50 nucleotides, based on their sequences only, and larger molecules if supplied with appropriate distance restraints. The user can specify various types of restraints, including restraints on secondary structure, distance and position.

SimRNA can be used for systems composed of several chains of RNA. It is also able to fold/refine structures with irregular (non-helical) geometry of the backbone (RNA pseudo-knots, coaxial stacking, bulges, etc.). As SimRNA is based on folding simulations, it also allows for examining folding pathways, getting an approximate view of the energy landscapes, and investigating of the thermodynamics of RNA systems.

P360

Cartesian Rotamers Library for protein structure prediction.

Aleksandra Dawid¹, Dominik Gront¹

¹Laboratory of Theory of Biopolymers, Faculty of Chemistry, University of Warsaw, Poland

A. Dawid <aleksandra.dawid@student.uw.edu.pl>

One of the most important elements in protein folding is appropriate packing of amino acid side chains. They do not have free mobility, because prefer distinct conformations according to physicochemical first principles. These low energy side chain conformations are usually called rotamers. The concise description of all side-chain conformational preferences defines a library of rotamers. It is generally prepared using theoretical calculations and the empirical observation that the side chains of amino acids in protein structures avoid most of the available conformational space and appear frequently as clusters in chi-angle space.

In this work we present a novel library of coarse-grained (CG) rotamers, defined in the Cartesian space. We decided to use a continuous-rotamer

model. For each CG virtual pseudo atom representing an amino acid side chain, its the centre of mass locations are described by a three-dimensional Gaussian distribution.

The input conformations were derived from a large set of non-redundant high quality structures (PISCES, identity cutoff is 60%, the resolution cutoff is 1.6 Angstroms) and successfully correlated with a small database of unique crystal structures of very high quality (Top500, atomic resolution 1.0 Angstroms).

Parameters of the distributions were estimated using the calc::clustering module from Bioshell package. Resulting rotamers library may be utilised by CABS and Rosetta packages for coarse-grained modelling of protein structures.

P361

Tree-generalized hypergeometric score for detection of drug resistance-associated mutations

Michał Woźniak¹, Limsoon Wong², Jerzy Tiuryn¹

¹The Faculty of Mathematics, Informatics and Mechanics, University of Warsaw, Poland, ²School of Computing, National University of Singapore, Singapore
M. Woźniak <m.wozniak@mimuw.edu.pl>

Drug resistance in bacterial pathogens is an increasing problem that stimulates research. However, still our understanding of drug resistance mechanisms remains incomplete. One promising approach to deepen our understanding of drug resistance mechanisms is to use whole-genome sequences to identify genetic mutations associated with drug resistance phenotypes in bacterial strains.

In this work, we present a new statistical measure, called tree-generalized hypergeometric (TGH) score, for detection of drug resistance-associated mutations in bacterial strains. This score generalizes the standard hypergeometric test by using a phylogenetic tree to capture the dependencies between the strains in question. When the input phylogenetic tree is flat our method is equivalent to the standard hypergeometric test. The presented scoring scheme is a part of the GWA-MAR pipeline we have developed for detection of drug resistance-associated mutations.

In order to test our approach we run it on two

datasets for *M. tuberculosis*. The first dataset consist of genotype and phenotype data we have collected from publications for a set of 173 fully sequenced *M. tuberculosis* strains and 10 commonly used drugs. The second dataset has been publicly deposited by The Broad Institute in the Tuberculosis Drug Resistance Mutation Database. Our computational experiments show that our method is capable of identifying drug resistance-associated mutations more accurately than standard hypergeometric test, as well as other methods which do not incorporate phylogenetic information. Applying our methodology we have identified some novel putative drug resistance-associated mutations.

P362

Measuring information transmission from single-cell heterogenous dynamical responses

Tomasz Jetka¹, Tomasz Winarski², Michał Komorowski¹

¹Institute of Fundamental Technological Research, Poland, ²Medical University of Warsaw, Poland
T. Jetka <t.jetka@gmail.com>

All biological organisms need to sense and response to their environment. At the level of single cells, surface receptors convert extracellular cues into activation of transcription factors that control cellular decisions. A considerable unresolved issue is how information about ligand binding is encoded into nuclear activity of the transcription factors. A growing number of studies supports the hypothesis that this is achieved by temporal regulation of their activities. The current challenge is to recognise the features of temporal activity profile that represent information about a given stimulus.

A natural strategy to decipher this temporal coding is to scan cellular responses across a range of considered stimuli and identify most sensitive features of temporal profiles. Methods however to quantify sensitivity and information capacity at the single cell level, where stochastic effects play a major role, are virtually missing. We have developed a statistical framework to measure information of cellular outcomes from time-resolved, single cell, heterogeneous responses.

We use the method to analyse nuclear translocation of the NF- κ B transcription factor upon TNF

stimulation in mouse embryonic fibroblasts. We identified how the information capacity of the system changes with inclusion of time series data and indicate the essential features of the nuclear NF- κ B temporal profile. Our method provides essential methodological advancement needed to gain understanding how temporal activity profiles encode information about a given stimulus.

P363

Automated predication of drinking categories in monkeys undergoing chronic alcohol self-administration

Alex Salo¹, Steven Gonzales², Kathy Grant², Erich Baker¹

¹Baylor University, United States, ²Oregon Health & Science University, United States
E. Baker <Erich_Baker@Baylor.edu>

We have recently demonstrated that the non-human primate model provides robust alignment to drinking categories reflective of human drinking populations. During animal protocols that provide open-access to ethanol we can quantitate Very Heavy Drinking (VHD), Heavy Drinking (HD), Binge Drinking (BD), and Low Drinking (LD) individuals based on consumption patterns over 12 to 18 months. Here, we propose a process to predict categorical behavior during a pre-open access drinking induction period, where ethanol consumption is strictly controlled. Data collected during this episodic and highly structured drinking period is standardized across numerous animal cohorts and collected as part of The Monkey Alcohol and Tissue Research Resource (MATRR). This resource allows us to consolidate numerous animal cohorts and leverage a wide range of animals and data attributes. These factors include behavioral characteristics, including alcohol consumption patterns, sleeping patterns, food and water intake, individual blood ethanol concentrations, hormone levels, and molecular indicators among others. Sinusoidal regression models of ethanol consumption were used to determine sex differences using the autoregressive integrated moving average model, while several model classifiers were employed to examine Gaussian discriminant analysis on the entire set of collected animal data. We are able to demonstrate sex differences with high accuracy using periodic ethanol consumption and offer a range of potential models for the predictive categorical outcome of

animals using induction phase drinking data.

P364

A pilot study for the construction of Korean-specific exome-variome database

Young Park¹, Sunhye Park¹, Yoon Cho², Kiejung Park¹, Insong Koh¹

¹Hanyang University, The Republic of Korea, ²Hallym University, The Republic of Korea

K. Park <kjpark63@gmail.com>

I. Koh <insong@hanyang.ac.kr>

Human genome researches have been main topics in genomics. International big projects, including HapMap project, 1000 genomes project, ENCODE project, IHEC project, have been progressing, which can be reference information for human genome studies. However, more specific reference sets have been required for more specific studies. Korean genome references have been also required and their corresponding studies, including a few reference genomes, Korean SNP databases based on SNP chips, Korean CNV databases based on CNV chips, have been conducted. And a few Korean whole genome projects have been being carried out.

We have analyzed Korean variation information with Korean exomes. Using read data of 100 Korean exome from KNIH, we have mapped each exome data of them on hg19, collected them, and extracted SNP and indel information. After getting initial SNPs and indels with our own analysis pipeline, we have masked dbSNP, where the variation information of 1,000 genomes are included, to get Korean specific variations. We have compared the analysis results with quality values and MAF distribution, through repeated application of variant calling programs with several option values. We have constructed a Korean exome variation pilot-database with a browsing interface. Currently, the Korean-specific exome-variomes shows similar quality to dbSNP, in spite of lacking a lot of variations due to the small size of our samples. The database will be contributed to Korean exome-based variation studies. And it will be extended to larger exome samples and to Korean whole genome scale with more precision

P365

Onctopus: A New Model for Subclonal Composition Reconstruction

Linda K. Sundermann¹, Amit G. Deshwar², Quaid Morris², Gunnar Rätsch³

¹Bielefeld University, Germany, ²University of Toronto, Canada, ³Memorial Sloan Kettering Cancer Center, United States

L.K. Sundermann <lsunderm@cebitec.uni-bielefeld.de>

Cancer samples are often genetically heterogeneous, harboring subclonal populations with different mutations. Information about copy number variations (CNVs) or simple somatic mutations (SSMs; i.e., single nucleotide variants and small indels) in the subpopulations can help to identify driver mutations or to choose targeted therapies. As methods that analyze individual cells with the help of fluorescent markers or single cell sequencing still have various drawbacks (for instance, limited number of features or high cost), bulk tumor sequencing is often used.

Recently, several methods that attempt to infer the genotype of subclonal populations using either CNVs, SSMs, or both have been published. Here, we present Onctopus, a new method to reconstruct the subclonal composition of bulk tumor samples in terms of SSMs and CNVs, using information about read depth and variant count data of heterozygous germline SNPs and SSMs, as well as information about segments affected by CNVs. We define a subclonal lineage as the set of subclonal populations that contains a subclonal population and all its descendants. We model the tumor as consisting of a mixture of these lineages where each comprises a characteristic set of CNVs and SSMs. Onctopus is designed to infer a partial order on these lineages while also simultaneously phasing SSMs and SNPs whose copy number is altered by CNVs. Currently, we are refining our model and implementation and compare it to similar tools on simulated data, plan to test it on real data sets and validate it with single cell sequencing data.

