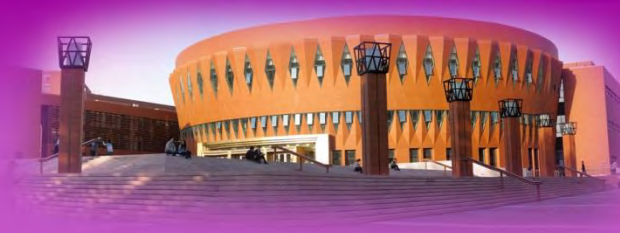
RECOMB 2013

17th Annual International Conference
on Research in Computational Molecular Biology

Beijing | April 7-10, 2013



RECOMB 2013

Book of Abstracts

Conference Chair:

Xuegong Zhang, Tsinghua University

Program Committee Chair:

Fengzhu Sun, University of Southern California

Organizing Committee Chairs:

Xuegong Zhang, Tsinghua University

Minghua Deng, Peking University

Rui Jiang, Tsinghua University

Bioinformatics Division / Center for Synthetic and Systems Biology
TNLIST, Tsinghua University

Program at a Glance

April 7-10, 2013, the New Tsinghua Auditorium, Tsinghua University

April 6 (Saturday)

18:00-20:00	Reception (all registered participants), Wenjin Hotel

April 7 (Sunday)

10:00-10:10	Opening Remarks
10:10-11:10	Keynote Talk by Deborah A. Nickerson
11:10-12:10	Session 1: Population Genomics
13:30-14:30	Poster Session 1
14:30-15:30	Keynote Talk by Chung-I Wu
15:30-16:10	Session 2: Evolutionary Genomics
16:30-17:50	Session 3: Computational Methods for Genomics

April 8 (Monday)

9:00-10:00	Keynote Talk by Takashi Gojobori
10:00-10:40	Session 4: Comparative Genomics
11:00-12:20	Session 5: Cancer Genomics
13:30-14:30	Poster Session 2
14:30-16:10	Session 6: RNA and Chromosome Structure I
16:30-17:10	Session 7: RNA and Chromosome Structure II

April 9 (Tuesday)

9:00-10:00	Keynote Talk by Scott Fraser
10:00-10:40	Session 8: Molecular Networks I
11:00-12:20	Session 9: Molecular Networks II
13:30-17:30	Great Wall Tour (all registered participants, reservation required)
18:30-20:30	Banquet (full registration only)

April 10 (Wednesday)

9:00-10:00	Keynote Talk by Nadia Rosenthal
10:00-10:40	Session 10: Epigenomics
11:00-12:20	Session 11: Protein Structure
13:40-14:40	Keynote Talk by Xiaoliang Sunney Xie
14:40-15:20	Session 12: Image Analysis and Comparative Epigenomics
15:40-17:00	Session 13: Proteomics
17:00-17:10	Closing Ceremony

April 11-12, RECOMB-Seq, Lecture Hall, FIT Building, Tsinghua University

Preface

The RECOMB conference series, with the full name of the Annual International Conference on Research in Computational Molecular Biology, started in 1997 by Sorin Istrail, Pavel Pevzner, and Michael Waterman. The 17th Annual International Conference on Research in Computational Molecular Biology or RECOMB 2013 was held at Tsinghua University, Beijing, China, hosted by the Bioinformatics Division of Tsinghua National Laboratory for Information Science and Technology (TNLIST), Tsinghua University.

This year's RECOMB conference featured 6 invited keynote talks by leading scientists in life sciences in the world. The keynote speakers were Scott Fraser (University of Southern California, USA), Takashi Gojobori (National Institute of Genetics, Japan), Deborah Nickerson (University of Washington, USA), Nadia A. Rosenthal (Monash University in Melbourne, Australia), Chung-I Wu (Beijing Institute of Genomics, Chinese Academy of Sciences, China), and Xiaoliang Sunny Xie (Harvard University, USA).

The conference features three complementary tracks. The proceeding track invites submissions of new research in all areas of computational biology and bioinformatics. A total of 167 extended abstracts were submitted to RECOMB2013 and 32 of them were selected for oral presentation at the conference after extensive reviews. Twenty three accepted extended abstracts and 9 shortened 2-page abstracts were published on *Lecture Notes in Bioinformatics* (Volume 7821). All the authors of the accepted extended abstracts were invited to submit their full papers to the Journal of Computational Biology and the authors of 2-page abstracts decided to publish their full papers on other journals.

The highlight track invites submissions of abstracts of papers published between October 2011 and February 2013. A total of 48 submissions were received and 9 of them were selected for oral presentation chosen by the program chair and the RECOMB steering committee. Papers were selected based on the topics that are complementary to the accepted extended abstracts, impact of the work on the field, interests to a broad audience, and the likelihood that the work will make a good presentation.

The poster track invites abstracts of ongoing and published researches that use mathematical, statistical and computational approaches to solve biological and biomedical problems. A total of 200 submissions were received. A group of poster committee members read and evaluated the submissions and gave suggestions to some of the posters. All the submissions contribute to our understanding of biological and/or biomedical researches using computational approaches.

This book contains the abstracts of 6 keynote speakers, 9 highlight track talks and 200 poster abstracts. The success of RECOMB depends on the efforts, dedication and devotion of many colleagues who spent countless of hours on the organization of the conference. The steering committee consisting of Vineet Bafna, Serafim Batzoglou, Bonnie Berger, Sorin Istrail, Michal Linial, and Martin Vingron (Chair) gave many excellent suggestions on the organization of the conference. We thank the PC members and the external reviewers for the timely review of the assigned papers despite their busy schedules. We also thank all the authors for submitting their excellent work to RECOMB. We would like to personally thank the local organizing committee members especially the co-chairs Xuegong Zhang, Minghua Deng and Rui Jiang, and the local secretary Zhuwei Joan Zhang for their efforts that insured smooth cooperation on the administrative and logistic details. Various organizations including Tsinghua University, TNLIST, the National Science Foundation of China (NSFC), the US National Science Foundation (NSF), the International Society of Computational Biology (ISCB) and all the industry sponsors for their financial support. Dr. Mona Singh (Princeton University) helped with the application for the US NSF student support. Finally, we thank the authors of the papers and posters and all the attendees for their enthusiastic participation of the conference.

January 2013
Fengzhu Sun
Professor, University of Southern California, USA
Chair, RECOMB 2013 Program Committee

Table of Contents

Program

Program

April 7-10, 2013, the New Tsinghua Auditorium, Tsinghua University

April 6 (Saturday)	
18:00-20:00	**Reception (all registered participants), Wenjin Hotel**

April 7 (Sunday)	
9:00-18:00	On-site registration
10:00-10:10	**Opening Remarks** **Xuegong Zhang (Tsinghua University, China)**
10:10-11:10	**Keynote Talk. Chair: Sorin Istrail (Brown University, USA)** **Deborah A. Nickerson** **University of Washington, USA** ***Next-Generation Human Genetics***
11:10-12:10	**Session 1: Population Genomics** **Chair: Francis Chin (University of Hong Kong, China)**
11:10-11:30	Emrah Kostem and Eleazar Eskin *Efficiently Identifying Significant Associations in Genome-wide Association Studies*
11:30-11:50	Jesse Rodriguez, Serafim Batzoglou and Sivan Bercovici *An Accurate Method for Inferring Relatedness in Large Datasets of Unphased Genotypes via an Embedded Likelihood-Ratio Test*
11:50-12:10	Itamar Eskin, Farhad Hormozdiari, Lucia Conde, Chris Skibola, Jacques Riby, Eleazar Eskin and Eran Halperin *eALPS: Estimating Abundance Levels in Pooled Sequencing Using Available Genotyping Data*
12:10-14:30	**Lunch Break and Poster Session**
12:10-13:30	**Poster Setup**
13:30-14:30	**Poster Session 1**
14:30-15:30	**Keynote Talk. Chair: Xuegong Zhang (Tsinghua University, China)** **Chung-I Wu** **Chinese Academy of Sciences, China** **University of Chicago, USA** ***Active Migration Driving Evolution --- Adaptive Invasion and Diversification of Multifocal Tumors***
15:30-16:10	**Session 2: Evolutionary Genomics** **Chair: Louxin Zhang (National University of Singapore)**
15:30-15:50	Yufeng Wu *An Algorithm for Constructing Parsimonious Hybridization Networks with Multiple Phylogenetic Trees*

15:50-16:10	Jesper Jansson, Chuanqi Shen and Wing-Kin Sung *An Optimal Algorithm for Building the Majority Rule Consensus Tree*
16:10-16:30	Tea Break
16:30-17:50	**Session 3: Computational Methods for Genomics** **Co-Chairs: Roded Sharan (Tel Aviv University, Israel) and Lei Li (Chinese Academy of Sciences, China)**
16:30-16:50	Dan He, Zhanyong Wang, Laxmi Parida and Eleazar Eskin *IPED: Inheritance Path based Pedigree Reconstruction Algorithm using Genotype Data*
16:50-17:10	Shijian Chen, Anqi Wang and Lei Li *SEME: A Fast Mapper of Illumina Sequencing Reads with Statistical Evaluation*
17:10-17:30	Alina Munteanu and Raluca Gordan *Distinguishing between Genomic Regions Bound by Paralogous Transcription Factors*
17:30-17:50	Po-Ru Loh, Michael Baym and Bonnie Berger (highlight) *Compressive Genomics*
17:50	**Removal of Poster Session 1**

April 8 (Monday)

9:00-18:00	On-site registration
9:00-10:00	**Keynote Talk. Chair: Martin Vingron (Max-Planck-Institute for Molecular Genetics, Germany)** **Takashi Gojobori** **National Institute of Genetics, Japan** ***Time-Course Big Data in Environmental Metagenomics for Monitoring Dynamic Changes of Marine Microorganism Diversity***
10:00-10:40	**Session 4: Comparative Genomics** **Chair: Xiujie Wang (Institute of Genetics and Developmental Biology, China)**
10:00-10:20	Mukul S. Bansal, Eric J Alm and Manolis Kellis *Reconciliation Revisited: Handling Multiple Optima when Reconciling with Duplication, Transfer, and Loss*
10:20-10:40	Roy Ronen, Nitin Udpa, Eran Halperin and Vineet Bafna *Learning Natural Selection from the Site Frequency Spectrum*
10:40-11:00	Tea Break
11:00-12:20	**Session 5: Cancer Genomics** **Co-Chairs: Ben Raphael (Brown University, USA) and Jingdong Han (Chinese Academy of Sciences, China)**
11:00-11:20	Dong-Yeon Cho and Teresa Przytycka *Dissecting Cancer Heterogeneity with a Probabilistic Genotype-Phenotype Model*

11:20-11:40	Fabio Vandin, Alexandra Papoutsaki, Ben Raphael and Eli Upfal *Genome-Wide Survival Analysis of Somatic Mutations in Cancer*
11:40-12:00	Raheleh Salari, Syed Shayon Saleh, Dorna Kashef-Haghighi, David Khavari, Daniel E Newburger, Robert E West, Arend Sidow and Serafim Batzoglou *Inference of Tumor Phylogenies with Improved Somatic Mutation Discovery*
12:00-12:20	Layla Oesper, Ahmad Mahmoody and Ben Raphael *Estimating Tumor Purity and Cancer Subpopulations from High-Throughput DNA Sequencing Data*
12:20-14:30	**Lunch Break and Poster Session**
12:20-13:30	**Poster Setup**
13:30-14:30	**Poster Session 2**
14:30-16:10	**Session 6: RNA and Chromosome Structure I** **Co-Chairs: Michal Linial (Hebrew University of Jerusalem, Israel) and Xiaowo Wang (Tsinghua University, China)**
14:30-14:50	Raheleh Salari, Damian Wojtowicz, Jie Zheng, David Levens, Yitzhak Pilpel and Teresa Przytycka (highlight) *Teasing Apart Translational and Transcriptional Components of Stochastic Variations in Eukaryotic Gene Expression*
14:50-15:10	Sebastian Will, Christina Schmiedl, Milad Miladi, Mathias Mohl and Rolf Backofen *SPARSE: Quadratic Time Simultaneous Alignment and Folding of RNAs Without Sequence-Based Heuristics*
15:10-15:30	Evan Senter, Saad Sheikh, Ivan Dotu, Yann Ponty and Peter Clote *Using the Fast Fourier Transform to Accelerate the Computational Search for RNA Conformational Switches*
15:30-15:50	Vladimir Reinharz, Yann Ponty and Jerume Waldispuhl *A Linear Inside-Outside Algorithm for Correcting Sequencing Errors in Structured RNA Sequences*
15:50-16:10	Yitzhak Freindman, Ohad Balaga and Michal Linial (highlight) *Cellular Regulation by Teamwork of microRNAs*
16:10-16:30	Tea Break
16:30-17:10	**Session 7: RNA and Chromosome Structure II** **Chair: Xie Zhen (Tsinghua University, China)**
16:30-16:50	Audrey M. Michel, Kingshuk Roy Choudhury, Andrew E. Firth, Nicholas T. Ingolia, John F. Atkins and Pavel V. Baranov (highlight) *Observation of Dually Decoded Regions of the Human Genome Using Ribosome Profiling Data*
16:50-17:10	Zhizhuo Zhang, Guoliang Li, Kim-Chuan Toh and Wing-Kin Sung *Inference of Spatial Organizations of Chromosomes Using Semi-definite Embedding Approach and Hi-C Data*
17:10-18:00	**Business meeting (all participants)**
18:00	**Removal of Poster Session 2**