P366**Computational detection of DNA double-stranded breaks with nucleotide resolution using deep sequencing data**

Norbert Dojer¹, Abhishek Mitra¹, Yea-Lih Lin², Anna Kubicka³, Magdalena Skrzypczak³, Krzysztof Ginalski³, Philippe Pasero², Maga Rowicka¹

¹University of Texas Medical Branch, United States,

²Institut de Genetique Humaine, CNRS, France,

³University of Warsaw, Poland

N. Dojer <dojer@mimuw.edu.pl>

M. Rowicka <merowick@utmb.edu>

Double-stranded DNA breaks (DSBs) are a dangerous form of DNA damage. The damage to both DNA strands precludes the straightforward use of the complementary strand as a template for repair, resulting in mutagenic lesions. Despite many studies on the mechanisms of DSB formation, our knowledge of them is very incomplete. A main reason for our limited knowledge is that, to date, DSB formation has been extensively studied only at specific loci but remains largely unexplored at the genome-wide level.

We recently developed a method to label DSBs in situ followed by deep sequencing (BLESS), and used it to map DSBs in human cells [1] with a resolution 2-3 orders of magnitude better than previously achieved. Although our protocol detects free ends of DNA with extreme single nucleotide precision, the inference of original positions of DSBs remains challenging. This problem is due to unavoidable sequencing of DNA repair intermediaries (end resection), which effectively lowers our detection resolution by several orders of magnitudes. Another challenge is that DSBs are rare events and sequencing signal originating from them is easily overpowered by background signal, such as copy number variation.

Here, we show how DSBs can be detected computationally with nucleotide resolution. First, we learnt characteristic sequencing read pattern in the vicinity of a DSB using experimental data with DSBs induced in pre-determined positions. On these data, by scanning the genome with our pattern, we were able to detect DSB locations with 2nt positional accuracy and precision of over 90%. Thus derived pattern was then applied to detection of DSBs in typical experiments, where DSB positions are not

known a priori. We used other data known to correlate with DSB locations to confirm high quality of our predictions. Finally, we analyzed read patterns in larger areas surrounding DSBs and their relation to the corresponding DNA damage and repair mechanisms.

[1] Crosetto N, Mitra A, Silva MJ, Bienko M, Dojer N, Wang Q, et al. Nucleotide-resolution DNA double-strand break mapping by next-generation sequencing. *Nat Methods*. 2013;10(4):361-5.

P367**Correcting for Cryptic Relatedness in Genome-Wide Association Studies**

Prajwal Devkota¹, Bonnie Kirkpatrick¹, Susan Blanton², Alexandre Bouchard-Cote³

¹Computer Science, University of Miami, United States,

²Hussman Institute of Human Genomics, University of Miami, United States, ³Statistics, University of British Columbia, Canada

B. Kirkpatrick <bbkirk@cs.miami.edu>

Although individuals in a genome wide association study might not be related to each other, there can exist a distant relationship between these individuals. This cryptic relationship violates the assumption regarding the independence of the subject genomes, causing results to be both false positive and false negative. Hence we present a method to correct these cryptic relationships.

We start by accurately detecting distant relationship using an expectation maximization (EM) algorithm for identity coefficients (a refinement of kinship coefficient). After this, we compute the kinship coefficient and apply a kinship-corrected association test.

We analyze and show that genome simulated from Wright-Fisher pedigree, our approach converges quickly and is accurate requiring relatively small number of sites. To assess the kinship-corrected association test, we simulated individuals from deep pedigrees and drew one site which recessively determined the disease status. Once we estimated the kinship coefficient through our method, we kinship-adjusted test, where the results were favorable to our method compared to the state-of-the-art covariance-based approach.

Use of our method to find cryptic relationships

and for corrected association tests is advantageous, because it is easy to interpret through the use of identity states as latent variables, and its results provides state-of-the-accuracy when compared to models that only consider kinship coefficients.

P368

Mathematical model of human gene regulatory network

Junil Kim¹, Jeong-Rae Kim², Seong-Jin Kim¹

¹CHA University, The Republic of Korea, ²University of Seoul, The Republic of Korea

J. Kim <jikim@uos.ac.kr>

S. Kim <kimsj@cha.ac.kr>

The cellular state is known to be determined by the gene regulatory network. In order to reveal the dynamics of human cellular state, we constructed a mathematical model of the human gene regulatory network (GRN) based on a Boolean network model. We first surveyed the human gene regulation information from hundreds of literatures and then constructed the human GRN composed of 241 nodes and 465 links. To reduce the computation time, we identified the core human GRN composed of 73 nodes and 195 links by removing leaf nodes. Second, we obtained 80 tissue attractors composed of 69 Boolean variables from public microarray data [Cell, Vol. 122, 947-956, Sep. 23, 2005]. Each Boolean variable represents each node in the core human GRN. Third, we obtained GRN attractors by simulating the core human GRN using Boolean network model based on a simple rule without any weight. From the simulation, we found that the average of Boolean values in the 80 tissue attractors is positively correlated with the average of Boolean values in the attractors based on the simple rule. Lastly, we constructed 80 Boolean network models optimized for representing each tissue attractor by evolving weights of Boolean rules. We adopted genetic algorithm (GA) to optimize weights of Boolean rules.

Acknowledgement

This work was supported by the Bio-Synergy Research Project (NRF-2012M3A9C4048735) and NRF-2012R1A1A2007188 of the Ministry of Science, ICT and Future Planning through the National Research Foundation, Korea.

P369

Inferring direction of replication fork and mechanism of DNA damage using sequencing data

Maga Rowicka¹, Norbert Dojer¹, Ji Li¹, Yea-Lih Lin², Magdalena Skrzypczak³, Anna Kubicka³, Krzysztof Ginalski³, Philippe Pasero⁴

¹University of Texas Medical Branch at Galveston, United States, ²IGH, Centre National de la Recherche Scientifique, France, ³Centre of New Technologies, University of Warsaw, Poland, ⁴IGH Centre National de la Recherche Scientifique, France

M. Rowicka <merowick@utmb.edu>

N. Dojer <nodojer@utmb.edu>

Double-stranded DNA breaks (DSBs) are a genotoxic form of DNA damage. The damage to both DNA strands precludes the straightforward use of the complementary strand as a template for repair, resulting in mutagenic lesions. Despite many studies on the mechanisms of DSB formation, our knowledge of them is very incomplete. A main reason for our limited knowledge is that, to date, DSB formation has been extensively studied only at specific loci but remains largely unexplored at the genome-wide level. We recently developed a method to label DSBs in situ followed by deep sequencing (BLESS), and used it to map DSBs in human cells [1] with a resolution 2-3 orders of magnitude better than previously achieved.

There are many factors inducing DSBs, including replication stress, oxidative stress and irradiation. Most of them cause two-ended DSBs (having two free ends of DNA), the only exception is replication stress which usually induces one-ended DSBs (caused by replication fork stalling and collapse). We use this observation to infer DSBs resulting from replication stress and to analyze chromatin context and sequence features related to replication stress-induced DSBs. Moreover, we show how to reconstruct the direction of replication fork movement from BLESS-Seq read pattern. We apply this concept to infer replication domain boundaries for several cell lines and conditions and to analyze how they change upon treatments and vary between cell lines.

References:

[1] Crosetto N, Mitra A, Silva MJ, Bienko M, Dojer N, Wang Q, et al. Nucleotide-resolution DNA double-strand break mapping by next-generation sequencing. *Nat Methods*. 2013;10(4):361-5.

P370

Using cell lines and patient samples to improve patients' drug response prediction

Cheng Zhao^{1,2}, Ying Li^{2,3}, Benjamin Haibe-Kains³, Anna Goldenberg^{1,2}

¹University of Toronto, Canada, ²Hospital for Sick Children, Canada, ³University Health Network, Canada

C. Zhao <cheng.zhao@mail.utoronto.ca>

B. Haibe-Kains <benjamin.haibe.kains@utoronto.ca>

A. Goldenberg <anna.goldenberg@utoronto.ca>

Recent advancements in high-throughput technologies facilitated collection of a large number of transcriptionally profiled cell lines with drug response measured for thousands of drugs. The computational challenge now is to realize the potential of this data in predicting the patients' response to these drugs in the clinic. Geeleher et al (2014) proposed to predict patient response directly from cell lines using before-treatment gene expression data. In our work, we examine the spectrum of prediction of patient response directly from cell lines (C2P), cell lines and patients combined (CP2P) and patients only (P2P). The key to C2P and CP2P models is to properly combine cell line and patient mRNA transcription levels. We used surrogate variable analysis to normalize cell lines for different tissue types and when combining the cell line and patients. We tested 21 supervised prediction methods in 4 drugs for which clinical trial data was available: bortezomib, erlotinib, docetaxel and epirubicin. We came to several interesting conclusions: 1) CP2P models are either on par or most often significantly outperform C2P and P2P models even if hundreds of patients are available; 2) using L1000 (1000 cancer genes that can capture variation in all genes) performs worse than doing feature selection (such as 1000 minimum redundancy maximum relevance features); 3) SVM with a linear kernel and ridge models were consistently among the top performing models.

P371

SV-Bay: structural variant detection in cancer genomes using a Bayesian approach with correction for GC-content and read mappability

Daria Iakovishina¹, Mireille Regnier¹, Valentina Boeva², Emmanuel Barillot³

¹INRIA projet AMIB, Ecole Polytechnique, France,

²Institut Curie, Centre de Recherche, France, ³Institut Curie, Inserm U900, Mines ParisTech, France

D. Iakovishina <yakovishinad@gmail.com>

V. Boeva <Valentina.Boeva@curie.fr>

Motivation: Whole genome sequencing of paired-end reads can be applied to characterize the landscape of large somatic rearrangements of cancer genomes. Several methods for detecting structural variants with whole genome sequencing data have been developed. So far, none of these methods has combined information about abnormally mapped read pairs connecting rearranged regions and associated copy number changes using GC-content and information about mappability of the region. Our aim was create a computational method that could use both types of information, i.e., normal and abnormal reads and takes into account all possible errors that can be caused by both sequencing machines and mapping tools, and demonstrate that by doing so we can highly improve both sensitivity and specificity rates of structural variant prediction.