April 9 (Tuesday)	
9:00-13:30	On-site registration
9:00-10:00	**Keynote Talk. Chair: Michael Waterman (University of Southern California, USA)** **Scott Fraser** **University of Southern California, USA** ***Pushing the Limits of Biological Imaging, Sensing and Manipulation***
10:00-10:40	**Session 8: Molecular Networks I** **Chair: Minghua Deng (Peking University, China)**
10:00-10:20	Mohammad Javad Sadeh, Giusi Moffa and Rainer Spang *Considering Unknown Unknowns - Reconstruction of Non-confoundable Causal Relations in Biological Networks*
10:20-10:40	Ngoc Hieu Tran, Kwok Pui Choi and Louxin Zhang *Counting Motifs in the Entire Biological Network from Noisy and Incomplete Data*
10:40-11:00	Tea Break
11:00-12:20	**Session 9: Molecular Networks II** **Chairs: Hyunju Lee (Gwangju Institute of Science and Technology, Korea) and Rui Jiang (Tsinghua University, China)**
11:00-11:20	Dkriti Puniyani and Eric Xing *NP-MuScL: Unsupervised Global Prediction of Interaction Networks From Multiple Data Sources*
11:20-11:40	Ali A Faruqi, William A Bryant and John W Pinney *Analysis of Metabolic Evolution in Bacteria Using Whole-Genome Metabolic Models*
11:40-12:00	Qiangfeng Cliff Zhang, Donald Petrey, Lei Deng, Andrea Califano and Barry Honig (highlight) *On the Integration of Structural and Systems Biology: Structure-Based Prediction of Protein–Protein Interactions on a Genome-Wide Scale*
12:00-12:20	Shihua Zhang, Chun-Chi Liu, Wenyuan Li, Hui Shen, Peter W. Laird and Xianghong Jasmine Zhou (highlight) *Discovery of Multi-Dimensional Modules by Integrative Analysis of Cancer Genomic Data*
12:20-12:40	Bartek Wilczynski, Ya-Hsin Liu, Zhen Xuan Yeo and Eileen E. M. Furlong (highlight) *Predicting Spatial and Temporal Gene Expression Using an Integrative Model of Transcription Factor Occupancy and Chromatin State*
12:40-13:30	**Lunch Break**
13:30-17:30	**Great Wall Tour (all registered participants, reservation required)**
18:30-20:30	**Banquet (full registration only)**

April 10 (Wednesday)

9:00-17:10	On-site registration
9:00-10:00	**Keynote Talk. Chair: Fengzhu Sun (University of Southern California, USA)** **Nadia Rosenthal** **Monash University in Melbourne, Australia** ***Enhancing Mammalian Regeneration***
10:00-10:40	**Session 10: Epigenomics** **Chair: Teresa Przytycka (NIH, USA)**
10:00-10:20	Christopher Reeder and David Gifford *High Resolution Modeling of Chromatin Interactions*
10:20-10:40	Michael Stevens, Jeffrey Cheng, Mingchao Xie, Joseph Costello and Ting Wang *MethylCRF: an Algorithm for Estimating Absolute Methylation Levels at Single CpG Resolution from Methylation Enrichment and Restriction Enzyme Sequencing Methods*
10:40-11:00	Tea Break
11:00-12:20	**Session 11: Protein Structure** **Co-Chairs: Knut Reinert (Freie Universität Berlin, Germany) and Lusheng Wang (City University of Hong Kong, China)**
11:00-11:20	Chittaranjan Tripathy, Anthony Yan, Pei Zhou and Bruce Donald *Extracting Structural Information from Residual Chemical Shift Anisotropy: Analytic Solutions for Peptide Plane Orientations and Applications to Determine Protein Structure*
11:20-11:40	Fei Guo, Shuai Cheng Li, Wenji Ma and Lusheng Wang *Detecting Protein Conformational Changes in Interactions via Scaling Known Structures*
11:40-12:00	Dong Xu, Hua Li and Yang Zhang *Fast and Accurate Calculation of Protein Depth by Euclidean Distance Transform*
12:00-12:20	Shuigeng Zhou *Boosting Prediction Performance of Protein-Protein Interaction Hot Spots by Using Structural Neighborhood Properties*
12:20-13:15	**Lunch Break**
13:15-13:35	**Industry Talk. Chair: Jin Gu (Tsinghua University, China)** Hubert Ding Healthcare Leading Architect, Healthcare Strategy and Solution, Intel Corporation *Compute for Personalized Medicine*
13:40-14:40	**Keynote Talk. Chair: Ron Shamir (Tel Aviv University, Israel)** **Xiaoliang Sunney Xie** **Harvard University, USA** ***Life at the Single Molecule Level***

14:40-15:20	**Session 12: Image Analysis and Comparative Epigenomics** **Chair: Russell Schwartz (Carnegie Mellon University, USA)**
14:40-15:00	Yinyin Yuan, Henrik Failmezger, Oscar M. Rueda, H. Raza Ali, Stefan Gräf, Suet-Feung Chin, Roland F. Schwarz, Christina Curtis, Mark J. Dunning, Helen Bardwell, Nicola Johnson, Sarah Doyle, Gulisa Turashvili, Elen Provenzano, Sam Aparicio, Carlos Caldas and Florian Markowetz (highlight) *Quantitative Image Analysis of Cellular Heterogeneity in Breast Tumors Complements Genomic Profiling*
15:00-15:20	Sheng Zhong (highlight) *Comparative epigenomics*
15:20-15:40	Tea Break
15:40-17:00	**Session 13: Proteomics** **Co-Chairs: Simin He (Chinese Academy of Sciences, China) and Sungroh Yoon (Seoul National University, Korea)**
15:40-16:00	Kyowon Jeong, Sangtae Kim and Pavel Pevzner *UniNovo: a Universal Tool for De Novo Peptide Sequencing*
16:00-16:20	Sergey Nurk, Anton Bankevich, Dmitry Antipov,Alexey Gurevich, Anton Korobeynikov, Alla Lapidus,Andrey Prjibelsky, Alexey Pyshkin, Alexander Sirotkin,Yakov Sirotkin, Ramunas Stepanauskas, Jeffrey McLean, Roger Lasken, Scott R. Clingenpeel, Tanja Woyke, Glenn Tesler, Max A. Alekseyev, and Pavel A. Pevzner *Assembling Genomes and Mini-metagenomes from Highly Chimeric Reads*
16:20-16:40	Mingxun Wang and Nuno Bandeira *Spectral Library Generating Function for Assessing Spectrum-Spectrum Match Significance*
16:40-17:00	Xiaowen Liu, Shawna Hengel, Si Wu, Nikola Tolic, Ljiljana Pasa-Tolic and Pavel Pevzner *Identification of Proteins with Multiple Post-Translational Modifications Using Top-Down Tandem Mass Spectra*
17:00-17:10	**Closing Ceremony**

Committees

Conference Chair

Xuegong Zhang	Tsinghua University, China

Program Chair

Fengzhu Sun	University of Southern California, USA

Steering Committee

Vineet Bafna	University of California, San Diego, USA
Serafim Batzoglou	Stanford University, USA
Bonnie Berger	Massachusetts Institute of Technology, USA
Sorin Istrail	Brown University, USA
Michal Linial	The Hebrew University of Jerusalem, Israel
Martin Vingron (Chair)	Max Planck Institute for Molecular Genetics, Germany

Organizing Committee

Xuegong Zhang (co-chair)	Tsinghua University, China
Minghua Deng (co-chair)	Peking University, China
Rui Jiang (co-chair)	Tsinghua University, China
Jin Gu	Tsinghua University, China
Jingdong Han	PICB, CAS, China
Ruiqiang Li	Peking University, China
Xuan Li	SIBS, CAS, China
Yixue Li	Shanghai Bioinformatics Center, China
Jingchu Luo	Peking University, China
Geng Tian	Tsinghua University, China
Xiujie Wang	Institute of Genetics and Development, CAS, China
Peiheng Zhang	Institute of Computing, CAS, China

Program Committee

Tatsuya Akutsu	Kyoto University, Japan
Frank Alber	University of Southern California, USA
Max Alekseyev	University of South Carolina, USA
Kiyoshi Asai	University of Tokyo, Japan
Joel Bader	Johns Hopkins University, USA
Vineet Bafna	University of California, San Diego, USA
Ziv Bar-Joseph	Carnegie Mellon University , USA

Nuno Bandeira	University of California, San Diego, USA
Serafim Batzoglou	Stanford University, USA
Bonnie Berger	Massachusetts Institute of Technology, USA
Sebastian Böcker	Jena University, Demark
Michael Brent	Washington University, USA
Michael Brudno	University of Toronto, Canada
Dongbo Bu	Chinese Academy of Sciences, China
Kun-Mao Chao	National Taiwan University, Taiwan
Brian Chen	Lehigh University, Canada
Luonan Chen	Chinese Academy of Sciences, China
Phoebe Chen	La Trobe University, Australia
Ting Chen	University of Southern California, USA
Francis Chin	The University of Hong Kong, Hong Kong
Minghua Deng	Peking University, China
Nadia El-Mabrouk	University of Montreal, Canada
Mikhail Gelfand	Institute for Information Transmission Problems RAS, Russia
Eran Halperin	Tel-Aviv University, Israel
Si-Min He	Chinese Academy of Sciences, China
Wen-Lian Hsu	Academia Sinica, Taiwan
Heng Huang	University of Texas, USA
Daniel Huson	University at Tubingen, Germany
Sorin Istrail	Brown University, USA
Daniel Huson	University at Tubingen, Germany
Rui Jiang	Tsinghua University, China
Simon Kasif	Boston University, USA
Daniel Huson	University at Tubingen, Germany
Jens Lagergren	Royal Institute of Technology, Sweden
Doheon Lee	Korea Advanced Institute of Science and Technology, Korea
Hyunju Lee	Gwangju Institute of Science and Technology, Korea
Thomas Lengauer	Max Planck Institute for Informatics, Germany
Lei M Li	Chinese Academy of Sciences, China
Ming Li	University of Waterloo, Canada
Yixue Li	Shanghai Center for Bioinformation Technology, China
Michal Linial	Hebrew University, Israel
Jinze Liu	University of Kenturky, USA
Stefano Lonardi	University of California, Riverside, USA
Satoru Miyano	Tokyo University, Japan
Bernard Moret	Swiss Federal Institutes of Technology, Switzerland
William Noble	University of Washington, USA
Arlindo Oliveira	INESC-ID, Portugal

Teresa Przytcka	NIH NCBI, USA
Ben J Raphael	Brown University, USA
Knut Reinert	Freie Universitat Berlin, Germany
Marie-France Sagot	INRIA, France
Cenk Sahinalp	Simon Fraisher University, Canada
Russell Schwartz	Carnegie Mellon University, USA
Roded Sharan	Tel Aviv University, Israel
Mona Singh	Princeton University, USA
Steven Skiena	SUNY, USA
Andrew Smith	University of Southern California, USA
Peter Stadler	Universitat Leipzig, Germany
Yuzhen Ye	Indiana University, USA
Kai Tan	University of Iowa, USA
Chao Tang	Peking University, China
Martin Vingron	Max Planck Institute for Molecular Genetics, Germany
Jerome Waldispuhl	McGill University, Canada
Lusheng Wang	City University of Hongkong, China
Limsoog Wong	National University of Singapore, Singapore
Xiaohui Xie	UC Irvine, USA
Dong Xu	University of Missuri, USA
Yuzhen Ye	Indiana University, USA
Sungroh Yoon	Seoul National University, Korea
Shibu Yooseph	JCVI, USA
Louxin Zhang	National University of Singapore, Singapore
Michael Q Zhang	UT Dallas, USA and Tsinghua University, China
Xuegong Zhang	Tsinghua University, China

External Reviewers

Aguiar, Derek
Ahrne, Erik
Aliphanahi, Babak
Andreotti, Sandro
Antipov, Dmitry
Arndt, Peter
Askenazi, Manot
Atias, Nir
Bagherian, Misagh
Bahrami, Emad
Bandyopadhyay, Nirmalya
Bankevich, Anton
Baran, Yael
Baudet, Christian
Bazykin, Yegor
Becerra, David
Beglov, Dmitri
Behnam, Ehsan
Bernstein, Laurence
Bienkowska, Jadwiga
Blin, Guillaume
Bozdag, Serdar
Bryant, David
Buske, Orion
Chang, Chia-Jung
Chen, Ching-Tai
Chen, Tiffany
Chen, Yi-Ching
Cheng, Cheng-Wei
Cheng, Chia-Ying
Chiu, Ka Ho

Cho, Dongyeon
Choi, Jeong-Hyeon
Choi, Kwok Pui
Chowdhury, Salim
Chu, An-Chiang
Clevert, Djork-Arné
Costello, James
Daley, Timothy
Daniels, Noah
Dao, Phuong
David, Matei
Dieterich, Christoph
Dondi, Riccardo
Donmez, Nilgun
Doose, Gero
Duma, Denisa
Ermakova, Ekaterina
Eskin, Itamar
Fischer, Martina
Fiser, Andras
Fleischauer, Markus
Frånberg, Mattias
Fu, Yan
Fuentes, Gloria
Gao, Jianjiong
Gao, Xin
Gautheret, Daniel
Gitter, Anthony
Golan, David
Guthals, Adrian
Hach, Faraz
Hajirasouliha, Iman
Halldorsson, Bjarni
Halloran, John
Han, Buhm
Hao, Xiaolin
Harris, Elena
He, Danning
He, Xin
He, Zengyou
He, Zhiquan
Henry, Henry
Hoffman, Michael
Holtby, Dan
Hosur, Raghavendra
Howbert, Jeff
Hu, Jialu
Hu, Yin
Huang, Yan
Hwang, Woochang
Irannia, Zohreh
Jiang, Shuai
Joshi, Trupti
Kaell, Lukas
Kalinina, Olga
Kaplan, Tommy
Katenka, Natallia
Kehr, Birte
Khrameeva, Ekaterina
Kim, Yoo-Ah
Kim, Younghoon
Kirkpatrick, Bonnie
Klau, Gunnar W.
Kochetov, Alex
Korkin, Dmitry
Korobeynikov, Anton
Kozakov, Dima
Kulikov, Alexander
Kyriazopoulou-Panagiotopoulou, Sofia
Lacroix, Vincent
Lam, Henry
Le, Hai-Son
Lee, Sael
Lee, Sejoon
Lee, Sunjae
Lehmann, Kjong
Leiserson, Mark
Lemaitre, Claire
Leung, Henry
Li, Ning
Li, Shuai Cheng
Li, Wei
Li, Weizhong
Liao, Chung-Shou
Libbrecht, Max
Lin, Wei-Yin
Lin, Yen Yi
Liu, Yan
Liu, Yizhou
Love, Michael
Lu, Bingwen
Lynn, Ke-Shiuan
Ma, Bin
Ma, Wenxiu
Madhusudhan, M.S.
Mahlab, Shelly
Makeev, Vsevolod
Mammana, Alessandro
Markowetz, Florian
Marti-Renom, Marc
Mazza, Arnon
Medvedev, Paul
Meusel, Marvin
Mezlini, Aziz
Mirebrahim, Seyed
Mironov, Andrey
Misra, Navodit
Molla, Michael
Navlakha, Saket
Ng, Kal Yen Kaow
Nikolenko, Sergey
Numanagic, Ibrahim
Nurk, Sergey
Oesper, Layla
Panchin, Alexander
Parviainen, Pekka
Pasaniuc, Bogdan
Pelossof, Raphael
Peng, Jian
Pfeifer, Nico
Pham, Son
Polishko, Anton
Qu, Jenny
Rahnenführer, Jörg
Rajasekaran, Rajalakshmi
Rampasek, Ladislav
Rappoport, Nadav
Ray, Pradipta
Rho, Mina
Richard, Hugues