Results: We developed a computational method, SV-Bay, to detect structural variants from whole genome sequencing mate-pair or paired-end data using a probabilistic Bayesian approach. This approach takes into account depth of coverage by normal reads and abnormalities in read pair mappings. To estimate the model likelihood, SV-Bay considers GC-content and read mappability of the genome, thus making important corrections to the expected read count. For the detection of somatic variants, SV-Bay makes use of a matched normal sample when it is available. We validated SV-Bay on simulated datasets and an experimental mate-pair dataset for the CLB-GA neuroblastoma cell line. The comparison of SV-Bay with several other methods for structural variant detection demonstrated that SV-Bay has better prediction accuracy both in terms of sensitivity and false positive detection rate.

P372**HapIso : An accurate method for the haplotype-specific isoforms reconstruction from long single-molecule reads****Serghei Mangul¹, Farhad Hormozdiari¹, Elizabeth Tseng², Alex Zelikovsky³, Eleazar Eskin¹**¹University of California, Los Angeles, United States,²University of Washington, United States, ³GSU, United States

S. Mangul <serghei@cs.ucla.edu>

Sequencing of the RNA cells provides exceptional possibility to study the individual transcriptome landscape and determine allelic expression ratios. Multi-kilabase reads delivered by the single-molecule protocols are within the size distribution of the most transcripts allowing to sequence full-length haplotype isoforms in a single pass providing clear discrimination of the reads into two parental haplotypes. While the read length of the single-molecule protocols is high enough to bridge the single nucleotide variation loci, the relatively high error rates limit the ability to accurately detect the genetic variants and assemble them into the haplotype-specific isoforms. In this paper, we present HapIso (Haplotype-specific Isoform Reconstruction), a method able to tolerate the relatively high error-rate of the single-molecule platform and discriminate the isoform reads into the parental alleles. Phasing the reads according to the allele of origin allows our method efficiently distinguish between the read errors and the true biological mutations. HapIso uses k-means cluster algorithm aiming to group the reads into two meaningful clusters maximizing the similarity of the reads within cluster, and minimizing the similarity of the reads from different clusters. Each cluster corresponds to the parental haplotype. We use family pedigree information to examine the assembled haplotype isoforms to follow rules of Mendelian inheritance. Short read RNA-Seq reads are used to assess the inter platform concordance of the detected heterozygous loci. Experimental validation suggests that HapIso is able to tolerate the relatively high error-rate and accurately discriminate the reads into the paternal and maternal alleles of the isoform transcript.

P373**Analysis of rhythms using R: Chronomics Analysis Toolkit (CAT)****Lee Gierke Cathy¹**¹University of MN, United States

L.G. Cathy <leegi001@umn.edu>

Aim is to examine the Chronomics Analysis Toolkit (CAT), a freely available package written in R, which performs analysis of periodic rhythms in time series. It is especially suited to the often scarce, frequently noisy, biological data.

A network of biological oscillators modulates genetic, molecular, physiological, and behavioral rhythms. Characterizing rhythmicity has contributed to our understanding of the role oscillator mechanisms play in organizing these complex systems. CAT is a flexible, robust, and open access computational suite, providing visualization tools as well as quantitative assessment, by cosinor, of mean, amplitude and phase at an assumed period, with a measure of uncertainty for each parameter.

Visualization tools allow inspection of raw data and smoothed data; an actogram at any selected period; graphs of autocorrelation and crosscorrelation r values; and a periodogram. Each is customizable.

Fast Fourier transform (FFT) has long been available to identify rhythms, but it requires equidistant data, whereas Cosinor can be performed on non-equidistant data, making it a robust tool for analyzing biological data. When data is equidistant, Cosinor and Fourier analysis produce equivalent results. Cosinor, however, has an additional advantage. Fourier periodogram is limited to discrete Fourier frequencies. The Cosinor can calculate frequencies intermediate to the Fourier frequencies. CATCosinor displays plots of the data, and pertinent parameters: MESOR, Amplitude, Phase and P.

Where data is non-stationary, CATCosinor can be used to analyze progressive segments through the data, identifying changing rhythm dynamics over time. A heat map can be generated of Amplitudes over time and frequency.

A library of easy to use R scripts demonstrate CAT usage.

P374**A Bayesian model for the number of motif hits in finite double-stranded DNA sequences**Wolfgang Kopp¹, Martin Vingron¹¹Max Planck Institute for Molecular Genetics, Germany
W. Kopp <kopp@molgen.mpg.de>

Transcription factors (TFs) play a crucial role in gene regulation by binding to TF binding sites (TFBSs), which are usually enriched in promoter or enhancer regions. Thus, one may hypothesize that sequences which exhibit an overrepresentation of TFBSs are actually bound and therefore functionally important. In order to determine statistical enrichment of TFBSs in a sequences, we propose a novel Bayesian model for the number of motif hits in double-stranded random DNA sequences which we assume to be distributed according to a higher order Markov model. We have compared our model with the binomial distribution and a compound Poisson model, which where previously proposed for that purpose. Especially in the regime where the Poisson assumption is violated for the latter two models the Bayesian model yields better results.

P375**Novel inhibitors of ErmC' methyltransferase responsible for resistance to MLSB antibiotics**Ilona Foik¹, Irena Tuszyńska¹, Marcin Feder², Elżbieta Purta¹, Janusz Bujnicki¹¹International Institute of Molecular and Cell Biology, Poland, ²Adamed Sp. z o.o., Poland

I. Foik <idomagala@genesilico.pl>

I. Tuszyńska <irena@genesilico.pl>

J. Bujnicki <iamb@genesilico.pl>

Bacterial infectious diseases became a public concern due to the growing resistance of these pathogens to clinically used antibiotics. Bacteria have developed antibiotic resistance through numerous mechanisms, one of which is the reduction of antibiotic's affinity to the target molecule caused by enzymatic alteration of the target. In erythromycin-resistance bacteria, the N6 position of A2058 in 23S rRNA is mono- or dimethylated by Erm methyltransferase (MTase). This modification results in cross-resistance to macrolides, lincosamides and streptogramin B (MLSB phenotype). Thus far, no inhibitors of Erm MTases have

been identified or designed that could effectively abolish bacterial resistance in vivo. We attempted to sensitize the resistant bacteria to antibiotics by inhibition of the activity of an ErmC' MTase.

We integrated bioinformatics and experimental methods to identify new chemical compounds able to inhibit the activity of ErmC. Based on the virtual screening of a ZINC database composed of 17 mln lead-like molecules, 29 compounds were chosen for experimental verification. 23 compounds decreased the minimal inhibitory concentration of erythromycin in *Escherichia coli* strain that overexpresses ErmC'. Among them inhibitor TC28 (ZINC code 32747906) with IC₅₀ 100 μM was non-toxic to HEK 293 cells. It served as a template for similarity-based virtual screening, which led to obtaining two derivatives TC3s (ZINC code 62022572) and TC4s (ZINC code 49032257) with IC₅₀ 116 μM and 110 μM, respectively. Analysis of models of TC3s and TC4s docked to the ErmC' structure revealed a competitive mode of inhibition while TC28 a non-competitive. The binding of TC28 is predicted to disrupt the substrate by the ErmC'. Our results provide a basis for the development of inhibitors against the Erm-family enzymes.

P377**Probabilistic genotyping without alignment**Jonas A. Sibbesen¹, Lasse Maretty¹, Anders Krogh¹¹University of Copenhagen, Denmark

A. Krogh <krogh@binf.ku.dk>

We propose a novel probabilistic approach to population genotyping from known variants that does not rely on alignment.

The algorithm first constructs a table containing the counts of all k-mers found in the sequencing reads for each individual. Next, variant loci less than k apart are combined into clusters and all possible haplotypes enumerated except for larger clusters, where a heuristic based on sample k-mer occurrences is used to limit the number of haplotypes. Finally, the multiset containing all k-mers found in haplotypes are enumerated and combined with the corresponding sample table to provide a vector containing the occurrences of haplotype k-mers for each sample.

These count vectors are then used as a basis for probabilistic inference. More specifically, we model the observed counts as generated by combining counts obtained from an individual's diplotype with counts originating from noise. An individual's diplotype is in turn modelled as drawn from a shared population of haplotypes whose frequencies are modelled using a novel sparse prior. The posterior distribution over genotypes is inferred using collapsed Gibbs sampling of diplotypes and haplotype frequencies.

We demonstrate that our method can accurately and rapidly estimate genotypes composed of arbitrarily complex alleles (i.e. single nucleotide, short indel or larger structural variants including nested arrangements) on both simulated and real data.

We suggest that our method may be used when the results from multiple variant discovery tools as well as previously annotated variants need to be integrated to provide final genotype estimates.

P378

Haplotype reconstruction for polyploid organisms

Mohammadhossein Moeinzadeh¹, Jun Yang¹, Martin Vingron¹

¹Max-Planck-Institute for Molecular Genetic, Germany
M. Moeinzadeh <moeinzad@molgen.mpg.de>

Reconstructing haplotypes from next generation sequencing data is one of the major challenges in molecular biology. Although much effort has been done on haplotype assembly of diploid organisms, specifically human, haplotyping for polyploid organisms still remains in its early stages due to the following reasons: 1) lack of a well-established reference genome, 2) higher computational complexity having polyploidy, 3) different rates of evolutionary events such as higher rate of duplication in most polyploid organisms, which not only make the haplotype assembly but also the genome assembly more complicated than the diploid genome. On the other hand, some properties of the polyploid genome, like the higher heterozygosity, could lead polyploid haplotype assembly forward. In this work, we design a pipeline for haplotype assembly during genome assembly and apply the methods on the sweet potato genome data, which is a hexaploid organism. Firstly, in silico haplotype reconstruc-

tion from Illumina reads is done. Then, in order to evaluate, the results are compared with longer 454 reads. Generally, we show that haplotype reconstruction for an acceptable part of polyploid genomes is possible to be done with current high throughput sequencing technologies.

P379

Targeted enrichment sequencing of the large genomic region in glioblastoma

Bartosz Wojtas¹, Bartłomiej Gielniewski¹, Marta Maleszewska¹, Mateusz Bujko², Janusz Siedlecki², Katarzyna Kotulska³, Wiesława Grajkowska³, Bożena Kamińska¹

¹Nencki Institute of Experimental Biology, Poland,

²Maria Skłodowska-Curie Oncology Center, Poland,

³The Children's Memorial Health Institute, Poland

B. Wojtas <b.wojtas@nencki.gov.pl>

Dissecting and identifying mutations and genomic structural variants that accompany tumorigenesis by next generation sequencing holds promise to reveal the genetic landscape of glioblastoma (GBM). GBM is the most deadly brain tumor with 14 months mean survival and high resistance to conventional therapies. Fifteen freshly frozen GBM samples were processed with NimbleGen custom designed enrichment kit, an extension of SeqCap EZ Design – Comprehensive Cancer Design kit with genes and flanking regions for encoding epigenetic enzymes. This custom designed kit spans a human genomic region of around 7 Mb (7 000 000 base pairs). Two independent hybridizations were performed (7 samples multiplexed versus 8 samples multiplexed) and subsequently sequenced with HiSeq 1500 platform technology. GBM DNAs used for a library preparation were precisely evaluated quantitatively and qualitatively during all processing. Sample were measured by Thermo Scientific Nanodrop to control yield and alcohol contamination, by Promega Quantus to measure long intact double DNA fragments concentration, by Agilent Bioanalyzer to evaluate DNA fragmentation and quality and by quantitative qPCR to evaluate library molarity and fold enrichment (estimation of enrichment process efficacy) of prepared libraries. Sequencing run was performed with the rapid mode of HiSeq 1500 sequencing of two independent hybridizations. Sequencing run was done with indexed paired end chemistry of 76 base pairs sequenced

from both sides. It resulted in >165M reads (PF-past filter) from each lane, 95% of base pairs had quality above Q30 (0.001% error probability), with total yield of 52.1 G. To analyze sequencing outcome, GATK Framework Depth Of Coverage and Picard Calculate HsMetrics tools were used to evaluate target enrichment. Surprisingly, despite optimal outcome of sequencing run, hybridization of two different batches did differ in respect of target enrichment in terms of mean target coverage, mean on-target rate and coverage uniformity. Detailed troubleshooting and correlation of target enrichment performance with qualitative and quantitative measurements performed in a process of library preparation allowed to determine causes for differences in target enrichment performance which could be useful in further attempts to perform successful library enrichment of clinical samples.

Work is supported by 2013/09/B/NZ3/01402 grant from National Science Centre.

P380

A method for combining multiple genomic and clinical datatypes to predict recurrence grade in gliomas

Isaac Joseph¹, Shannon McCurdy¹, Joseph F. Costello², Lior Pachter¹

¹University of California, Berkeley, United States,

²University of California, San Francisco, United States

I. Joseph <ijoseph@berkeley.edu>

A central goal of oncology is making treatment decisions that result in the best patient outcome, which must be consequently predicted. In gliomas, a common type of brain tumor with a nearly 100% recurrence rate following initial surgery, recurrence grade is an impactful, yet difficult to predict outcome. Molecular evidence suggests that recurrence grade may be increased by application of chemotherapy agent Temozolomide, modulated by various known molecular mutation-related and adduct-removal pathways. Consequently, the usage of high-throughput sequencing (HTS) genomic data of multiple types may improve our ability to predict recurrence grade by measuring related molecular mutations and epigenetic alterations. Framed as a machine-learning problem, the high-dimensionality of genomic data poses a prediction accuracy challenge due to overfitting, and consequently limits transferability of findings to further patients. To

reduce overfitting, we developed and implemented a novel statistical dimensionality reduction method that relies on the following assumptions: (1) correlation of genomic modalities being informative for outcome, (2) linear relationship of predictors to response, and (3) collinearity of genomic/ clinical predictors. Initial tests of an approximation of the method shows an improved generalizability and interpretability of the relationship between high-dimensional genomics data and clinical outcomes within data from The Cancer Genome Atlas and UCSF Department of Neurooncology. Successful application of the method results in identification of multi-HTS-assay signatures of recurrence grade, useful for understanding mechanisms and treatment decisions in gliomas.

P382

The DNA sequence features of non-methylated islands across six vertebrates

Matt Huska¹, Martin Vingron¹

¹MPI for Molecular Genetics, Germany

M. Huska <huska@molgen.mpg.de>

DNA methylation has long been suspected to play a role in vertebrate gene regulation. While most of the genome tends to be methylated, a series of experiments using methylation-sensitive restriction enzymes that were carried out over 30 years ago identified regions that make up a small percentage of the genome and that were unexpectedly free of methylation. As DNA sequencing technology improved these non-methylated regions were found to be rich in CpG dinucleotides, and this sequence-based definition of non-methylated regions, or CpG islands, has become commonplace. Recently, experimental methods have been developed that can identify non-methylated regions genome-wide, and surprisingly these regions have little overlap with CpG islands, leading some to believe that accurate prediction of non-methylated regions from sequence only is not possible. In this work we explore the sequence of non-methylated regions. First, we show that we can greatly improve the prediction of non-methylated regions by changing the way they are predicted in two ways: learning from example regions rather than predicting the regions from scratch (supervised learning rather than unsupervised learning), and considering longer regions of nucleotides rather than simple dinucleotide counts.

Next, we look at the sequence features that differentiate between non-methylated regions in different tissues as well as across six different vertebrate species.

P383 **Three dimensional threading for protein structure modelling**

Dominik Gront¹

¹Warsaw University, Faculty of Chemistry, Poland

D. Gront <dgront@chem.uw.edu.pl>

Template-based modelling still remains the fastest, most accurate and most reliable method to predict the three dimensional structure of a protein from its sequence. Outcomes of the procedure greatly depend on a selected template structure as well as on the alignment between the template and the query sequence.

In this contribution we present a new 3D threading algorithm that may be used to refine sequence-to-structure alignments. The program employs Monte Carlo sampling schemes such as Parallel Tempering to explore the conformational space of all possible alignments. The energy function combines terms commonly used for alignment calculations such as sequence gap penalty or profile-profile match score with scores based on the template structure. The latter are more important than the former ones and greatly increase the accuracy of the method. Structure-based scores however violates assumptions of dynamic programming approach, so the 3D threading approach is the only way to exploit this advantage.

P385 **A more powerful test for identification of differentially methylated regions**

Przemek Biecek¹

¹Interdisciplinary Centre for Mathematical and Computational Modelling University of Warsaw, Poland

P. Biecek <przemyslaw.biecek@gmail.com>

High-throughput bisulfite sequencing is a useful tool for genome-scale DNA methylation profiling. Data from RRBS (reduced representation bisulfite sequencing) [1] are often used to identify DMRs (differentially methylated regions), i.e. regions with different methylation between two or more condi-

tions.

Currently for such analyses it is common to use Fisher test or logistic regression test. The tested hypothesis is that proportion of methylated sites in the first condition is different than proportion of methylated sites in the second condition [2]. In classical approach methylations for all CpG sites in a considered region are pulled together. Yet, in some conditions it leads to high number of false positives (see Fig 1).

We propose an alternative test that take into account the lack of balance and differences in coverage of different sites in different conditions and find regions with consistent differences in methylation profiles. We also show that this approach has much lower false positive rate than the classical one.

Fig 1.:

<http://bit.ly/1M59a3u>

An example in which classical tests fail. X axis stands for the genome position and the presented region is 2000bp long. Y axis shows the fraction of methylated sites in a given position in a given experimental condition. Color stands for experimental condition. Horizontal bars show the average methylation within condition. The presented region will be considered as significantly changed due to differences between averages. However the true reason for this difference is that readings for these two conditions comes from different sites. So for most methods it is a false positive but our method will correctly classify it.

[1] Hongcang Gu, Zachary Smith, Christoph Bock, Patrick Boyle, Andreas Gnirk, Alexander Meissner (2011)

Preparation of reduced representation bisulfite sequencing libraries for genome-scale DNA methylation profiling

Nature Protocols 6, 468–481

[2] Altuna Akalin, Matthias Kormaksson, Sheng Li, Francine E. Garrett-Bakelman, Maria E. Figueroa, Ari Melnick, Christopher E. Mason. (2012).

methylKit: A comprehensive R package for the analysis of genome-wide DNA methylation profiles. Genome Biology, 13:R87.

P386 **Time-varying Gene Interaction Network Modeling by Sequential Monte Carlo**

Sergiy Ancherbak¹, Ercan Kuruoglu^{1,2}, Martin Vingron²

¹ISTI-CNR, Italy, ²Max Planck Institute for Molecular Genetics, Germany

E. Kuruoglu <ercan.kuruoglu@isti.cnr.it>

Most existing methods used for gene regulatory network modeling are dedicated to inference of steady state networks, which are prevalent over all time instants. However, gene interactions evolve over time. Information about the gene interactions in different stages of the life cycle of a cell or an organism is of high importance for biology. In the statistical graphical models literature one can find a number of methods for network modelling while the study of time varying networks is rather recent. Using synthetic time series dataset for a gene network, we show that a sequential Monte Carlo method, namely Particle Filtering method is capable of tracking gene expression data and infer time-varying networks online.