Ritz, Anna
Rodriguez, Jesse
Roman, Theodore
Rozov, Roye
Sacomoto, Gustavo
Salari, Raheleh
Sayyed, Auwn
Scheubert, Kerstin
Schulz, Marcel
Sheng, Quanhu
Sheridan, Paul
Shibuya, Tetsuo
Shimamura, Teppei
Shiraishi, Yuichi
Silverbush, Dana
Sinaimeri, Blerina
Sindi, Suzanne
Sjöstrand, Joel
Snedecor, June
Souaiaia, Tade
Steffen, Martin
Stegle, Oliver
Subramanian, Ayshwarya
Sul, Jae Hoon
Sun, Ruping
Sun, Shiwei
Swenson, Krister
Tannier, Eric
Thomas-Chollier, Morgane
Tjong, Harianto
Tofigh, Ali
Tran, Ngoc Hieu
Uren, Philip
Vandin, Fabio
Varoquaux, Nelle
Wan, Lin
Wang, Hung-Lung
Wang, Jian
Wang, Kendric
Wang, Mingxun
Wang, Yi
Wang, Yunfei
Wise, Aaron
Wojtowicz, Damian
Wojtowicz, Danian
Xu, Jinbo
Yamaguchi, Rui
Yan, Xifeng
Yeger-Lotem, Esti
Yu, Zhaoxia
Zeng, Feng
Zhang, Chao
Zhang, Jingfen
Zhang, Jiyang
Zhang, Shihua
Zheng, Jie
Zheng, Yu
Zhong, Shan
Zinman, Guy

Keynote Speakers

Deborah Nickerson

Professor of Genome Sciences
Adjunct Professor of Bioengineering
University of Washington School of Medicine
USA

Next-generation Human Genetics

The application of massively parallel sequencing is rapidly providing new insights into the genetic basis of both rare and common human diseases. I will highlight findings from the analysis of rare Mendelian diseases from the Centers for Mendelian Genomics, where new candidate genes are being uncovered at an unprecedented pace. For common, complex human diseases, I will provide an overview of the strategies that are being applied and some initial findings emerging from the large-scale application of next-generation sequencing. I will also highlight where novel algorithms could impact these approaches. Lastly, large-scale sequencing is providing new insights into the demographic and evolutionary forces that have shaped the allelic architecture of the protein coding regions in humans. I will discuss the importance of these findings for personalized human genetics.

Chung-I Wu

Professor
Director, Beijing Institute of Genomics
Chinese Academy of Sciences
China
Department of Ecology and Evolution
University of Chicago
USA

Active Migration Driving Evolution - Adaptive Invasion and Diversification of Multifocal Tumors

The ability to migrate leading to range expansion should be beneficial to populations of individuals or cells. However, since emigration to a new site is rarely adaptive for the migrants themselves, the selective advantage of migration-prone genotypes remains unclear. In this report, we study cell migration in hepatocellular carcinoma (HCC), which can be multifocal with several tumors in the same liver. Whole-exome sequencing was applied to 12 such cases with 84 sections sequenced and/or genotyped. The clonal relationships in 75% of the cases follow a spatial pattern in which cell migration precedes clonal expansion and lineage diversification. Early emigration before tumor growth raises questions about the advantage of cell migration, which is classified as either simple dispersal or invasion. In invasion, the colonizing cells, much like invasive species, have higher fitness than those of the native site. In dispersal, fitness remains unchanged. By proposing a new method for inferring adaptive evolution in cell lineages, we show that invasive migration characterizes 7 of the 8 informative HCC cases. Each proliferation event is adaptively driven by a different set of mutations. The selective advantage of migration is due to a mutual reinforcement of tumor-growth and cell-motility mutations. The mutual reinforcement leads to a process of adaptive diversification that accelerates as tumors evolve. This evolutionarily interesting process is also clinically relevant.

Biosketch:

Prof. Wu obtained his Bachelor's degree from Tunghai Univ, Taiwan in 1976, and his Ph.D. degree from Univ of British Columbia, Canada in 1982. In 1998, he became a professor at the Univ of Chicago and served as the chair of the Department of Ecology and Evolution untill 2006. He was appointed the director of Beijing Institute of Genomics, CAS in 2008.

Takashi Gojobori

Professor
Center for Information Biology
National Institute of Genetics, Mishima
Japan

Time-Course Big Data in Environmental Metagenomics for Monitoring Dynamic Changes of Marine Microorganism Diversity

Environmental metagenomics is a genomic approach in which genomic fragments of any species contained in environmental samples such as a bottle of sea water and a cup of land soil are sequenced without morphological identification of those species in order to observe ecological features of a diversity of microorganisms. In this practice, specific primers such as 16S and 18S rRNA genes are usually used for amplification so that species identification can be easily done through the database. Moreover, when this kind of conventional metagenomics is applied for studies of marine microorganism diversity, it has been usual to observe species composition of microorganisms at a given single time in given locations.

Here, we developed a new approach for metagenomics called the "Digital DNA Chip," in which we can identify a compositional profile in silico of a large set of DNA sequences obtained from shot gun sequencing of any genomic segments of microorganisms in the sea water. We also developed a water-floating device for continuously taking the water samples from the sea water in different depths at a series of time points in various locations in a given sea area. Using this new device, we have initiated a project for conducting environmental evaluation of the sea water over continuous time points, in addition to understanding of dynamic features of marine microorganism diversity.

Here, we present the progress of our project, particularly focusing on how to handle the Big Data that are obviously produced in the present project.

Biosketch:

Takashi Gojobori is a Vice-Director of National Institute of Genetics (NIG) and Professor at Center for Information Biology and DNA Data Bank of Japan (DDBJ) in NIG, Mishima, Japan.

After finishing his Ph.D.(1979) at Kyushu University, Japan, he has been Research Associate and Research Assistant Professor at the University of Texas at Houston for 4 years (1979-1983). He has also experienced a Visiting Assistant Professor at Washington University in St. Louis (1985, 1986) and a Visiting Research Fellow at Imperial Cancer Research Fund (ICRF) in London (1989).

He has been elected as Member of Academia Europaea in 2012. He has received The Medal with Purple Ribbon from the Government of Japan in 2009. He has been an Academician member of the Pontifical Academy of Science in Vatican (2007) and a Foreign Honorary Member of American Academy of Arts and Sciences (2006) and as a Fellow of American Association for the Advancement of Science (AAAS) (2006). He has received the Salvatore Gold Medal from Italy (2004) and some of other Japanese awards.

He is a Section Editor of BMC Genomics, the Founding Editor of Genome Biology and Evolution, the Editor of GENE and FEBS Letters, and Associate Editor of Molecular Biology and Evolution and PLoS Genetics. He has also served on the editorial boards of 6 international journals. He was an Editor of Journal of Molecular Evolution for 8 years (1995-2003).

He has about 400 publications in the peer-reviewed international journals on comparative and evolutionary genomics. He has worked extensively on the rates of synonymous and nonsynonymous substitutions, positive selection, horizontal gene transfer, viral evolution, genomic evolution, and comparative gene expressionics. In recent years, he has focused on evolution of the central nervous system and sensory organs. He has also contributed to the DDBJ/GenBank/EMBL database construction as well as the H-Invitational human gene database.

Scott E. Fraser

Professor
Department of Biological Sciences and
Department of Biomedical Engineering
University of Southern California
USA

Pushing the Limits of Biological Imaging, Sensing and Manipulation

Advances in cell biology, genomics and proteomics offer unprecedented knowledge of the constituents within cells and the means by which the cells can interact. The triumphs of these reductionistic approaches are not without limitations: "-omics" approaches typically analyze homogenates of millions of cells, washing out heterogeneities between cells in the population; physical interactions between molecular components are deduced from cross-linking or co-precipitation; the approaches miss the dynamics of interactions, signals and motions that provide important insights into biological processes. Advanced imaging techniques provide powerful means of answering these challenges, and integrating the growing genome-scale data sets into a mechanistic understanding of the molecular and cellular mechanisms that underlie key biological events.

Our own work has concentrated on developing instruments capable of sensing cellular and molecular events during complex processes such as embryogenesis and organogenesis, with the goals of studying the events in the most relevant setting of the intact system. Such intravital imaging is challenged by major tradeoffs between resolution, the rapidity of data collection and the limited photon budget.

Intravital techniques for imaging, combined with better sensors and true molecular imaging offer the promise of testing the predictions made by high throughput approaches. This work offers the promise of the direct measurement of cellular and molecular events, harvesting meaning from the temporal and spatial heterogeneities that plague modern "-omics" analyses.

Nadia Rosenthal

Professor
Australian Regenerative Medicine Institute
Monash University, Melbourne
Australia
National Heart and Lung Institute
Imperial College London
UK

Enhancing Mammalian Regeneration

What lies behind the remarkable potential of some organisms to rebuild themselves after injury, and why aren't we better at it? Our approach has been to tinker with pathways activated in the response to damage, disease and ageing in mammals, reducing the impediments to effective regeneration. Using mouse genetics we have investigated the role of growth factors and resident cells in the resolution of tissue injury, uncovering a complex interaction between local repair mechanisms and cells of the immune system, which participate in the removal of necrotic cells but may also provide support for the action of stem and progenitor cells in the repair process. These interventions support the feasibility of improving human regenerative capacity by modulating key signaling pathways controlled by specific components of the immune system, providing new targets for clinical intervention and improving prospects for molecular and cellular combination therapies.

Biosketch:

Nadia Rosenthal obtained her PhD from Harvard Medical School, where she later directed a biomedical research laboratory, serving for a decade at the New England Journal of Medicine as editor of the Molecular Medicine series. She headed the EMBL Mouse Biology program in Rome from 2001-2012 and holds a Professorship of Cardiovascular Science at Imperial College London. She is an EMBO member, was awarded the Ferrari-Soave Prize in Cell Biology and Doctors Honoris Causa from the Pierre and Marie Curie University in Paris and the University of Amsterdam. She is currently Founding Director of the Australian Regenerative Medicine Institute at Monash University and Scientific Head of EMBL Australia. Her internationally recognized research program focuses on the role of growth factors and stem cells in tissue regeneration. Professor Rosenthal is an NH&MRC Australia Fellow.

Xiaoliang Sunney Xie

Professor
Department of Chemistry and Chemical Biology
Harvard University
USA

Life at the Single Molecule Level

DNA, which pass genetic information from generation to generation, are single molecules in individual cells. Consequently, gene expression is stochastic. Recent single-molecule live-cell experiments have allowed the mechanisms of stochastic gene expression to be understood at the molecular level. Point mutation and copy number variation, which are two major stochastic changes of DNA, can now be studied at the single cell level by advances in whole genome amplification and sequencing.

Biosketch:

Prof. Xie recieved his B.S. degree in chemistry from Peking University, Beijing, and his Ph.D. Degree in chemistry from University of California, San Diego. He was a postdoctoral fellow of University of Chicago in 1990-1992, and a senior research scientist to chief scientist in the Pacific Northwest National Laboratory in 1992-1998. Dr. Xie was a professor of chemistry of Harvard University in 1999-2008, and became a Cheung Kong Visiting Professor of the Biodynamic Optical Imaging Center of Peking University in 2009 and a Mallinckrodt Professor of Chemistry and Chemical Biology of Harvard University since 2009.

Accepted Papers

Accepted Papers

Reconciliation Revisited: Handling Multiple Optima when Reconciling with Duplication, Transfer, and Loss
Mukul S. Bansal, Eric J Alm and Manolis Kellis

SEME: A Fast Mapper of Illumina Sequencing Reads with Statistical Evaluation
Shijian Chen, Anqi Wang and Lei Li

Dissecting Cancer Heterogeneity with a Probabilistic Genotype-Phenotype Model
Dong-Yeon Cho and Teresa Przytycka

eALPS: Estimating Abundance Levels in Pooled Sequencing Using Available Genotyping Data
Itamar Eskin, Farhad Hormozdiari, Lucia Conde, Chris Skibola, Jacques Riby, Eleazar Eskin and Eran Halperin

Analysis of Metabolic Evolution in Bacteria Using Whole-Genome Metabolic Models
Ali A Faruqi, William A Bryant and John W Pinney

Detecting Protein Conformational Changes in Interactions via Scaling Known Structures
Fei Guo, Shuai Cheng Li, Wenji Ma and Lusheng Wang

IPED: Inheritance Path based Pedigree Reconstruction Algorithm using Genotype Data
Dan He, Zhanyong Wang, Laxmi Parida and Eleazar Eskin

An Optimal Algorithm for Building the Majority Rule Consensus Tree
Jesper Jansson, Chuanqi Shen and Wing-Kin Sung

UniNovo: A Universal Tool for *de Novo* Peptide Sequencing
Kyowon Jeong, Sangtae Kim and Pavel Pevzner

Efficiently Identifying Significant Associations in Genome-wide Association Studies
Emrah Kostem and Eleazar Eskin

Identification of Proteins with Multiple Post-Translational Modifications Using Top-Down Tandem Mass Spectra
Xiaowen Liu, Shawna Hengel, Si Wu, Nikola Tolic, Ljiljana Pasa-Tolic and Pavel Pevzner

Distinguishing between Genomic Regions Bound by Paralogous Transcription Factors
Alina Munteanu and Raluca Gordan

Assembling Genomes and Mini-metagenomes from Highly Chimeric Reads
Sergey Nurk, Anton Bankevich, Dmitry Antipov,Alexey Gurevich, Anton Korobeynikov, Alla Lapidus,Andrey Prjibelsky, Alexey Pyshkin, Alexander