P388 **Comprehensive selection and analysis of C2H2-ZF DNA-binding domains**

Mona Singh¹

¹Princeton University, United States

M. Singh <mona@cs.princeton.edu>

Cys2His2 zinc fingers (C2H2-ZFs) comprise the largest class of metazoan DNA-binding domains. Despite a well-defined interaction interface, much remains unknown about the DNA-binding landscape of this domain. We have screened large synthetic libraries of C2H2-ZFs for members able to bind each possible three base pair target, thereby providing one of the most comprehensive investigations of C2H2-ZF DNA-binding interactions to date. The resulting data consist of >160,000 unique domain-DNA interactions. An integrated analysis of these independent screens yielded DNA-binding profiles for tens of thousands of domains, and led to the successful design and prediction of C2H2-ZF DNA-binding specificities. Computational analyses uncovered fundamental aspects of C2H2-ZF domain-DNA interaction, including the important roles of within-finger context and domain position

influences on base recognition. We observed the existence of various distinct binding strategies for each possible DNA target. This comprehensive dataset allows a better understanding of the complex binding landscape of C2H2-ZF domains and provides a foundation for efforts in predicting and engineering their DNA-binding specificities.

(This is joint work with Anton V. Persikov, Joshua Wetzel, Elizabeth F. Rowland, Benjamin L. Oakes, Denise J. Xu, and Marcus B. Noyes.)

P389 **Uncertain Biological Networks: A Dynamical Modeling Methodology** **Daniel N. Mohsenizadeh¹, Jianping Hua¹, Michael Bittner¹, Edward R. Dougherty¹**

¹Texas A&M University, United States

D.N. Mohsenizadeh <danielmz@tamu.edu>

This paper presents a new methodology to develop dynamical models for uncertain biological networks. Most current models are built upon one of two methodologies, one process-based and the other based on logical (Boolean) networks. These cannot be effectively used for experimental design purposes in the laboratory. The first requires comprehensive a priori knowledge of the parameters involved in all biological processes, whereas results from the second method may not possess a clear biological correspondence and thus cannot be laboratory tested. Our new methodology does not require full knowledge of the network and can generate a dynamical model based on an available small data set and the results from experimental design can be examined in the laboratory. The method translates a biological database into an interaction-based network containing a set of nodes (biological entities) N , and a set of edges (biological processes) E . Each node has a value, where the abstract term value is interpreted based on the characteristics of that biological entity; for instance, it may represent the concentration level of a protein or the expression level of a gene. Available a priori knowledge can be incorporated into the model by assigning appropriate edge labels that reduce model uncertainty. Our novel algorithm is capable of updating state variables using synchronous and asynchronous approaches, thus leading to a dynamical model that can be used for further analyses and design purposes, such as forward prediction, backward inference, and experimental design.

Author index

A

- Abbasi L. 42 (P238)
- Acheson-Field H. 94 (P346)
- Adamiak R. 50 (P254)
- Aerts S. 42 (P238)
- Akbari A. 31 (T138)
- Ala-Korpela M. 78 (P311)
- Alipanahi B. 24 (H180)
- Alkhateeb A. 67 (P292)
- Alsbeih G. 84 (P325)
- Aluru S. 31 (T129)
- Amir N. 29 (T76)
- Ancherbak S. 112 (P386)
- Andreotti S. 31 (T146)
- Ansari S. 61 (P280)
- Antczak M. 38 (P229), 50 (P254)
- Apostolico A. 31 (T129)
- Ara I. 86 (P328)
- Arnosti D. 63 (P285)
- Arodz T. 65 (P287)
- Aryee M. 23 (H177)

B

- Babu M.M. 17 (K3)
- Badie C. 84 (P325), 85 (P326), 90 (P336)
- Bafna V. 31 (T138)
- Bai Y. 69 (P297)
- Baker E. 102 (P363)
- Baranowski M. 46 (P246)
- Barash Y. 24 (H180)
- Barillot E. 106 (P371)
- Barjaktarovic Z. 90 (P336)
- Bartosik D. 79 (P313)
- Basu A.K. 76 (P308)
- Basu S. 27 (T36)
- Basu S. 45 (P244)
- Batzoglou S. 28 (T46)
- Bayzid M.S. 23 (H179)
- Beaulieu P. 49 (P252)
- Bednarz P. 40 (P234)
- Beerenwinkel N. 34 (P217), 75 (P306)
- Ben-Bassat I. 29 (T93)
- Benham C.J. 86 (P329)
- Benotmane R. 90 (P336)
- Berger B. 16 (K2), 29 (T95), 32 (T160)
- Berger E. 29 (T95)
- Bernaer J. 29 (T89)
- Biecek P. 111 (P385)
- Binczyk F. 91 (P338)
- Birol I. 72 (P302)
- Bishara A. 28 (T46)
- Biswas S. 32 (T167)
- Bittner M. 112 (P389)
- Blachowicz A. 90 (P336)
- Blanton S. 104 (P367)
- Błaszczuk M. 83 (P322)
- Blencowe B. 24 (H180)
- Blumcke I. 80 (P316)
- Bobek-Billewicz B. 91 (P338)
- Bodył A. 44 (P242)
- Boeva V. 44 (P243), 106 (P371)
- Bolotin E. 38 (P230)
- Boniecki A. 45 (P244)
- Boniecki M. 61 (P279), 100 (P359)
- Bonissone S. 30 (T126)
- Borowski M. 37 (P225)
- Bot A. 80 (P316)
- Botzman M. 53 (P262)
- Bouchard-Cote A. 104 (P367)
- Boudelloua I. 92 (P341)
- Bouffler S. 90 (P336)
- Boussau B. 23 (H179)
- Boža V. 57 (P271)
- Brandes N. 41 (P235)
- Brejová B. 43 (P240), 57 (P271)
- Bretschneider H. 24 (H180)
- Brodz A. 53 (P262)
- Brudno M. 26 (H215)
- Bujko M. 109 (P379)
- Bujnicki J.M. 61 (P279), 89 (P334), 99 (P356), 100 (P359), 108 (P375)
- Burdukiewicz M. 45 (P245), 46 (P247)
- Butte A.J. 76 (P308)
- Böcker S. 29 (T74)
- Błażej P. 45 (P245), 48 (P250), 50 (P256)
- Błażewicz J. 36 (P224), 37 (P226), 38 (P229), 50 (P254)
- Břinda K. 44 (P243), 75 (P307)

C

- | | | | |
|-------------------|---------------------------------|--------------------|---------------------------------|
| Cai Y. | 42 (P238) | Devkota P. | 104 (P367) |
| Campbell C. | 59 (P274) | Dewey C. | 69 (P297) |
| Camphausen K. | 53 (P263) | Dhabalia A. | 38 (P230) |
| Candéias S. | 85 (P326) | Dharanipragada P. | 85 (P327) |
| Canzar S. | 31 (T146) | Di Giallonardo F. | 34 (P217) |
| Carmi S. | 30 (T120) | Diamanti K. | 52 (P261), 62 (P281) |
| Carrieri A.P. | 27 (T36) | Diament A. | 24 (H191), 82 (P321) |
| Casellas R.C. | 86 (P329) | Ding L. | 25 (H212) |
| Cathy L.G. | 107 (P373) | Dobson J. | 25 (H212) |
| Cavalli M. | 52 (P261) | Doig A.J. | 58 (P272) |
| Cavallo-Medved D. | 67 (P292) | Dojer N. | 104 (P366), 105 (P369) |
| Cebrat S. | 65 (P289) | Dolbniak M. | 37 (P227), 84 (P325) |
| Celiku O. | 53 (P263) | Donald B.R. | 27 (T41), 30 (T114) |
| Charzewski Ł. | 77 (P310) | Donato M. | 61 (P280) |
| Charzyńska A. | 68 (P294) | Dougherty E.R. | 112 (P389) |
| Chauve C. | 32 (T170) | Draghici S. | 61 (P280) |
| Chen C. | 38 (P230) | Draminski M. | 62 (P281) |
| Chen C. | 27 (T19) | Drechsel O. | 63 (P284) |
| Chen Y. | 33 (P190) | Drouin S. | 49 (P252) |
| Cheng Y. | 25 (H212) | Duggal G. | 31 (T150) |
| Cho H. | 32 (T160) | Dunin-Horkawicz S. | 89 (P334) |
| Cho S.Y. | 33 (P178) | Durbin R. | 69 (P296) |
| Cho Y. | 103 (P364) | Dziedzic A. | 80 (P315) |
| Chor B. | 29 (T93) | Dziubiński M. | 74 (P305), 95 (P347), 96 (P350) |
| Chu J. | 72 (P302) | Dührkop K. | 29 (T74) |
| Chuang K. | 33 (P190) | Dębski K. | 80 (P316) |
| Chung H. | 98 (P353) | | |
| Ciach M. | 78 (P312) | E | |
| Cichonska A. | 78 (P311) | Ehrenberg M. | 93 (P344) |
| Cohen D. | 29 (T76) | El-Kebir M. | 94 (P346) |
| Collins M. | 69 (P297) | El-Osta A. | 80 (P316) |
| Cooper D. | 59 (P274) | Eldridge J. | 25 (H212) |
| Costello J.F. | 110 (P380) | Eskin E. | 27 (T30), 107 (P372) |
| Cramer P. | 40 (P233) | | |
| Crowther J. | 42 (P238) | F | |
| Cserző M. | 59 (P275) | Feder M. | 108 (P375) |
| Cudek P. | 81 (P318) | Fedichev P. | 49 (P253) |
| Czaplewski C. | 54 (P265), 67 (P291) | Fillmore N. | 69 (P297) |
| | | Finnon R. | 90 (P336) |
| D | | Firlik G. | 95 (P347) |
| Dabrowski M.J. | 62 (P281), 97 (P352) | Fiume M. | 26 (H215) |
| Dangl J. | 32 (T167) | Flowers M. | 65 (P288) |
| Daniluk P. | 74 (P305), 95 (P347), 96 (P350) | Fogtman A. | 98 (P354) |
| David E. | 53 (P262) | Foik I. | 108 (P375) |
| Dawid A. | 101 (P360) | Fonseca R. | 29 (T89) |
| Dawson W. | 61 (P279), 100 (P359) | Formanowicz D. | 56 (P269), 56 (P270) |
| Day I. | 59 (P274) | Formanowicz P. | 37 (P225), 56 (P269), 56 (P270) |
| Demir F. | 26 (H215) | Franco Pérez E. | 33 (P184) |
| Dempster S. | 63 (P283) | Frey B. | 24 (H180) |
| Deng X. | 53 (P263) | Frishberg A. | 52 (P259) |
| Deshwar A.G. | 103 (P365) | Fujarewicz K. | 37 (P227) |

Furlotte N.	27 (T30)	Hallen M.	27 (T41)
G		Handelman S.	89 (P335)
Gabaldón T.	35 (P221)	Hansen P.	64 (P286)
Gagat P.	44 (P242)	Harel T.	54 (P264)
Gagneur J.	40 (P233)	Hartmann K.	90 (P337)
Gambette P.	32 (T173)	Hausmanowa-Petrusewicz I.	77 (P310)
Gambin A.	60 (P278), 68 (P294), 73 (P303), 83 (P323)	Hałabis A.	67 (P291)
Gao X.	65 (P287)	Healy J.	49 (P252)
Garbulowski M.	92 (P342)	Hecht J.	64 (P286)
Gat-Viks I.	28 (T55), 52 (P259), 53 (P262), 54 (P264)	Heckerman D.	23 (H177)
Gaunt T.	59 (P274)	Helling K.	65 (P288)
Georgiev I.	30 (T114)	Hentges K.E.	58 (P272)
Getz G.	25 (H212)	Herencsár A.	43 (P240)
Gielniewski B.	109 (P379)	Herman-Izycka J.	94 (P345)
Giełdoń A.	56 (P268)	Hong C.P.	79 (P314)
Gilad Y.	88 (P333)	Hormozdiari F.	107 (P372)
Ginalski K.	104 (P366), 105 (P369)	Hua J.	112 (P389)
Gogolewski K.	60 (P278)	Hua Y.	24 (H180)
Gogolinska A.	55 (P266)	Hughes T.	24 (H180)
Golas E.	54 (P265)	Hunyady L.	59 (P275)
Goldenberg A.	26 (H215), 106 (P370)	Huska M.	110 (P382)
Gong H.	30 (T98)	Huusko N.	80 (P316)
Gong H.	30 (T98)	Huynh-Thu V.A.	42 (P239)
Gonzales S.	102 (P363)	Hvidsten T.	39 (P231)
Gonzalez-Perez A.	25 (H212)	I	
Gough J.	59 (P274)	Iakovishina D.	106 (P371)
Grabherr M.	97 (P352)	Ibrahim D.	64 (P286)
Grabińska M.	48 (P250)	Iwanicka-Nowicka R.	98 (P354)
Grabowicz I.	51 (P258)	J	
Grad Ł.	73 (P304)	Jackman S.D.	72 (P302)
Grajkowska W.	109 (P379)	Jagiela D.	67 (P291)
Grant K.	102 (P363)	Jain S.	30 (T114)
Gront D.	101 (P360), 111 (P383)	Jaksik R.	37 (P227), 97 (P351)
Grover M.P.	34 (P219)	Jakubowski R.	88 (P332)
Gruca A.	93 (P343)	Jamroz M.	83 (P322), 96 (P349)
Grzebelus D.	60 (P278)	Jang I.	50 (P255)
Guerousov S.	24 (H180)	Jankowski A.	70 (P298)
Gunawan A.	32 (T173)	Jarmolińska A.	99 (P355)
Górecki P.	87 (P330), 87 (P331)	Jasiński M.	47 (P249)
Górska A.	47 (P249)	Jernigan R.	27 (T31)
Götz M.	91 (P338)	Jerzmanowski A.	98 (P354)
Günthard H.F.	34 (P217)	Jetka T.	102 (P362)
Głowacki T.	37 (P225)	Ji H.	82 (P319)
H		Jojic N.	24 (H180)
Haibe-Kains B.	26 (H215), 106 (P370)	Jojic V.	32 (T167)
Haj Dezfulian M.	66 (P290)	Joo J.W.	27 (T30)
Hallay-Suszek M.	77 (P310)	Joseph I.	110 (P380)
		Jou J.D.	30 (T114)
		Jursa J.	57 (P271)

Järvelin M.	78 (P311)	Krogh A.	108 (P377)
K		Kruczyk M.	52 (P261)
Kabir M.	58 (P272)	Krupa P.	58 (P273)
Kalinowski M.	71 (P300)	Krzyśko K.	77 (P310)
Kambadur P.	28 (T59)	Krüger T.	75 (P306)
Kaminska B.	97 (P352), 109 (P379)	Kubicka A.	104 (P366), 105 (P369)
Kaminska K.H.	89 (P334)	Kucherov G.	44 (P243), 75 (P307)
Kang E.Y.	27 (T30)	Kuhn D.	27 (T36)
Kangas A.J.	78 (P311)	Kulik M.	43 (P241)
Karczyńska A.	58 (P273)	Kumar K.	30 (T119)
Karro J.	65 (P288)	Kurcinski M.	83 (P322)
Kashef-Haghighi D.	28 (T46)	Kuruoglu E.	112 (P386)
Kaspi A.	80 (P316)	Kwiatkowska A.	98 (P354)
Katiyar U.	67 (P292)	L	
Keich U.	32 (T166)	Labaj W.	82 (P320)
Kelly J.	67 (P292)	Labarre A.	32 (T173)
Keshava N.	63 (P283)	Lajoie M.	49 (P252)
Kierczak M.	62 (P282)	Lauber C.	46 (P247)
Kim J.	105 (P368)	Lawrence M.	25 (H212)
Kim J.	105 (P368)	Le Rouzic A.	60 (P278), 83 (P323)
Kim S.	105 (P368)	Lee B.	42 (P237), 50 (P255)
Kim Y.	25 (H212)	Lee L.	24 (H180)
Kimmel M.	69 (P295)	Lee S.	28 (T59)
Kingsford C.	25 (H214), 31 (T150)	Lee Y.S.	33 (P178)
Kirkpatrick B.	104 (P367)	Leelananda S.	27 (T31)
Klau G.W.	31 (T146)	Lehtimäki T.	78 (P311)
Kloczkowski A.	27 (T31), 89 (P335)	Leiserson M.	25 (H212), 31 (T156)
Kmieciak S.	83 (P322), 96 (P349)	Lesyng B.	71 (P300), 74 (P305), 77 (P310)
Ko J.	79 (P314)	Levens D.	86 (P329)
Koblowska M.	98 (P354)	Levitt M.	15 (K1)
Kobow K.	80 (P316)	Lewandowska A.	67 (P291)
Kogan V.	49 (P253)	Li B.	69 (P297)
Koh I.	103 (P364)	Li J.	105 (P369)
Kolinski A.	83 (P322)	Li Y.	88 (P333)
Komorowski J.	52 (P261), 62 (P281), 97 (P352)	Li Y.	27 (T19)
Komorowski M.	60 (P277), 102 (P362)	Li Y.	106 (P370)
Konarska M.M.	20 (K6)	Lidschreiber M.	40 (P233)
Konopka B.	70 (P299)	Lin C.	33 (P190)
Kopp W.	108 (P374)	Lin Y.	104 (P366), 105 (P369)
Koronacki J.	62 (P281)	Linial M.	41 (P235), 76 (P309)
Kotas J.	85 (P326)	Lippert C.	23 (H177)
Kotliński M.	98 (P354)	Lipska A.G.	56 (P268)
Kotulska K.	109 (P379)	Listgarten J.	23 (H177)
Kotulska M.	48 (P251), 52 (P260), 70 (P299)	Litova N.	68 (P293)
Kouzine F.	86 (P329)	Liu Y.	28 (T46)
Kozłowski L.	99 (P356)	Liwo A.	39 (P232), 54 (P265), 56 (P268), 58 (P273), 67 (P291), 68 (P293)
Krainer A.	24 (H180)	Lopez-Bigas N.	25 (H212)
Krannich A.	64 (P286)	Lozano A.	28 (T59)
Kreil D.P.	95 (P348)	Lukasiak P.	38 (P229), 50 (P254)
Kroczak A.	51 (P257)		