Sirotkin,Yakov Sirotkin, Ramunas Stepanauskas, Jeffrey McLean, Roger Lasken, Scott R. Clingenpeel, Tanja Woyke, Glenn Tesler, Max A. Alekseyev, and Pavel A. Pevzner

Inferring Intra-Tumor Heterogeneity from High-Throughput DNA Sequencing Data
Layla Oesper, Ahmad Mahmoody and Ben Raphael

NP-MuScL: Unsupervised Global Prediction of Interaction Networks from Multiple Data sources
Kriti Puniyani and Eric Xing

High Resolution Modeling of Chromatin Interactions
Christopher Reeder and David Gifford

A Linear Inside-Outside Algorithm for Correcting Sequencing Errors in Structured RNA Sequences
Vladimir Reinharz, Yann Ponty and Jérôme Waldispühl

An Accurate Method for Inferring Relatedness in Large Datasets of Unphased Genotypes via an Embedded Likelihood-Ratio Test
Jesse Rodriguez, Serafim Batzoglou and Sivan Bercovici

Learning Natural Selection from the Site Frequency Spectrum
Roy Ronen, Nitin Udpa, Eran Halperin and Vineet Bafna

Considering Unknown Unknowns - Reconstruction of Non-confoundable Causal Relations in Biological Networks
Mohammad Javad Sadeh, Giusi Moffa and Rainer Spang

Inference of Tumor Phylogenies with Improved Somatic Mutation Discovery
Raheleh Salari, Syed Shayon Saleh, Dorna Kashef-Haghighi, David Khavari, Daniel E Newburger, Robert E West, Arend Sidow and Serafim Batzoglou

Using the Fast Fourier Transform to Accelerate the Computational Search For RNA Conformational Switches
Evan Senter, Saad Sheikh, Ivan Dotu, Yann Ponty and Peter Clote

MethylCRF, an Algorithm for Estimating Absolute Methylation Levels at Single CpG Resolution from Methylation Enrichment and Restriction Enzyme Sequencing Methods
Michael Stevens, Jeffrey Cheng, Mingchao Xie, Joseph Costello and Ting Wang

Counting Motifs in the Entire Biological Network from Noisy and Incomplete Data
Ngoc Hieu Tran, Kwok Pui Choi and Louxin Zhang

Extracting Structural Information from Residual Chemical Shift Anisotropy: Analytic Solutions for Peptide Plane Orientations and Applications to Determine Protein Structure
Chittaranjan Tripathy, Anthony Yan, Pei Zhou and Bruce Donald

Genome-Wide Survival Analysis of Somatic Mutations in Cancer
Fabio Vandin, Alexandra Papoutsaki, Ben Raphael and Eli Upfal

Spectral Library Generating Function for Assessing Spectrum-Spectrum Match Significance
Mingxun Wang and Nuno Bandeira

SPARSE: Quadratic Time Simultaneous Alignment and Folding of RNAs without Sequence-Based Heuristics
Sebastian Will, Christina Schmiedl, Milad Miladi, Mathias Möhl and Rolf Backofen

An Algorithm for Constructing Parsimonious Hybridization Networks with Multiple Phylogenetic Trees
Yufeng Wu

Fast and Accurate Calculation of Protein Depth by Euclidean Distance Transform
Dong Xu, Hua Li and Yang Zhang

Inference of Spatial Organizations of Chromosomes Using Semi-definite Embedding Approach and Hi-C Data
Zhizhuo Zhang, Guoliang Li, Kim-Chuan Toh and Wing-Kin Sung

Boosting Prediction Performance of Protein-Protein Interaction Hot Spots by Using Structural Neighborhood Properties
Shuigeng Zhou

Accepted Highlights

Accepted Highlights

Compressive Genomics

Po-Ru Loh[1,†], Michael Baym[1,2,†,*], & Bonnie Berger[1,*]

[1]: Department of Mathematics and Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA.

[2]: Department of Systems Biology, Harvard Medical School, Boston, Massachusetts, USA.

[†]: P.-R.L. and M.B. contributed equally to this work..

[*]: To whom correspondence should be addressed.

Emails: P.-R.L. (ploh@mit.edu); M.B. (baym@mit.edu); B.B. (bab@mit.edu)

The past two decades have seen an exponential increase in sequencing capabilities, outstripping advances in computing power. Extracting new insights from the datasets currently being generated will require not only faster computers; it will require smarter algorithms. However, most genomes currently sequenced are highly similar to ones already collected; thus the amount of novel sequence information is growing much more slowly. We show that this redundancy can be exploited by compressing the data in a way as to allow direct computation on the compressed data. This approach reduces the computational task of operating on many similar genomes to only slightly more than that of operating on just one. Moreover, its relative advantage over existing algorithms grows with the accumulation of future genomic data. We demonstrate here this compressive architecture by implementing versions of both BLAST and BLAT, and emphasize how compressive genomics, more generally, will enable biologists to keep pace with current data.

Reference

1. Po-Ru Loh, Michael Baym, & Bonnie Berger. Compressive Genomics. *Nature Biotechnology*, **30**: 627-630, 2012.

Teasing Apart Translational and Transcriptional Components of Stochastic Variations in Eukaryotic Gene Expression

Raheleh Salari [1,#], Damian Wojtowicz [2,3,#], Jie Zheng [4,5,#], David Levens [6], Yitzhak Pilpel [7], Teresa M. Przytycka [2,*]

[1]: Department of Computer Science, Stanford University, California, USA.
[2]: National Center for Biotechnological Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland, USA.
[3]: Institute of Informatics, University of Warsaw, Warsaw, Poland.
[4]: Bioinformatics Research Centre (BIRC), School of Computer Engineering, Nanyang Technological University, Singapore.
[5]: Genome Institute of Singapore, A*STAR, Biopolis, Singapore.
[6]: Laboratory of Pathology, National Cancer Institute, Bethesda, Maryland, USA.
[7]: Department of Molecular Genetics, Weizmann Institute of Science, Rehovot, Israel.
[#]: These authors contributed equally to this work.
[*]: To whom correspondence should be addressed.
Emails: RS (rahelehs@stanford.edu); DW (wojtowda@ncbi.nlm.nih.gov); JZ (zhengjie@ntu.edu.sg); DL (levens@helix.nih.gov); YP (pilpel@weizmann.ac.il); TMP (przytyck@ncbi.nlm.nih.gov)

The intrinsic stochasticity of gene expression leads to cell-to-cell variations, noise, in protein abundance. Several processes, including transcription, translation, and degradation of mRNA and proteins, can contribute to these variations. Recent single cell analyses of gene expression in yeast have uncovered a general trend where expression noise scales with protein abundance. This trend is consistent with a stochastic model of gene expression where mRNA copy number follows the random birth and death process. However, some deviations from this basic trend have also been observed, prompting questions about the contribution of gene-specific features to such deviations. For example, recent studies have pointed to the TATA box as a sequence feature that can influence expression noise by facilitating expression bursts. Transcription-originated noise can be potentially further amplified in translation. Therefore, we asked the question of to what extent sequence features known or postulated to accompany translation efficiency can also be associated with increase in noise strength and, on average, how such increase compares to the amplification associated with the TATA box. Untangling different components of expression noise is highly nontrivial, as they may be gene or gene-module specific. In particular, focusing on codon usage as one of the sequence features associated with efficient translation, we found that ribosomal genes display a different relationship between expression noise and codon usage as compared to other genes. Within nonribosomal genes we found that sequence high codon usage is correlated with increased noise relative to the average noise of proteins with the same abundance. Interestingly, by projecting the data on a theoretical model of gene expression, we found that the amplification of noise strength associated with codon usage is comparable to that of the TATA box, suggesting that the effect of translation on noise in eukaryotic gene expression might be more prominent than previously appreciated.

References

1. Raheleh Salari, Damian Wojtowicz, Jie Zheng, David Levens, Yitzhak Pilpel, Teresa M. Przytycka. Teasing Apart Translational and Transcriptional Components of Stochastic Variations in Eukaryotic Gene Expression. *PLoS Computational Biology*, 8(8): e1002644, 2012.

Cellular regulation by teamwork of microRNAs

Yitzhak Friedman[1], Ohad Balaga [2], Michal Linial[1,*]

[1]: Department of Biological Chemistry, Institute of Life Sciences, The Hebrew University of Jerusalem, Israel

[2]: School of Computer Science and Engineering, The Hebrew University of Jerusalem, Israel

[*]: To whom correspondence should be addressed.

Emails: YF (yitzhak.friedman@mail.huji.ac.il); OB (ohad.balaga@mail.huji.ac.il); ML (michall@cc.huji.ac.il)

A longstanding goal in biology is the detailed understanding of the principles that govern gene regulation. To date, our understanding of the principles that underlie regulation by microRNAs (miRNAs) remains fragmentary. MicroRNAs (miRNAs) negatively regulate the levels of messenger RNA (mRNA) post-transcriptionally. Recent advances in CLIP technology allowed the capturing of miRNAs with their cognate mRNAs. Consequently, thousands of validated mRNA–miRNA pairs have been revealed. Our publication[1] proposes a paradigm shift accordingly; although individual miRNAs are not particularly specific to their targets, system-wide efficiency at the cellular level results from a synergistic playing-together phenomenon. We compare the predictive power of miRror that is an inherently combinatorial miRNA prediction tool, in view of raw data from CLIP-Seq and microarray experiments. We extend this view to suggest that miRNA regulation is achieved by a combination of miRNAs acting on specific pathways. Specifically, the potency of global miRNA regulation can be determined by understanding the pathway graph topology. A systematic approach that identifies pairs and triplets of miRNAs that maximally affect cellular pathways will be presented.

We implemented combinatorial and statistical constraints in the miRror2.0 algorithm. miRror estimates the likelihood of combinatorial miRNA activity in explaining the observed data. We tested the success of miRror in recovering the correct miRNA from transcriptomic profiles of cells overexpressing a miRNA, and in identifying hundreds of genes from miRNA sets, which are observed in CLIP experiments. We demonstrate that the success of miRror in recovering the miRNA regulation is superior in respect to a dozen leading miRNA-target prediction algorithms. We further describe the balance between alternative modes of joint regulation that are executed by pairs of miRNAs. Finally, manipulated cells were tested for the possible involvement of miRNA in shaping their transcriptomes. We identify instances in which the observed transcriptome can be explained by a combinatorial regulation of miRNA pairs. We conclude that the joint operation of miRNAs is an attractive strategy to maintaine cell homeostasis and overcome the low specificity inherent in individual miRNA–mRNA interaction.

References

1. Yitzhak Friedman, Ohad Balaga, Michal Linial. Toward a combinatorial nature of microRNA regulation in human cells. *Nucleic Acids Research* 40(19) 9404–9416, 2012.

Observation of dually decoded regions of the human genome using ribosome profiling data.

Audrey M. Michel [1], Kingshuk Roy Choudhury [2], Andrew E. Firth [3], Nicholas T. Ingolia [4], John F. Atkins [1,5], Pavel V. Baranov [1,*]

[1]: Biochemistry Department, University College Cork, Ireland.

[2]: Department of Statistics, University College Cork, Ireland.

[3]: Department of Pathology, University of Cambridge, Cambridge, UK.

[4]: Department of Embryology, Carnegie Institution for Science, Baltimore, MD, USA.

[5]: Human Genetics Department, University of Utah, Salt Lake City, UT, USA.

*: To whom correspondence should be addressed.

Emails: AMM (a.mannionmichel@umail.ucc.ie); KRC (kingshuk@ucc.ie); AEF (aef24@cam.ac.uk); NTI (ingolia@ciwemb.edu); JFA (atkins@genetics.utah.edu); PVB (P.Baranov@ucc.ie)

The recently developed ribosome profiling technique (Ribo-Seq) allows mapping of the locations of translating ribosomes on mRNAs with sub-codon precision. When ribosome protected fragments (RPFs) are aligned to mRNA, a characteristic triplet periodicity pattern is revealed. We utilized the triplet periodicity of RPFs to develop a computational method for detecting transitions between reading frames that occur during programmed ribosomal frameshifting or in dual coding regions where the same nucleotide sequence codes for multiple proteins in different reading frames. Application of this method to ribosome profiling data obtained for human cells allowed us to detect several human genes where the same genomic segment is translated in more than one reading frame (from different transcripts as well as from the same mRNA) and revealed the translation of hitherto unpredicted coding open reading frames.

Reference

1. Audrey M. Michel, Kingshuk Roy Choudhury, Andrew E. Firth, Nicholas T. Ingolia, John F. Atkins, Pavel V. Baranov. Observation of dually decoded regions of the human genome using ribosome profiling data. *Genome Res*. 22: 2219-2229, 2012.

On the integration of structural and systems biology: structure-based studies of protein-protein interactions on a genome-wide scale

Qiangfeng Cliff Zhang[1,2,3 *], Donald Petrey[1,2,3], Lei Deng[2,3,4], Andrea Califano[2,3,5,6] & Barry Honig[1,2,3]

[1]: Howard Hughes Medical Institute, Columbia University, New York, New York 10032, USA.

[2]: Department of Biochemistry and Molecular Biophysics, Columbia University, New York, New York 10032, USA.

[3]: Center for Computational Biology and Bioinformatics, Columbia Initiative in Systems Biology, Columbia University, New York, New York 10032, USA.

[4]: Department of Computer Science and Technology, Tongji University, Shanghai 201804, China.

[5]: Institute of Cancer Genetics, Columbia University, New York, New York 10032, USA.

[6]: Department of Biomedical Informatics, Columbia University, New York, New York 10032, USA.

[*]: Present address: CCSR 2155, 269 Campus Drive, Stanford, CA 94305-5168

Emails: QCZ (qczhang@stanford.edu), DP (dsp18@columbia.edu), LD (leideng9@gmail.com), AC (califano@c2b2.columbia.edu) & BH (bh6@columbia.edu)

Knowledge of protein-protein interactions (PPIs) is essential to understanding cell regulatory mechanisms. Much of presently known PPIs derive from high-throughput techniques as well as from manual curations of experiments on individual systems. However, comparative studies suggest that these data sets contain many false interactions and are largely incomplete in that many interactions have not been identified. The same is true for PPIs inferred from the many computational tools that have been developed. Here we show that geometric relationships between protein structures, including both PDB structures and homology models, can be used to accurately predict PPIs on a genome-wide scale. Furthermore, an algorithm, PrePPI, that combines structural information with non-structural clues yields predictions of comparable quality to high-throughput experiments. Experimental tests of a number of predictions demonstrated that the PrePPI algorithm can identify unsuspected PPIs in biological systems of considerable interest. The surprising effectiveness of three-dimensional structural information can be attributed to the use of homology models and the exploitation of both close and remote geometric relationships between proteins. Our results constitute a significant paradigm shift in both structural and systems biology and suggest that they can be integrated to an extent that has not been possible in the past.

References

1. Qiangfeng Cliff Zhang, Donald Petrey, Lei Deng, Li Qiang, Yu Shi, Chan Aye Thu, Brygida Bisikirska, Celine Lefebvre, Domenico Accili, Tony Hunter, Tom Maniatis, Andrea Califano & Barry Honig. Structure-based prediction of protein–protein interactions on a genome-wide scale. *Nature* 490, 556–560, 2012.

Discovery of multi-dimensional modules by integrative analysis of cancer genomic data

Shihua Zhang[1,*], Chun-Chi Liu[2], Wenyuan Li[3], Hui Shen[4], Peter W. Laird[4] and Xianghong Jasmine Zhou[3,*]

[1]: Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China
[2]: Institute of Genomics and Bioinformatics, National Chung Hsing University, Taiwan, ROC
[3]: Program in Molecular and Computational Biology, University of Southern California, Los Angeles, CA, USA
[4]: USC Epigenome Center, University of Southern California, Los Angeles, CA, USA
[*]: To whom correspondence should be addressed.

Emails: SZ (zsh@amss.ac.cn); CCL (jimliu@nchu.edu.tw); WL (wel@usc.edu); HS (shenhui1986@gmail.com); PWL (plaird@usc.edu); XJZ (xjzhou@usc.edu)

Recent technology has made it possible to simultaneously perform multi-platform genomic profiling (e.g., DNA methylation, and gene expression) of biological samples, resulting in so-called ``multi-dimensional genomic data". Such data provide unique opportunities to study the coordination between regulatory mechanisms on multiple levels. However, integrative analysis of multi-dimensional genomics data for the discovery of combinatorial patterns is currently lacking. Here, we adopt a joint matrix factorization technique to address this challenge. This method projects multiple types of genomic data onto a common coordinate system, in which heterogeneous variables weighted highly in the same projected direction form a multi-dimensional module. Genomic variables in such modules are characterized by significant correlations and likely functional associations. We applied this method to the DNA methylation, gene expression, and microRNA expression data of 385 ovarian cancer samples from the TCGA project. These multi-dimensional modules revealed perturbed pathways that would have been overlooked with only a single type of data, uncovered associations between different layers of cellular activities, and allowed the identification of clinically distinct patient subgroups. Our study provides a useful protocol for uncovering hidden patterns and their biological implications in multi-dimensional ``omic" data.

References

1. Shihua Zhang, Chun-Chi Liu, Wenyuan Li, Hui Shen, Peter Laird, Xianghong Jasmine Zhou. Discovery of multi-dimensional modules by integrative analysis of cancer genomic data. *Nucleic Acids Research*, 40(19): 9379-9391, 2012.

2. Shihua Zhang, Qingjiao Li, Juan Liu, Xianghong Jasmine Zhou. A novel computational framework for simultaneous integration of multiple functional genomic data to identify microRNA-gene regulatory modules. *Bioinformatics (ISMB2011)*, 27:i401-i409, 2011.

Predicting Spatial and Temporal Gene Expression Using an Integrative Model of Transcription Factor Occupancy and Chromatin State

Bartek Wilczynski [1,2*], Ya-Hsin Liu[1], Zhen Xuan Yeo[1], Eileen E. M. Furlong [1*]

1. EMBL Heidelberg, Meyerhofstrasse 1, 69117 Heidelberg, Germany
2. Institute of Informatics, University of Warsaw, Banacha 2, 02-089 Warsaw, Poland

*: To whom correspondence should be addressed

e-mails: Bartek Wilczynski (bartek@mimuw.edu.pl), Eileen Furlong (furlong@embl.de)

Abstract:

Precise patterns of spatial and temporal gene expression are central to metazoan complexity and act as a driving force for embryonic development. While there has been substantial progress in dissecting and predicting *cis*-regulatory activity, our understanding of how information from multiple enhancer elements converge to regulate a gene's expression remains elusive. This is in large part due to the number of different biological processes involved in mediating regulation as well as limited availability of experimental measurements for many of them. Here, we used a Bayesian approach to model diverse experimental regulatory data, leading to accurate predictions of both spatial and temporal aspects of gene expression. We integrated whole-embryo information on transcription factor recruitment to multiple *cis*-regulatory modules, insulator binding and histone modification status in the vicinity of individual gene loci, at a genome-wide scale during *Drosophila* development. The model uses Bayesian networks to represent the relation between transcription factor occupancy and enhancer activity in specific tissues and stages. All parameters are optimized in an Expectation Maximization procedure providing a model capable of predicting tissue- and stage-specific activity of new, previously unassayed genes. Performing the optimization with subsets of input data demonstrated that neither enhancer occupancy nor chromatin state alone can explain all gene expression patterns, but taken together allow for accurate predictions of spatio-temporal activity. Model predictions were validated using the expression patterns of more than 600 genes recently made available by the BDGP consortium, demonstrating an average 15-fold enrichment of genes expressed in the predicted tissue over a naïve model. We further validated the model by experimentally testing the expression of 20 predicted target genes of unknown expression, resulting in an accuracy of 95% for temporal predictions and 50% for spatial.

References

1. Bartek Wilczynski, Ya-Hsin Liu, Zhen Xuan Yeo, Eileen E. M. Furlong, Predicting Spatial and Temporal Gene Expression Using an Integrative Model of Transcription Factor Occupancy and Chromatin State, PLoS Computational Biology, 8(12): e1002798, 2012, doi:10.1371/journal.pcbi.1002798

Quantitative image analysis of cellular heterogeneity in breast tumors complements genomics profiling

Yinyin Yuan[1,2,a†,], Henrik Failmezger[3,4*], Oscar M. Rueda[1,2*], H. Raza Ali[1,2*], Stefan Gräf[1,2,b], Suet-Feung Chin[1,2], Roland F. Schwarz[1,2], Christina Curtis[5], Mark J. Dunning[1], Helen Bardwell[1], Nicola Johnson[6], Sarah Doyle[6], Gulisa Turashvili[7,8], Elena Provenzano[9], Sam Aparicio[7,8], Carlos Caldas[1,2,9,10], and Florian Markowetz[1,2†]

[1] Cancer Research UK Cambridge Research Institute, Cambridge, CB2 0RE, UK;
[2] Department of Oncology, University of Cambridge, Cambridge, CB2 2XZ, UK;
[3] Max Planck Institute for Plant Breeding Research, Carl-von-Linné-Weg 10, 50829 Cologne, Germany;
[4] Gene Center and Department of Biochemistry, Center for Integrated Protein Science CIPSM, Ludwigs-Maximilian University, Munich, 81377, Germany;
[5] Department of Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles, CA 90033, USA;
[6] Department of Histopathology, Cambridge University Hospital NHS Foundation Trust (Addenbrooke's Hospital), Cambridge, CB2 OQQ, UK;
[7] Department of Pathology and Laboratory Medicine, University of British Columbia, Vancouver, G227-2211, Canada;
[8] Molecular Oncology, British Columbia Cancer Research Centre, Vancouver, V5Z 1L3, Canada;
[9] Cambridge Breast Unit, Addenbrooke's Hospital, Cambridge University Hospital NHS Foundation Trust and NIHR Cambridge Biomedical Research Centre, Cambridge, CB2 2QQ, UK;
[10] Cambridge Experimental Cancer Medicine Centre, Cambridge, CB2 0RE, UK;
[a] The Institute of Cancer Research, London, SW3 6JB, UK;
□ These authors contributed equally
† E-mail: florian.markowetz@cancer.org.uk, yinyin.yuan@icr.ac.uk

Solid tumors are heterogeneous tissues composed of a mixture of cancer and normal cells, which complicates the interpretation of their molecular profiles. Furthermore, tissue architecture is generally not reflected in molecular assays, rendering this rich information underused. To address these challenges, we developed a computational approach based on standard hematoxylin and eosin–stained tissue sections and demonstrated its power in a discovery and validation cohort of 323 and 241 breast tumors, respectively. To deconvolute cellular heterogeneity and detect subtle genomic aberrations, we introduced an algorithm based on tumor cellularity to increase the comparability of copy number profiles between samples. We next devised a predictor for survival in estrogen receptor–negative breast cancer that integrated both image-based and gene expression analyses and significantly outperformed classifiers that use single data types, such as microarray expression signatures. Image processing also allowed us to describe and validate an independent prognostic factor based on quantitative analysis of spatial pat- terns between stromal cells, which are not detectable by molecular assays. Our quantitative, image-based method could benefit any large-scale cancer study by refining and complementing molecular assays of tumor samples.

Reference

Yuan Y. et al., Quantitative image analysis of cellular heterogeneity in breast tumors complements genomics profiling. *Sci Transl Med*, 4, 157ra143, 2012.

Comparative Epigenomics

Sheng Zhong [1]

[1]: Department of Bioengineering, University of California San Diego, USA.

Email: SZ (szhong@ucsd.edu)

Despite the explosive growth of genomic data, functional annotation of regulatory sequences remains difficult. We contributed to introduce “comparative epigenomics” - interspecies comparison of DNA and histone modifications - as an approach for annotation of the regulatory genome. We assayed and compiled in human, mouse, and pig pluripotent stem cells the genomic distributions of cytosine methylation, H2A.Z, H3K4me1/2/3, H3K9me3, H3K27me3, H3K27ac, H3K36me3, transcribed RNAs, and P300, TAF1, OCT4, and NANOG binding. A Comparative Epigenome Browser was developed to allow the public to interact with the data (http://www.cepbrowser.org). We observed that epigenomic conservation was strong in both rapidly evolving and slowly evolving DNA sequences, but not in neutrally evolving sequences. We posit that the conserved colocalization of different epigenomic marks can be used to discover regulatory sequences. Indeed, seven pairs of epigenomic marks identified exhibited regulatory functions. Thus, comparative epigenomics reveals regulatory features of the genome that cannot be discerned from sequence comparisons alone[1].

It remains unclear what can be learned from the temporal changes of the epigenome. We developed a finite mixture model of HMMs to cluster genomic sequences based on the similarity of temporal changes of multiple epigenomic marks during a cellular differentiation process (open source software at http://systemsbio.ucsd.edu/GATE). We differentiated mouse embryonic stem (ES) cells into mesendoderm cells. At three time points during this differentiation process, we used high-throughput sequencing to measure 7 histone modifications and variants, 2 DNA modifications including 5-mC and 5-hmC, and transcribed mRNAs and non-coding RNAs (ncRNAs). Genomic sequences were clustered based on the spatiotemporal epigenomic information. These clusters not only clearly distinguished gene bodies, promoters, and enhancers, but also were predictive of bidirectional promoters, miRNA promoters, and piRNAs. Several rules on combinatorial epigenomic changes and their effects on mRNA expression and ncRNA expression were derived, including a simple rule governing the relationship between 5-hmC and gene expression levels. A Sox17 enhancer containing a FOXA2 binding site and a Foxa2 enhancer containing a SOX17 binding site were identified, suggesting a positive feedback loop between the two mesendoderm transcription factors. These data illustrate the power of using epigenome dynamics to investigate regulatory functions[2].

References

1. Shu Xiao, Dan Xie, Xiaoyi Cao, Pengfei Yu, Xiaoyun Xing, Chieh-Chun Chen, Meagan Musselman, Mingchao Xie, Franklin D. West, Harris A. Lewin, Ting Wang, Sheng Zhong. COMPARATIVE EPIGENOMIC ANNOTATION OF REGULATORY DNA. *Cell*, 149: 1381-1391, 2012.
2. Pengfei Yu, Shu Xiao, Xiaoyun Xin, Chun-Xiao Song, Wei Huang, Darina McDee, Tetsuya Tanaka, Ting Wang, Chuan He, Sheng Zhong. SPATIOTEMPORAL CLUSTERING OF EPIGENOME REVEALS RULES OF DYNAMIC GENE REGULATION. *Genome Research*, 23:352-384, 2013.

Accepted Posters

Poster Session 1

April 7 (12:10-13:30, setup. 13:30-14:30, session. 17:50, removal)

Poster Session 2

April 8 (12:20-13:30, setup. 13:30-14:30, session. 18:00, removal)

RseqFlow: A more easy and flexible pipeline for RNA-Seq data analysis

Ying Wang[1,4,*], Lin Liu[1], Meifang Zhu[1], Rajiv Mayani[2], James A Knowles[3], Ewa Deelman[2], Ting Chen[4]

[1]: Dept. of Automation, Xiamen University, China.

[2]: Information Science Institute, University of Southern California, USA.

[3]: Keck Medical School, University of Southern California, USA.

[4]: Program of Computational Molecular Biology, University of Southern California, USA.

[*]: Ying Wang

E-mail: Ying Wang (wangying@xmu.edu.cn)

We have developed an RNA-Seq analysis workflow for single-ended Illumina reads, termed RseqFlow. This workflow attempts to integrate more analytical functions than the previous tools, and at the same time, be flexible and easy to use.

The whole framework has four running branches, which can be run simultaneously or individually: Branch 1, based on the merging of alignments to transcriptome and genome, the Quality Control, SNP Calling can be implemented; Branch 2, based on only the alignment to transcriptome, expression level for Gene/Exon/Splice Junction will be produced; Branch 3, some files format conversion for easy store, backup and visualization; Branch 4, differentially expressed gene identification based on the output of Branch 2. The whole pipeline is applicable for various species with proper reference sequences and annotation files.

The pipeline offer two running modes: Mode 1, Unix running mode, targets the analysis of a small amount of datasets or trial. This running mode integrated all the coding with unix shell scripts. And all the required softwares are included in the downloaded package, which will installed easily. For each branch, one shell command line is implemented for the whole running. Mode 2, Virtual Machine running mode, targets the large amount of datasets or runnings. It manages the source code and whole running process with Pegasus Workflow Management Service. It helps workflow execute in different kinds of different environments including desktops, campus clusters, grids, and clouds.