- Lundberg D. 32 (T167) Moeinzadeh M. 109 (P378)
Lusis A. 27 (T30) Mohamadi H. 72 (P302)
L Mohsenizadeh D.N. 112 (P389)
Łabaj P.P. 95 (P348) Molodtsov I. 49 (P253)
Łach G. 100 (P359) Mondal S. 63 (P285)
Łukasiuk K. 80 (P316) Montesinos López O.A. 33 (P184)
Łukasz P. 100 (P359) Moret B. 29 (T66)
Łącki M.K. 73 (P303) Mori T. 43 (P241)
M Morris Q. 24 (H180), 103 (P365)
Ma J. 31 (T137) Mort M. 59 (P274)
Machnicka M.A. 89 (P334) Mount S. 25 (H214)
Macioszek A. 98 (P354) Mozolewska M. 58 (P273)
Mackiewicz D. 65 (P289) Mroczek T. 81 (P318)
Mackiewicz P. 44 (P242), 45 (P245), 46 (P247), 48 (P250), 50 (P256), 51 (P257) Murphy R.F. 28 (T57)
Madej-Pilarczyk A. 77 (P310) Muszewska A. 78 (P312), 80 (P315)
Mahmood U. 84 (P324) Mykowiecka A. 87 (P331), 95 (P347), 96 (P350)
Maier-Hein K. 91 (P338) Mäkinen V. 28 (T63)
Makowski M. 39 (P232) **N**
Maleki M. 66 (P290) Naik A.W. 28 (T57)
Maleszewska M. 109 (P379) Najafabadi H. 24 (H180)
Mammana A. 98 (P353) Nebel D. 38 (P229)
Mandric I. 32 (T161) Newburger D. 28 (T46)
Manescu D. 32 (T166) Nguyen N. 30 (T119)
Mangul S. 107 (P372) Nicosia G. 41 (P236)
Mar J.C. 72 (P301) Niebroj-Dobosz I. 77 (P310)
Marchel M. 77 (P310) Niemi A. 68 (P293)
Marczyk M. 81 (P317) Nienałtowski K. 60 (P277)
Maretty L. 108 (P377) Niu B. 25 (H212)
Marszalek J. 54 (P265) Nowak W. 41 (P236), 55 (P266), 88 (P332)
Martinez-Canales M. 76 (P308) **O**
Marttinen P. 78 (P311) Ochab M. 55 (P267)
Maskoni D. 67 (P292) Oesper L. 94 (P346)
Mazza A. 30 (T113) Oleniecki T. 95 (P347), 96 (P350)
McCurdy S. 110 (P380) Org E. 27 (T30)
McDonald M. 32 (T167) Ossowski S. 63 (P284)
McLellan M. 25 (H212) Ołdziej S. 46 (P246), 67 (P291)
McPherson A. 32 (T170) **P**
Meinzer H. 91 (P338) Pacholczyk M. 69 (P295)
Mellenius H. 93 (P344) Pachter L. 110 (P380)
Menshikov L. 49 (P253) Papież A. 81 (P317)
Merico D. 24 (H180) Papoutsaki A. 25 (H212)
Metzner K.J. 34 (P217) Parekh N. 85 (P327)
Mezlini A. 26 (H215) Parida L. 27 (T36)
Migacz S. 73 (P304) Park K. 103 (P364)
Mihasan M. 36 (P223) Park S. 33 (P178)
Mirarab S. 23 (H179), 30 (T119) Park S. 103 (P364)
Mitra A. 104 (P366) Park Y. 103 (P364)
Mnich M. 45 (P244) Parks B. 27 (T30)

Pasero P.	104 (P366), 105 (P369)	Rosenberg N.A.	31 (T138)
Paszek J.	87 (P330)	Roszak R.	70 (P299)
Patro R.	25 (H214)	Roth A.	32 (T170)
Paulino D.	72 (P302)	Rother K.	100 (P359)
Pe'er I.	30 (T120)	Rousu J.	78 (P311)
Peng J.	32 (T160)	Rowicka M.	104 (P366), 105 (P369)
Peplowski L.	88 (P332)	Rościszewski A.	73 (P304)
Pevzner P.	30 (T126)	Rudnicki W.	73 (P304), 81 (P318)
Piecuch K.	43 (P240)	Rueda L.	66 (P290), 67 (P292)
Pillai M.R.	35 (P220)	Ruppin E.	30 (T113)
Pinter R.Y.	24 (H191)	Rybarczyk A.	37 (P226), 56 (P269), 56 (P270)
Pirinen M.	78 (P311)	Rydzewski J.	41 (P236)
Pitkänen A.	80 (P316)	Ryslik G.	25 (H212)
Plewczyński D.	45 (P244), 62 (P282)	Ryu D.	79 (P314)
Podolsky D.	49 (P253)	Rzeszowska-Wolny J.	37 (P227)
Polańska J.	81 (P317), 84 (P325), 85 (P326), 90 (P336), 91 (P338)	Rätsch G.	103 (P365)
Polański A.	81 (P317), 82 (P320), 92 (P342)		
Popenda M.	50 (P254)	S	
Porter L.	67 (P292)	Sablina A.	42 (P238)
Posacki P.	47 (P248)	Sadee W.	89 (P335)
Prabhakar S.	70 (P298)	Sahinalp S.C.	32 (T170)
Pritchard J.	88 (P333)	Sahlin K.	28 (T63)
Prokop S.	59 (P275)	Salmela L.	28 (T63)
Pryszcz L.	35 (P221)	Salo A.	102 (P363)
Przychodzen B.	100 (P358)	Salomaa V.	78 (P311)
Przytycka T.M.	86 (P329)	Sanguinetti G.	42 (P239)
Pujols J.	96 (P349)	Scheraga H.A.	54 (P265), 56 (P268)
Purta E.	108 (P375)	Scherer S.	24 (H180)
Puszyński K.	55 (P267), 97 (P351)	Schwemmer M.	91 (P340)
		Schöpflin R.	59 (P276)
		Sefer E.	31 (T150)
R		Seidman S.R.	56 (P268)
Radom M.	56 (P269), 56 (P270)	Seifert D.	34 (P217)
Ragan M.A.	72 (P301)	Seong J.K.	33 (P178)
Raghavan K.	72 (P302)	Seweryn M.	89 (P335), 90 (P337), 91 (P340)
Raitakari O.T.	78 (P311)	Shahik S.M.	86 (P328)
Raj A.	88 (P333)	Shankavaram U.	53 (P263)
Raphael B.	29 (T73), 94 (P346)	Shao M.	29 (T66)
Raphael B.	25 (H212), 31 (T156)	Sharan R.	30 (T113)
Rappoport N.	76 (P309)	Shchur V.	69 (P296)
Raymond A.	72 (P302)	Shihab H.	59 (P274)
Reddy S.	67 (P292)	Shim H.	88 (P333)
Regnier M.	106 (P371)	Shmookler Reis R.J.	49 (P253)
Reinert K.	31 (T146)	Sibbesen J.A.	108 (P377)
Reshmi G.	35 (P220)	Sidow A.	28 (T46)
Rezaeian I.	67 (P292)	Siedlecki J.	109 (P379)
Riedel M.	38 (P229)	Sieradzan A.	67 (P291), 68 (P293)
Ripatti S.	78 (P311)	Sikora-Wohlfeld W.	76 (P308)
Robinson P.N.	64 (P286)	Sikora M.	93 (P343)
Rogers M.	59 (P274)	Singh M.	112 (P388)
Ronen R.	31 (T138)	Sinnett D.	49 (P252)

- Skrzypczak M. 104 (P366), 105 (P369)
- Smieszek S. 100 (P358)
- Smolińska K. 69 (P295)
- Sobczyk P. 45 (P245), 46 (P247)
- Sobieraj M. 71 (P300)
- Soininen P. 78 (P311)
- Sokołowska B. 77 (P310)
- Song S. 79 (P314)
- Sołtysiński T. 100 (P359)
- Srihari S. 72 (P301)
- St-Onge P. 49 (P252)
- Startek M. 60 (P278), 83 (P323)
- Stephens M. 88 (P333)
- Stewart R. 69 (P297)
- Stjeltjes B. 91 (P338)
- Stypka Ł. 93 (P343)
- Su J. 38 (P230)
- Sugita Y. 43 (P241)
- Sulkowska J.I. 99 (P355), 99 (P357)
- Sundermann L.K. 103 (P365)
- Susak H. 63 (P284)
- Sulecki A. 73 (P304)
- Sykulski M. 75 (P307)
- Szachniuk M. 37 (P226), 38 (P229), 50 (P254)
- Szalai B. 59 (P275)
- Szczasiuk A. 96 (P349)
- Szczurek E. 75 (P306)
- Szuplewska M. 79 (P313)
- Szybalski W. 18 (K4)
- Szóstak N. 37 (P226)
- Ś**
- Świstak M. 79 (P313)
- T**
- Tabaszewski P. 73 (P304)
- Taherian Fard A. 72 (P301)
- Tamborero D. 25 (H212)
- Tandle A. 53 (P263)
- Tapio S. 90 (P336)
- Tarkhov A. 49 (P253)
- Tarnawski R. 91 (P338)
- Tatjewski M. 45 (P244), 62 (P282)
- Temerinac-Ott M. 28 (T57)
- Tenzen T. 82 (P319)
- Tesler G. 31 (T138)
- Thankachan S.V. 31 (T129)
- Thomas J. 25 (H212)
- Thompson J. 69 (P297)
- Tiuryn J. 70 (P298), 101 (P361)
- Tomala K. 100 (P359)
- Tomescu A. 28 (T63)
- Torabi Moghadam B. 97 (P352)
- Tresch A. 40 (P233)
- Trylska J. 43 (P241), 47 (P249)
- Tseng E. 107 (P372)
- Tu Z. 26 (H215)
- Tuller T. 24 (H191), 82 (P321)
- Tuszynska I. 108 (P375)
- U**
- Ul-Haq Z. 84 (P324)
- Umer H.M. 52 (P261)
- Upfal E. 29 (T73)
- Urantówka A. 51 (P257)
- Utro F. 27 (T36)
- V**
- van den Bedem H. 29 (T89)
- Van Steensel B. 19 (K5)
- Vandervalk B.P. 72 (P302)
- Vandin F. 25 (H212), 29 (T73), 31 (T156)
- Ventura S. 96 (P349)
- Vialette S. 32 (T173)
- Villmann T. 38 (P229)
- Vinar T. 57 (P271)
- Vingron M. 59 (P276), 108 (P374), 109 (P378), 110 (P382), 112 (P386)
- Vivekanand A. 35 (P220)
- Várnai P. 59 (P275)
- W**
- Wadelius C. 52 (P261)
- Wagner A. 30 (T113)
- Wang B. 26 (H215)
- Wang D. 63 (P283)
- Wang J. 30 (T98)
- Wang S. 31 (T137)
- Wang S. 88 (P333)
- Wang Z. 31 (T137)
- Warnow T. 23 (H179), 30 (T119)
- Warren R.L. 72 (P302)
- Wasik S. 36 (P224)
- Wasserman W. 27 (T19)
- Weber C. 91 (P338)
- Weese D. 31 (T146)
- Wei Y. 63 (P285)
- Wei Y. 82 (P319)
- Weng Z. 28 (T46)
- West R. 28 (T46)
- Wilczyński B. 39 (P231), 40 (P234), 51 (P258), 63 (P285), 94 (P345), 98 (P354)