The RseqFlow is available from http://code.google.com/p/rseqflow/ and http://genomics.isi.edu/rnaseq/documents .

HMM for predicting single nucleotide variants from next generation sequencing

Jiawen Bian[1,2], Chenglin Liu[1], Hongyan Wang[1], Jing Xing[2] and Xiaobo Zhou[1,*]

[1]: Department of Radiology, The Methodist Hospital Research Institute, Weil Cornell Medical College, Houston, TX 77030, USA.

[2]: School of Mathematics and Physics, China University of Geosciences, Wuhan, China, 430074.

[*]: Corresponding author: XZhou@tmhs.org

The rapid development of next generation sequencing (NGS) technology provides a new chance for genomic exploration and research. Single nucleotide variants inferring from next generation sequencing is expected to reveal gene mutations in cancer. However, next generation sequencing has lower sequence coverage and poor single nucleotide variant (SNV) detection capability in the regulatory regions. Post probabilistic based methods are efficient for SNV-detection of high coverage region or sequence data with high depth. The performance of SNV detection for data with low sequencing depth remains poor and needs to be improved. We developed a new algorithm based on a discrete hidden Markov model (HMM) to infer the mutation for each position on the genome. We incorporated the mapping quality of each read covering the position on the genome and the corresponding base quality on each reads covering the stated position into the emission probability of HMM. As such, the context information from the whole observation as well as the confidence of the observations can both be made full use for mutation inferring on considered genome, which gains more probability power over the detection methods basing only on post probability, thus is very useful for SNV-detection for data with low sequencing depth. Moreover, HMM was validated by two sets of lobular breast tumor data and tested against two sets of Myelodysplastic Syndromes data, and compared with a recently published SNV calling algorithm SNVMix2(Goya, R. et al., 2010). HMM improved the performance of SNVMix2 largely when the sequencing depth is low and also outperformed SNVMix2 when SNVMix2 is well trained by a large dataset.

CASmap: Splitting Short Reads Alignment with FPGA-based Streamline Optimization

Shaoping Ling [1], Jiahan Liu [2], Lingtong Hao [1], Longhui Yin [3], Lili Dong [1], Lihua Cao [1], Wei Zou [1], Fen Xiao [3], Junsuo Zhao [2], Chung-I Wu [1,*], Xuemei Lu [1,*]

[1]: Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing, China.

[2]: Institute of Software, Chinese Academy of Sciences, Beijing, China.

[3]: Key Laboratory of Intelligent Computing & Information Processing of Ministry of Education, Xiangtan University, Xiangtan, China.

[*]: To whom correspondence should be addressed.

Emails: SL (spling@big.ac.cn); JL (liujiahan@gmail.com); LH (haolt@big.ac.cn) ; LY (yinlonghui.big@gmail.com) ; LD (donglili@big.ac.cn) ; LC (caolh@big.ac.cn) ; WZ (zouwei@big.ac.cn) ; FX (xiaof@xtu.edu.cn) ; JZ (junsuo@iscas.ac.cn) ; CW (wuci@big.ac.cn) ; XL (luxm@big.ac.cn)

Short reads alignment, as a core computational issue in HTS (High-Throughput Sequencing) data analysis, it is a bottle-neck in HTS data real-time clinical applications. We created a new alignment system (CASmap) which implemented BWT-based alignment algorithm in a customized desktop reconfigurable computer based on FPGA reconfigurable platform. It accelerated ~30X and ~2X higher than BWA (one thread) and SOAP3 in alignment in suffix array with the power of FPGA-based streamline optimization. Multi-threading parallelization of smith-waterman algorithm was implemented in multi-core host of CPU to verify the location of reads. CASmap achieved the high speed of ~310Gbp/day,cpu in aligning real human whole genome sequencing Hiseq pair-end reads (4 mismaches/read, 2×100bp) with the low power consumption of ~30w/Gb and high accuracy of 99%. CASmap, as an efficient reconfigurable heterogeneous computing system for short read alignment, provided a new green computing framework for HTS genome re-sequencing projects and a solution for real-time HTS data application.

References

1. Li H. and Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler Transform. *Bioinformatics, Epub.* (2010)
2. C. Liu, T. Wong, E. Wu, R. Luo, S. Yiu, Y. Li, B. Wang,C. Yu, X. Chu, K. Zhao, R. Li, and T. Lam, SOAP3:Ultra-fast GPU-based parallel alignment tool for short reads, *Bioinformatics*. (Jan. 2012)

GWIPS-viz: Development of a ribo-seq genome browser

Audrey M. Michel [1], Gearoid Fox [1,2], Pavel V. Baranov [1,*]

[1]: Biochemistry Department, University College Cork, Ireland.

[2]: Conway Institute, University College Dublin, Ireland.

[*]: To whom correspondence should be addressed.

Emails: AMM (a.mannionmichel@umail.ucc.ie); GF (gearoidfox@gmail.com); PVB (P.Baranov@ucc.ie)

Ribosome profiling (ribo-seq) is a recently developed technique that provides Genome Wide Information on Protein Synthesis (GWIPS) in vivo. It is based on the deep sequencing of ribosome protected mRNA fragments which allows the ribosome density along all mRNA transcripts present in the cell to be quantified. Since its inception, ribo-seq has been carried out in a number of eukaryotic and prokaryotic organisms. Due to the increasing interest in ribo-seq, there is a pertinent demand for a dedicated ribo-seq genome browser. Therefore we have developed GWIPS-viz, an online genome browser for visualizing ribosome profiling data (http://gwips.ucc.ie/). GWIPS-viz is based on the UCSC Genome Browser. Ribo-seq tracks coupled with mRNA-seq tracks are currently available for human, mouse, zebrafish, and yeast. Although still in early stage development, our objective is to continue incorporating ribo-seq datasets so that the wider community can readily visualize ribosome profiling information without the need to carry out computational processing.

Profiling Microbial Community Compositions of Lake Taihu with NGS

Yucai Fan[1], Liyang Liu[2], Junyi Zhang[1], Junfeng Li[2], Hongfei Cui[2], Kaixuan Tian[1], Xuegong Zhang[2*], Zuhong Lu[1*]

[1]: State Key Lab for Bioelectronics, School of Biological Science and Medical Engineering, Southeast University, Nanjing, China.

[2]: MOE Key Lab of Bioinformatics; Bioinformatics Division / Center for Synthetic and Systems Biology, TNLIST; Department of Automation, Tsinghua University, Beijing, China.

[*]: To whom correspondence should be addressed.

Emails: zhangxg (zhangxg@tsinghua.edu.cn); zhlu (zhlu@seu.edu.cn)

Fresh water is an important natural resource necessary for the survival of all ecosystems. Cyanobacterial bloom in lakes is a serious environmental problem in China. Lake Taihu, the third largest lake in China, is a typical shallow freshwater lake located in east China, which has serious environmental problem caused by cyanobacterial bloom [1].

The SSU rRNA gene (also known as the 16S rRNA gene) is widely used in studies of microbial ecology as a ‘‘barcode gene’’ to quantify microbial community structure and diversity. New developed high-throughput next-generation sequencing (NGS) technologies can provide a cost-effective means of identifying the microbial phylotypes and comparing microbiomes [2,3].

In this study, we intended to determine the composition of microbial community during the development, degradation and absence of cyanobacterial blooms of Taihu Lake by amplifying and sequencing of the V6 region of the 16S rRNA genes from the collected microbials. Totally 81 samples were collected at 9 well-selected sampling sites through a full year (9 saperate time points). By using a variant of the barcoding strategy and sequencing with the Illumina GAIIx platform, the samples can be analyzed in unprecedented depth. The QIIME software package is applied to analyze the results.

The previous studies have shown that many environmental factors such as pH, nutrient concentrations, temperature, and water flow covary, etc, will determine the distribution of taxa of microbials in freshwater systems. In order to reveal relationships between the appearance of bacterial groups and environmental variables from the lakes, a redundancy analysis (RDA) was used with the software CANOCO. We will also obtain possible determinative factors affecting the bacterioplankton fingerprints in the Taihu Lake, and explore relationship between the bacterial community composition and cyanobacterial bloom in the lake.

References

1. Wu, X., et al., Bacterial community composition of a shallow hypertrophic freshwater lake in China, revealed by 16S rRNA gene sequences. *FEMS Microbiology Ecology*. 61(1): 85-96, 2007.

2. Huber, J.A., et al., Microbial Population Structures in the Deep Marine Biosphere. *Science*, 318(5847): 97-100, 2007.

3. Jiang B. et al, Integrating next-generation sequencing and traditional tongue diagnosis to determine tongue coating microbiome, *Scientific Reports*, 2: 936, 2012.

VERSE: A Varying Effect Regression for Splicing Elements Discovery

Jing Zhang [1], C.-C. Jay Kuo [1], Liang Chen [2,*]

[1]: Ming Hsieh Department of Electrical Engineering, University of Southern California, Los Angeles, California.

[2]: Molecular and Computational Biology, Department of Biological Sciences, University of Southern California, Los Angeles, California.

[*]: To whom correspondence should be addressed.

Emails: Jing Zhang (zhang28@usc.edu); C.-C. Jay Kuo (cckuo@sipi.usc.edu); Liang Chen (liang.chen@usc.edu)

Identification of splicing regulatory elements (SREs) deserves special attention because these cis-acting short sequences are vital parts of splicing code. The fact that a variety of other biological signals cooperatively govern the splicing pattern indicates the necessity of developing novel tools to incorporate information from multiple sources to improve splicing factor binding sites prediction. Under this context, we proposed a Varying Effect Regression for Splicing Elements (VERSE) to discover intronic SREs in the proximity of exon junctions by integrating other biological features. As a result, 1562 intronic SREs were identified in 16 human tissues, many of which overlapped with experimentally verified binding motifs for several well-known splicing factors, including FOX-1, PTB, hnRNP A/B, hnRNP F/H, and so on. The discovered tissue, region, and conservation preferences of the putative motifs demonstrate that splice site selection is a complicated process that needs subtle and delicate regulation. VERSE may serve as a powerful tool to not only discover SREs by incorporating additional informative signals but also precisely quantify their varying contribution under different biological contexts.

An efficient random overlapping pool design for next generation sequencing based rare variant identification

Chang-Chang Cao, Xiao Sun*

State Key Laboratory of Bioelectronics, School of Biological Science and Medical Engineering, Southeast University, Nanjing, China.

*: To whom correspondence should be addressed.

Emails: Chang-Chang Cao (caochch@gmail.com); Xiao Sun (xsun@seu.edu.cn);

Motivation: Identification of rare variants through large scale resequencing is important for understanding complex disease. Population studies that involve sequencing thousands of individual genomes for charactering rare variants are still unaffordable until now, regardless of the drop in the price of next generation sequencing (NGS). Nevertheless group testing (GT) based overlapping pool design which greatly reduces sequencing pools to identify all the individuals helps to solve this problem.

Results: Here, taking advantage of quantitative information in sequencing results, we propose a random overlapping pool design algorithm that enables efficient recovery of variant carriers in groups of individuals with less cost. First of all, the optimal depths of coverage for pooled sequencing are computed based on a mathematic model. Random k-set pool design is used with appropriate selected parameters to guarantee the efficiency. Utilizing the information of reads number, we design a fast heuristic probability decoding algorithm to classify variant carriers. The results of simulation experiments with DNA preparation bias and sequencing errors indicate that our method performs similar or better compared with other previous algorithms in all aspects, including the amount of DNA libraries, data requirement and mixing procedure.

References

1. Shental, N., Amir, A. & Zuk, O. Identification of rare alleles and their carriers using compressed se (que) nsing. *Nucleic Acids Research* **38**, e179-e179, 2010.
2. Erlich, Y., Gordon, A., Brand, M., Hannon, G. J. & Mitra, P. P. Compressed genotyping. *Information Theory, IEEE Transactions on* **56**, 706-723, 2010.
3. Bruno, W. J. *et al.* Efficient pooling designs for library screening. *Genomics* **26**, 21-30, 1995.
4. Hwang, F. Random k-set pool designs with distinct columns. *Probability in the Engineering and Informational Sciences* **14**, 49-56, 2000.
5. Das, S. Binary Solutions for Overdetermined Systems of Linear Equations. *arXiv preprint arXiv:1101.3056* , 2011.

A Conditional Random Field-based model for clustering genes according to their RNA-Seq expression profile

Mohamed Nadhir DJEKIDEL [1], Xiaoning GAO[3], Yonghui LI[3], Yang CHEN [1], Li YU [3], Michael Q ZHANG [1,2,*]

[1]: MOE Key Laboratory of Bioinformatics; Bioinformatics Division/Center for Synthetic & System Biology, TNLIST; Department of Automation; Tsinghua University, Beijing 100084, China.

[2]: Department of Molecular and Cell Biology, Center for Systems Biology, The University of Texas at Dallas, Richardson, Texas 75080, USA.

[3]: Department of Hematology and BMT Center, Chinese PLA General Hospital, Beijing, China.

[*]: To whom correspondence should be addressed.

Emails: M.N.DJEKIDEL (nde12@mails.tsinghua.edu.cn); X.N.GAO (gaoxn@263.net); Y.H.LI (yonghuililab@yahoo.com.cn); Y.CHEN (yc@mail.tsinghua.edu.cn); L.YU (chunhuiliyu@yahoo.com); M.Q.ZHANG (michael.zhang@utdallas.edu)

In diseases or biological processes in general, genes that perform a similar cellular process tend to interact together. However, the identification of these groups is still a challenging question.

Consequently, techniques ranging from graph-based to biological-based methods have been developed to tackle this problem [1]. In this work, we developed a new clustering technique based on the new probabilistic graphical model paradigm of Conditional Random Field (CRF)[2] in order cluster an interaction network taking into account both the biological and topological characteristics of it nodes given their RNA-Seq interaction profile.

The expression profiles of genes in different samples were obtained from RNA-Req experiments, then, we used these datasets to filter-out edges in the protein interaction network based on their Jensen-Shannon distance between their expressions. After that, we clustered the network using a CRF-based model that considered three features: the density of the clusters, the intra-cluster GO similarity and inter-clusters GO dissimilarity.

We applied our method on six leukemia samples, in which 3 of them are t(8,21) translocation positive and 3 are negative. Our method was able to group many leukemia related genes reported in the literature. The comparison of our results to some classical clustering methods showed better topological and biological performance.

References

1. Xiaoli Li, Min Wu, Chee-Keong Kwoh, and See-Kiong Ng. Computational approaches for detecting protein complexes from protein interaction networks: a survey. *BMC Genomics*, 11(S3), 2010.
2. John D. Lafferty, Andrew McCallum, and Fernando C.N.Pereira. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. *In Proceedings of the Eighteenth International Conference on Machine Learning (ICML '01).* 282-289, 2001

A new method for STR genotyping based on NGS technology

Junji Li [1], Jing Tu [1], Zuhong Lu [1,*]

[1]: State Key Laboratory of Bioelectronics, School of Biological Science and Medical Engineering, Southeast University, Nanjing, 210096, China

[*]: To whom correspondence should be addressed. Email: zhlu@seu.edu.cn

With NGS technology entering in more and more fields of biological research and medical applications, the related detecting methods have been developing by taking advantages of the parallel detection of massive samples with high-speed and high-throughput characteristics. In this study, we proposed a new procedure based on NGS technology to detect and genotype classic STRs which is currently realized by capillary electrophoresis [1]. The new method was mainly divided into three steps: library preparation, repeat number detection and image analysis. In library preparation, magnetic beads were used to fix STRs onto separated locations in sequencing chip. Specific primers were modified on beads to guarantee the amplification of different STR fragments in a multiplex PCR system. Also, we added barcodes to distinguish STRs from different test samples.

A special detection proposal was designed to realize STR genotyping on NGS instruments. Two sequencing primers were hybridized with STR sequence at first; therefore, synthetic reaction would be limited to the core area of each STR where the counts of repeats were reasonable. As we used a back primer to end synthetic reaction, the enzyme should not have 5' exonuclease activity. During the sequencing detection, we added three natural dNTPs with a special modified dNTP. If the core repeat (reverse chain) was "AATG" for example, then the special dNTP would be dTTP or dGTP, which was single in the repeat, therefore, counting repeats could be simplified as counting the special nucleotide sites. For the efficiency of detection, we modified the special dNTP with a cleavable fluorophore and a 3' reversible terminator. Synthetic reaction would automatically pause at every site to be detected, then excited the fluorophore and took photos. At the end of each pause, cleaved fluorophore and restored extension until the entire core area sequenced.

Image results were consist of visions of fluorescence signals in every detecting cycles. Three types of signals were meaningful: ①bright spots represented the beads on which core repeats were still to be counted in current sequencing cycle. ②less bright spots meant that the STRs on these beads were alleles. For the shorter allele, repeat counting had already end, however, the longer one could still be detected. ③dark spots should be sorted as counting completed spots and invalid spots.

This new method can greatly increase the sample numbers of STR genotyping, thus greatly reducing the detection cost and time. That means the new method can realize applications which need to detect massive samples, such as criminal database establishment and genetic identity information acquisition. Also, the detection principle of this new method can be adjusted and applied to different NGS platform, or even some third-generation sequencing platforms, like SMRT.

References

1. Katherine L, et al. Genotyping of forensic short tandem repeat (STR) systems based on sizing precision in a capillary electrophoresis instrument. *Electrophoresis*, 19:86-93, 1998

RNAseqViewer: A New Software Program for RNA-seq Data Visualization

Xavier Rogé [1], Xuegong Zhang [1,2,*]

[1]: MOE Key Laboratory of Bioinformatics, Bioinformatics Division and Center for Synthetic and Systems Biology, TNLIST / Department of Automation, Tsinghua University, Beijing 100084, China.

[2]: School of Life Sciences, Tsinghua University, Beijing 100084, China.

[*]: To whom correspondence should be addressed.

Emails: XR (xavier.roge@gmail.com); XZ (zhangxg@tsinghua.edu.cn)

New advances in RNA sequencing have opened up new horizons in the field of transcriptomics and given access to new extensive data. The analysis of these data needs effective visualization tools, so as scientists can gain an insight into the data and are able to review the results of the computational tools. We developed a new software program, RNAseqViewer, to visualize the various data from the RNA-Seq analyzing process for single or multiple samples. By focusing on expression of genes and transcript isoforms, the program offers innovative ways to present the transcriptome data in a quantitative and interactive manner.

RNAseqViewer currently supports 7 types of data: read alignments (SAM/BAM format) and junction reads (BED), which can be provided by RNA-Seq mappers like TopHat; transcripts (GTF), which can be computed by tools like Cufflinks; numeric data (Wiggle); reference sequences (FASTA) and annotations (RefSeq); and generic BED tracks. Different types of view for each data set allow the visualization of different levels of information, including heatmap-like views for informative and yet very compact tracks, making possible to visualize dozens of samples simultaneously. Special attention has been given to the user interface, so that the data can be explored in a fast and intuitive way, and to the memory management, so that very large data sets can be visualized without exceeding memory limits nor affecting the fluidity of the user interface.

The software is a handy tool for scientists who use RNA-Seq data to compare multiple transcriptomes, for example, to compare gene expression and alternative splicing of cancer samples or of different development stages.

SWAP-Assembler: A Scalable De Bruijn Graph Based Assembler for Massive Genome Data

Jintao Meng [1,2], Bingqiang Wang [2], Yanjie Wei [1,*], Shengzhong Feng [1,*], Jiefeng Cheng [1], Pavan Balaji [3]

[1]: Shenzhen Institutes of Advanced Technology, CAS, Shenzhen, P.R. China.

[2]: Beijing Genomics Institutes Shenzhen, P.R. China.

[3]: Mathematics and Computer Science Division, Argonne National Laboratory, USA

[*]: Dr. Wei and Prof. Feng are corresponding authors.

Emails: Jintao Meng (jt.meng@siat.ac.cn); Bingqiang Wang (wangbingqiang@genomics.cn); Shengzhong Feng (sz.feng@siat.ac.cn); Jiefeng Cheng (jf.cheng@siat.ac.cn); Pavan Balaji (balaji@mcs.anl.gov)

Sequencing species with large genome can produce Tara bytes data, and the de bruijn graph constructed from these data - in some cases having ten billions of vertices and edges - poses challenges to genome assembly problem. This paper presents a multi-step bi-directed graph (MSG) to abstract the standard genome assembly (SGA) problem. With MSG, SGA can be decomposed into several edge merging operations, and this operation and the multi-step semi-extended edges are proved to be a semi-group. Afterwards a small world asynchronous parallel model (SWAP), which can automatically detect and make use of the locality of computation and communication in semi-group to maximize potential parallelism, is proposed for this type of computation. With MSG and SWAP, SWAP-assembler is developed, the scalability test shows that it can scale up to 1024 cores with improved performance, the 2008 Asian (YanHuang) genome can be assembled in 2 hours, which is 6 times faster than SOAPdenovo on one server with 32 cores, and about 24 times faster than ABySS with 1024 cores.

NURD: A New Tool for Estimating Isoform Expression from Non-Uniform RNA-Seq Data

Xinyun Ma [1], Xuegong Zhang [*1,2]

[1]: MOE Key Laboratory of Bioinformatics, Bioinformatics Division and Center for Synthetic and Systems Biology, TNLIST / Department of Automation, Tsinghua University, Beijing 100084, China.

[2]: School of Life Sciences, Tsinghua University, Beijing 100084, China.

[*]: To whom correspondence should be addressed.

Emails: XM (maxy218@gmail.com); XZ (zhangxg@tsinghua.edu.cn)

RNA-Seq technology has been used widely in transcriptome study and one of the most important applications is to estimate the expression level of genes and their alternative splicing isoforms. There have been several algorithms published to estimate the expression based on different models, but most of them assumed uniform distribution of sequencing reads. Recently Wu et. al. published an algorithm for non-uniform read distribution. It can accurately estimate isoform expression levels by modeling position-related sequencing biases in a nonparametric manner [1]. But an efficient program to implement the algorithm is still lacking.

We developed an efficient implementation of the algorithm in the program NURD. Our program can correct both the global tendency of sequencing bias in the data and local sequencing bias specific to each gene. The correction makes the isoform expression estimation more reliable. NURD is a free tool to estimate the isoform expression level from RNA-Seq data. Given the reads mapping result and gene annotation file, NURD will output the expression estimation result. NURD is proved to be both effective and efficient on isoform level expression estimation by experiments on simulated and real RNA-Seq data. NURD is mainly implemented in C++ code and can be run on Unix/Linux operation system with GCC/G++ compiler. The package is freely available for academic use at http://bioinfo.au.tsinghua.edu.cn/software/NURD/.

References

1. Zhengpeng Wu, Xi Wang, Xuegong Zhang. Using non-uniform read distribution models to improve isoform expression inference in RNA-Seq. *Bioinformatics*, 27(4): 502-508, 2011.

A statistical method to infer tumor purity, ploidy and absolute copy numbers from next generation sequencing data

Lei Bao [1,*], Minya Pu [1], Karen Messer [1,*]

[1]: Division of Biostatistics, Moores UCSD Cancer Center, University of California, San Diego, CA 92093.

[*]: To whom correspondence should be addressed.

Emails: LB (lebao@ucsd.edu); MP (mpu@ucsd.edu); KM (kmesser@ucsd.edu)

Detection and quantification of the DNA copy number alterations (CNAs) in tumor cells is challenging because the DNA specimen is extracted from a mixture of tumor and normal stromal cells. Estimates of tumor purity (0-100%) and ploidy (reference value of 2.0 for diploid genome) are necessary to correctly infer CNAs and ploidy may itself be a prognostic factor in cancer progression. As deep sequencing of the exome or genome has become routine for characterization of tumor samples, in this work we aim to develop a simple and robust algorithm to infer purity, ploidy and absolute copy numbers in whole numbers from sequencing data. The algorithm first extracts depth of coverage from the aligned reads using VARSCAN2 [1], then segments the read depth data, and finally fits a regression model to the segmented data. A simulation study shows that the algorithm is robust against the existence of subclonal populations. We compared our algorithm to a well established SNP array based method called ABSOLUTE [2] on eleven breast cancer samples. Our method has high concordance to ABSOLUTE with a mean squared error of 0.01 and 0.11 for purity and ploidy respectively. Our method hence may offer a simple solution to the CNA quantification for cancer sequencing projects.

References

1. Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L, Miller CA, Mardis ER, Ding L, Wilson RK: VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res* 2012, 22(3):568-576.
2. Carter SL, Cibulskis K, Helman E, McKenna A, Shen H, Zack T, Laird PW, Onofrio RC, Winckler W, Weir BA et al: Absolute quantification of somatic DNA alterations in human cancer. *Nat Biotechnol* 2012, 30(5):413-421.

A high-performance database framework for fast and easy prioritization of disease related variants from Next Generation Sequencing data

Bolan Linghu[1,*], Fan Yang[1], Robert Bruccoleri[1], Dan Spiewak [2], Joseph D. Szustakowski[1]

[1]: Translational Medicine, Novartis Institutes for BioMedical Research, Cambridge MA, USA.

[2]: Informatics and Technology, Novartis Institutes for BioMedical Research, Cambridge MA, USA.

[*]: To whom correspondence should be addressed.

Emails: LinghuBL (bolan.linghu@novartis.com); YangFY (fan.yang@novartis.com); BruccoleriRB (robert.bruccoleri@novartis.com); SpiewakDS (dan.spiewak@novartis.com); SzustakowskiJDS (joseph.szustakowski@novartis.com).

Exome sequencing (Exome Seq) has become a promising approach to identify disease related genetic variations. Pinpointing the small subset of pathogenic mutations amongst the thousands or millions of variants generated in an Exome Seq experiment remains a conceptual and computational challenge. One common approach is to use relational database systems to conveniently organize and query variant data for prioritization. However, traditional database systems often perform poorly when applied to such "Big Data". Recently, a number of high-performance database systems have been developed specifically to enable analysis of extremely large data sets. Here we describe applying one such system, namely Vertica, to prioritize disease variants. Our approach leverages Vertica's high performance capabilities to efficiently model, store, and query a comprehensive landscape of information including variant calls, variant quality metrics, predicted functional consequences, allele frequencies, disease prior knowledge, inheritance patterns, and clinical phenotypes. This framework enabled the convenient and efficient identification of candidate disease variants, with significant improvements over traditional databases. To our knowledge, this is the first demonstration that high-performance databases such as Vertica provide an efficient solution to prioritize variants from exome sequencing.

GPU-BLASTN: Accelerating Nucleotide Sequence Alignment by GPUs

Kaiyong Zhao [1], Xiaowen Chu [1, 2, *]

[1]: Department of Computer Science, Hong Kong Baptist University.

[2]: Institute of Computational and Theoretical Studies, Hong Kong Baptist University.

[*]: To whom correspondence should be addressed.

Emails: KZ (kyzhao@comp.hkbu.edu.hk); XC (chxw@comp.hkbu.edu.hk)

The BLAST software package for sequence alignment is one of the most fundamental and widely used bioinformatics tools [1] [2]. Given the large population of BLAST users, any improvement in the execution speed of BLAST will bring significant benefits to the bioinformatics community. Some research groups have used GPUs to accelerate the speed of BLAST. E.g., GPU-BLAST uses GPUs to accelerate BLASTP, and it achieves 3 to 4 times of speedup over single-thread CPU based NCBI-BLASTP [3]. GPUs have also been successfully used to accelerate other sequence alignment tools, e.g., [4].

In this poster, we show our design, implementation, optimization, and experimental results of GPU-BLASTN, a GPU-accelerated version of the widely used NCBI-BLASTN. To the best of our knowledge, this is the first work that provides a complete solution for accelerating BLASTN by GPUs. GPU-BLASTN can obtain identical results as NCBI-BLASTN, and its speed on a contemporary Nvidia GTX680 GPU card is about 10 to 20 times faster than the speed of single-thread NCBI-BLASTN running on Xeon E5620.

We evaluate GPU-BLASTN by running sequence search experiments against human build 36 and mouse build 36 genome databases that have been masked with WindowMasker. We use six sets of query sequences with different lengths ranging from hundreds to hundreds of thousands. We compare the results and running time of GPU-BLASTN with those of NCBI-BLASTN on both single-thread CPU and multi-thread CPU.

The GPU-BLASTN will be open source and freely available to the bioinformatics community.