Wilentzik R.	28 (T55)	Żmudzińska W.	67 (P291)
Winarski T.	102 (P362)	Żulpo M.	48 (P251)
Wiśniewska M.	39 (P232)		
Wnętrzak M.	50 (P256)		
Wojtas B.	109 (P379)		
Wolfson H.J.	29 (T76)		
Wong L.	101 (P361)		
Woźniak M.	101 (P361)		
Woźniak P.	52 (P260)		
Wu H.	25 (H212), 31 (T156)		
Wu Y.	28 (T54)		
Wójtowicz D.	86 (P329)		

X

Xing E.	28 (T59)
Xiong H.Y.	24 (H180)
Xu J.	31 (T137)

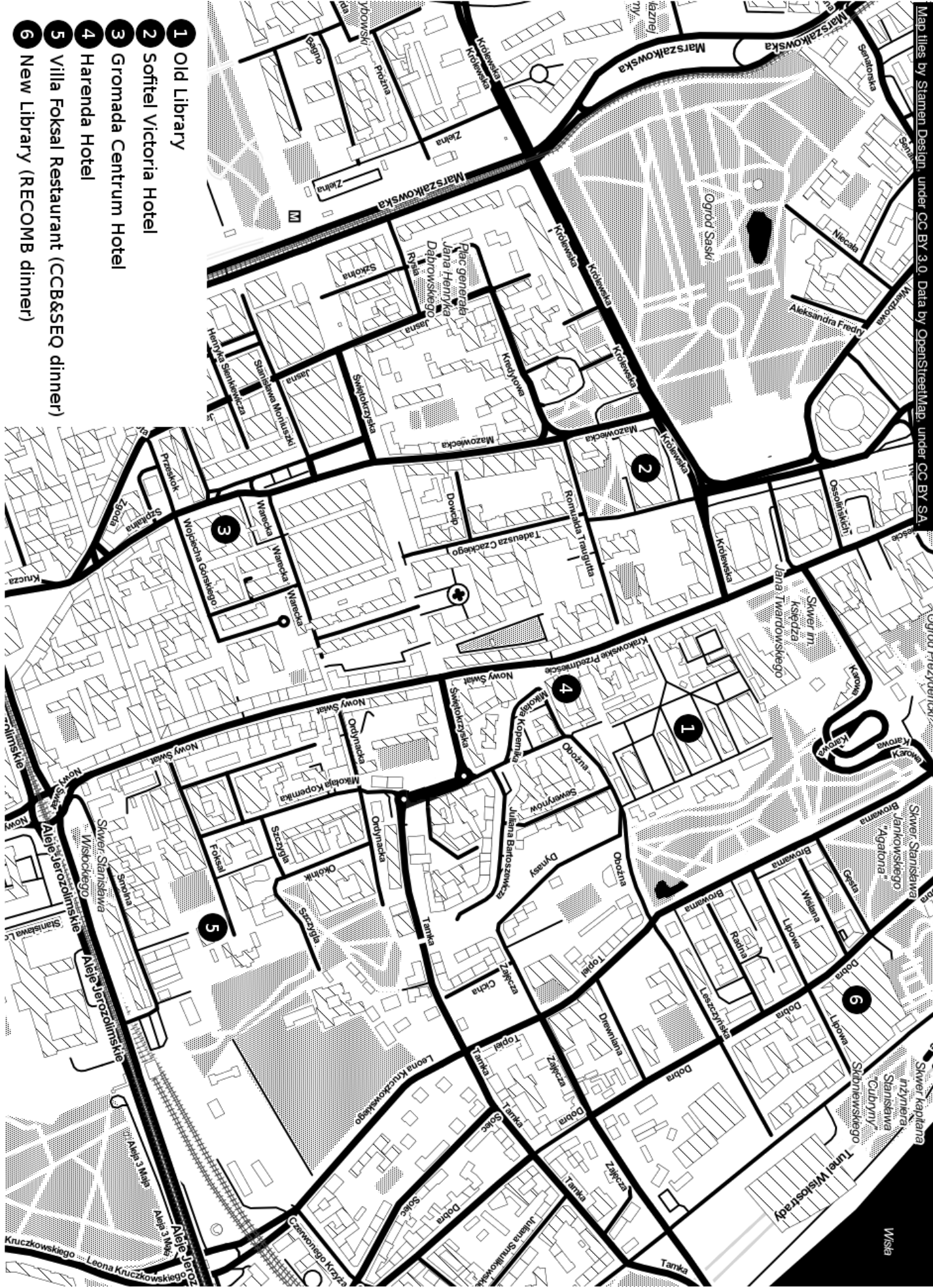
Y

Yamane A.	86 (P329)
Yang J.	109 (P378)
Yeang C.	33 (P190)
Yen S.	33 (P190)
Yorukoglu D.	27 (T36), 29 (T95)
Young S.	30 (T120)
Yuen R.	24 (H180)

Z

Zaborowski B.	67 (P291)
Zaborowski R.	39 (P231)
Zacher B.	40 (P233)
Zakov S.	31 (T138)
Zambrano R.	96 (P349)
Zapata L.	63 (P284)
Zaremba M.	73 (P304)
Zelikovsky A.	32 (T161), 107 (P372)
Zeng J.	28 (T54), 30 (T98)
Zenin A.	49 (P253)
Zhang J.	65 (P288)
Zhang L.	32 (T173)
Zhang S.	30 (T98)
Zhao C.	106 (P370)
Zhao S.	53 (P263)
Zhou Y.	28 (T54)
Ziemann M.	80 (P316)
Zok T.	38 (P229), 50 (P254)
Zou J.	23 (H177)
Zubek J.	45 (P244), 62 (P282)
Zyla J.	84 (P325)

Ż



- 1 Old Library
- 2 Softel Victoria Hotel
- 3 Gromada Centrum Hotel
- 4 Harenda Hotel
- 5 Villa Foksal Restaurant (CC&S&EQ dinner)
- 6 New Library (RECOMB dinner)

Platinum sponsors



Warsaw Center
of Mathematics
and Computer Science



Biogen[™]

Gold sponsors



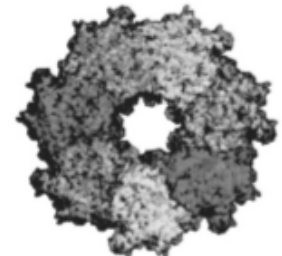
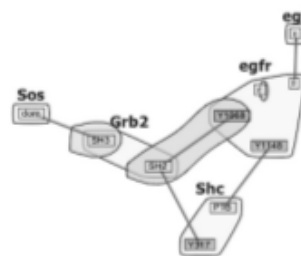
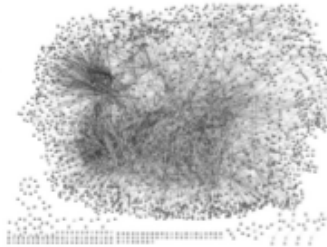
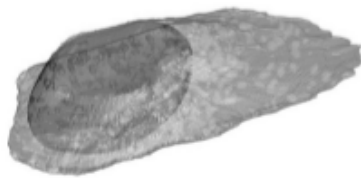
Ministerstwo Nauki
i Szkolnictwa Wyższego





Carnegie Mellon University | University of Pittsburgh

Ph.D Program in Computational Biology



Application Deadline: December 15, 2015
www.compbio.cmu.edu | www.compbio.pitt.edu

PROGRAM OVERVIEW

- Research program combining the strengths of the internationally renowned computational and biomedical research programs of Carnegie Mellon University and the University of Pittsburgh
- Broad, interdisciplinary curriculum and training
- All trainees have access to advisors, classes, and facilities at both institutions
- Competitive stipend and full tuition remission

CURRICULUM

- Core courses offer an overview of the current state-of-the-art in computational biology, and the fundamental concepts and approaches in its component life, physical, and computer sciences.
- A wide selection of advanced elective classes are available in computational biology, computer science, biomedicine, and related areas
- A seminar series, journal club, and research ethics training complement the coursework.

AREAS OF SPECIALIZATION

- Bioimage Informatics
- Cellular and Systems Modeling
- Computational Genomics
- Computational Structural Biology

STUDENTS WILL LEARN TO:

- Analyze gene sequences as part of the future of personalized medicine
- Construct models of cell-signaling networks that can identify novel drug tests
- Run large-scale simulations of molecular machines
- Analyze evolutionary history to predict critical pathways and pathologies
- Develop image-derived computational models of cell organization & cell functions
- Develop and apply algorithms for image analysis of cell, medical, and tissue samples

WHO SHOULD APPLY?

- Highly motivated students with interest in the field of Computational Biology are encouraged to apply.
- Cross-disciplinary training in biological and quantitative sciences is common among many of our applicants.
- We welcome applications from students with degrees in computational biology, bioinformatics, and related areas as well as those with cross-disciplinary coursework or research experience in biology, chemistry, computer science, engineering, mathematics, physics, statistics, and/or other STEM disciplines.

For more information contact: admissions@compbio.cmu.edu or admissions@compbio.pitt.edu



Supported by the
 HHMI-NIBIB
 Interfaces Initiative