References

1. Altschul,S.F. et al. (1990) Basic local alignment search tool. J. Mol. Biol., 215, 403–410.

2. Altschul,S.F. et al. (1997) Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. Nucleic Acids Res., 25, 3389–3402.

3. P. D. Vouzis and N. V. Sahinidis (2011) GPU-BLAST: using graphics processors to accelerate protein sequence alignment. Bioinformatics, 27(2), 182-188.

4. C. M. Liu et. al. (2012) SOAP3: ultra-fast GPU-based parallel alignment tool for short reads. Bioinformatics, 28(6), 878-879.

Preprocessing methods to enhance the quality of diversity estimation for pyrosequenced amplicon samples

Byunghan Lee [1] and Sungroh Yoon [1,2,*]

[1]: Electrical and Computer Engineering, Seoul National University, Seoul 151-744, Korea.

[2]: Bioinformatics Institute, Seoul National University, Seoul 151-747, Korea.

[*]: To whom correspondence should be addressed.

Emails: B. L. (styxkr@snu.ac.kr); S. Y. (sryoon@snu.ac.kr)

To analyze pyrosequenced amplicon data in metagenomics, we need to filter or denoise the sequencing data before estimating the diversity appearing in a given sample. Raw sequenced data often contain problematic sequences, which may lead to overestimation [1]. To avoid that, there exist computational tools for preprocessing (*i.e.*, filtering or denoising) pyrosequenced data. Most of these tools utilize nucleotide sequence data, whereas some techniques rely on flow data. In preprocessing, removing erroneous reads and filtering out duplicates can affect the diversity estimation. In addition to preprocessing, the procedure used to identify operational taxonomic units (OTUs) may also bias the estimation, but here we only consider the effect of preprocessing. We compare existing preprocessing approaches [2,3,4] and examine their effectiveness in terms of accuracy and efficiency.

References

1. Victor Kunin, Anna Engelbrektson, Howard Ochman, Philip Hugenholtz. Wrinkles in the rare biosphere: pyrosequencing errors can lead to artificial inflation of diversity estimates. *Environmental Microbiology*, 12: 118-123, 2010.
2. Susan M. Huse, Julie A. Huber, Hilary G. Morrison, Mitchell L. Sogin, David M. Welch. Accuracy and quality of massively parallel DNA pyrosequencing. *Genome Biology*, 8(7): R143, 2007.
3. Susanne Balzer, Ketil Malde, Markus A. Grohme, Inge Jonassen. Filtering duplicate reads from 454 pyrosequencing data. *Bioinformatics*, in press.
4. Weizhong Li, Limin Fu, Beifang Niu, Sitao Wu, John Wooley. Ultrafast Clustering Algorithms for Metagenomic Sequence Analysis. *Briefings in Bioinformatics*, 13(6): 656-668, 2012.

SIGMA: a Bayesian Model Based Clustering Approach for Reconstructing Individual Genomes from Shotgun Sequencing of Microbial Communities

M. Senthil Kumar[1,2,*], Denis Bertrand[1,*], Song Gao[3], Niranjan Nagarajan[1,#]

[1]: Computational and Systems Biology, Genome Institute of Singapore, Singapore 138672.

[2]: Graduate Program in Computational Biology and Bioinformatics, Center for Bioinformatics and Computational Biology, University of Maryland, College Park, MD 20742.

[3]: NUS Graduate School for Integrative Sciences and Engineering, National University of Singapore.

[*]: The two first authors contributed equally to this work

[#]: To whom correspondence should be addressed.

Email: NN (nagarajann@gis.a-star.edu.sg)

For complex microbial communities, the analysis of information-rich, whole-community shotgun sequencing datasets is often limited by the fragmentary nature of the assembly. While recent work has shown that near-complete genomes can be assembled from metagenomic data, a robust and fully-automated framework to consistently do so has yet to be established.

In this work, we show that the problem of metagenomic assembly can be accurately and automatically reduced to the single genome assembly problem by systematically exploiting genome coverage and assembly information. To do this, we introduce a model-based clustering approach called Sigma that finds an optimal solution based on the Bayesian Information Criterion (BIC).

We combined this method with an optimal single-genome scaffolder (Opera) to show that near-complete genomes can be automatically and accurately reconstructed even from shotgun sequencing of complex microbial communities. Comparisons on in silico and real datasets confirmed that our approach (OperaMS) consistently outperforms state-of-the-art single genome (Velvet, SOAPdenovo) and metagenomic (MetaVelvet, Bambus2) assembly tools on assembly contiguity and correctness statistics.

The Elbow Method on deciding significant fold change cutoffs of differentially expressed genes

Xiangli Zhang1, Natalie K Björklund1, Thomas Rydzak2, Richard Sparling2,Graham Alvare1, Brian Fristensky1,*

1: Departments of Plant Science, University of Manitoba, Winnipeg, Canada, R3T 2N2
2: Department of Microbiology, University of Manitoba, Winnipeg, Canada, R3T 2N2
*: To whom correspondence should be addressed.

Emails: XZ (zhangju@cc.umanitoba.ca); NB (Natalie.Bjorklund@ad.umanitoba.ca);
TR (tom8_98@hotmail.com); RS (Richard.Sparling@ad.umanitoba.ca);
GA (Graham.Alvare@ad.umanitoba.ca); BF (Brian.Fristensky@ad.umanitoba.ca)

Fold change test has often been used by biologists to identify differentially expressed genes between two conditions. Fold change test has also been integrated into automatic pipeline with p-value test, for example using Volcano plot with microarray data. However, the cutoff for significance is often arbitrarily set to 2.0 fold change i.e. set as 1 on a log2 scale. Because the chosen cutoffs will significantly affect the choice of genes deemed to have changed significantly, we have devised a method for assigning cutoffs that reflect the trends seen in the actual data set.

The elbow method sets the cut-offs by plotting ordered fold change values against their order creating a logit curve. The cutoffs of significance can be selected with the aid of amplified regional derivative plots with R. Automatic version is being designed and implemented with R. Calculation of tolerance limits (p= 50%, non central t distribution) gives values equivalent to inflections points (elbows) on the curve and can be calculated trivially in R by using the normtol.int function. Significance is based on being either above or below tolerance limits. Therefore, the elbow method produces fold change cutoffs for significance that are adjusted for data variation.

A RNA-seq data of *Clostridium thermocellum* ATCC 27405 was analyzed with elbow method. Cutoffs selected with the elbow method are far away from 2 fold change. With fold 2 cut-off, there are 996 genes of down regulated and 650 up regulated in the total 3238 genes. With elbow cutoffs of up 3.06 and down -2.79 of log2(normalized FPKM fold change), there are total 133 down regulated genes, and 92 up regulated genes from exponential phase to stationary phase. Several gene clusters were identified to be co-regulated. The biggest cluster with 43 genes featured ribosome proteins from Cthe_2891 to Cthe_2933 shows down-regulation. The results are in agreement with the findings by Raman 2011 based on one-way ANOVA analysis with p-value < 0.01.

References

1. Babu Raman, Catherine K McKeown, Miguel Rodriguez, Steven D Brown, Johathan R Mielenz. Transcriptomic analysis of *Clostridium thermocellum* ATCC 27405 cellulose fermentation. BMC Microbiology: 11:134, 2011

Accelerating Mass Spectrometry-Based Protein Identification Using GPUs

You Li [1], Leihao Xia [1], Hao Chi [2,3], Xiaowen Chu [1,4,*]

[1]: Department of Computer Science, Hong Kong Baptist University.

[2]: Key Lab of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China..

[3]: Graduate University of Chinese Academy of Sciences, Beijing 100049, China.

[4]: Institute of Computational and Theoretical Studies, Hong Kong Baptist University.

[*]: To whom correspondence should be addressed.

Emails: YL (youli@comp.hkbu.edu.hk); LX (10050760@life.hkbu.edu.hk); HC (hchi@jdl.ac.cn); XC (chxw@comp.hkbu.edu.hk)

Tandem mass spectrometry-based database searching is currently the principal method for protein identification in shotgun proteomics. The explosive growth of protein and peptide databases due to genome translations, enzymatic digestions, and post-translational modifications (PTMs), is making computational efficiency in database searching a serious challenge. Profile analysis shows that most search engines spend 50%-90% of their total time on the scoring module, and that the spectrum dot product (SDP) based scoring module is the most widely used. As a general purpose and high performance parallel hardware, graphics processing units (GPUs) are promising platforms for speeding up many bioinformatics tools [1] [2].

In this poster, we show our design and implementation of a parallel SDP-based scoring module on GPUs that exploits the efficient use of GPU registers and shared memory. Compared with the CPU-based version, we achieved a 30 to 60 times speedup using a single GPU. We also implemented our algorithm on a GPU cluster and achieved an approximately linear speedup.

Our GPU-based SDP algorithm can significantly improve the speed of the scoring module in mass spectrometry-based protein identification. The algorithm can be easily implemented in many database search engines such as X!Tandem, SEQUEST, and pFind.

More details about this work can be found in [3]. A software tool implementing this algorithm is freely available at [4].

Acknowledgement

This work was supported by research grants FRG2/10-11/099 and FRG2/11-12/158 from Hong Kong Baptist University.

References

1. C. M. Liu et. al. SOAP3: ultra-fast GPU-based parallel alignment tool for short reads. Bioinformatics, 28(6), 878-879, 2012.
2. K. Zhao and X.-W. Chu. GPU-BLASTN: Accelerating Nucleotide Sequence Alignment by GPUs. Poster at RECOMB 2013.
3. Y. Li and X.-W. Chu. Speeding up Scoring Module of Mass Spectrometry Based Protein Identification by GPU, The Fifth International Symposium on Advances of High Performance Computing and Networking, Liverpool, UK, June 2012.
4. http://www.comp.hkbu.edu.hk/~chxw/ProteinByGPU.html

De novo transcript reconstruction and abundance estimation in eukaryotic RNA-Seq data analysis

Tianyang Li [1], Rui Jiang [1], Xuegong Zhang [1,*]

[1]: MOE Key Laboratory of Bioinformatics and Bioinformatics Division, TNLIST/Department of Automation, Tsinghua University.

[*]: To whom correspondence should be addressed.

Email: Xuegong Zhang (zhangxg@tsinghua.edu.cn)

RNA-Seq is a recent technology to identify and quantify transcriptomes through high-throughput sequencing. Transcript reconstruction and abundance estimation are important steps in RNA-Seq data analysis, and current approaches are limited. For example, current approaches for transcript abundance estimation require aligning RNA-Seq reads to a reference genome. And some approaches may be computationally inefficient for genes with a large number of exons, as they enumerate many possible isoforms for these genes. Some approaches may also fail to discover certain isoforms whose start or end sites are contained within other isoforms. Here, we propose a Bayesian statistical method for *de novo* transcript reconstruction and abundance estimation in eukaryotic RNA-Seq data analysis. Our method uses a Markov chain Monte Carlo method to sample the posterior probability. An advantage of our method is that it can discover isoforms whose start or end sites are contained within other isoforms. And instead of enumerating a gene's set of possible isoforms, we initially start with a small set of possible isoforms and make modifications to this set in Markov chain Monte Carlo iterations. A tool implementing our method is currently in development, and is available at https://github.com/tianyang-li/de-novo-rna-seq-quant-1.

A Random-Permutations-Based Approach to Fast Read Processing

Roy Lederman [1]

[1]: Applied Mathematics Program, Yale University.

Email: roy.lederman@yale.edu

Read alignment and assembly are computationally expensive steps in the processing of NGS reads. Existing read alignment programs use prefix-tree algorithms and hash-table algorithms.

We present a new approach to read alignment which uses random permutations of strings. This randomized approach is flexible, accurate and fast. We present experimental results to demonstrate that random-permutations-based algorithms can successfully align significantly more reads than comparable programs in significantly shorter run times.

We demonstrate the flexibility of permutations-based algorithms by extending them to other applications, such as assembly.

We also describe "homopolymer-length-filters," a separate method of read processing which allows random-permutations-based algorithms to also process 454/IonTorrent reads rapidly and accurately.

Our paper "A Random-Permutations-Based Approach to Fast Read Alignment" will be presented at RECOMB-seq.
Technical reports and more information: http://alignment.commons.yale.edu.

FaSD: a efficient model to detect SNPs for NGS data

Feng Xu[1,2,†], Weixin Wang[1,2,†], Panwen Wang[1,2], Mulin Jun Li[1,2], Pak Chung Sham[3,4,5], and Junwen Wang[1,2,3,6*]

[1]Department of Biochemistry, LKS Faculty of Medicine, The University of Hong Kong, Hong Kong SAR, China.

[2]Shenzhen Institute of Research and Innovation, The University of Hong Kong, Shenzhen, China.

[3]Centre for Genomic Sciences, LKS Faculty of Medicine, The University of Hong Kong, Hong Kong SAR, China.

[4]Department of Psychiatry, LKS Faculty of Medicine, The University of Hong Kong, Hong Kong SAR, China.

[5]State Key Laboratory in Cognitive and Brain Sciences, The University of Hong Kong, Hong Kong SAR, China.

[6]HKU-BGI Bioinformatics Algorithms and Core Technology Research Laboratory, The University of Hong Kong, Hong Kong SAR, China.

*Correspondence should be addressed to Junwen Wang (Tel: +852 2831 5075; Fax: +852 28551254; Email: junwen@hku.hk)

[†]Both authors contributed equally to this work

Various methods have been developed for calling single nucleotide polymorphisms (SNPs) from next-generation sequencing (NGS) data. However, for satisfactory performance, most of these methods require expensive high-depth sequencing. Here, we propose a fast and accurate SNP detection (FaSD) program that uses a binomial distribution based algorithm and a mutation probability. We extensively assess this program on normal and cancer NGS data from The Cancer Genome Atlas project and pooled data from the 1000 Genomes Project. We also compare the performance of several state-of-the-art programs for SNP calling and evaluate their pros and cons. We demonstrate that FaSD is a fast and highly accurate SNP detection method, particularly when the sequence depth is low. FaSD can finish SNP calling within four hours for ten-fold human genome NGS data (30 gigabases) on a standard desktop computer.

References

Xu F, Wang W, Wang P, Li M J, Chung Sham P, Wang J. A fast and accurate SNP detection algorithm for next-generation sequencing data. *Nat Commun* **3**:1258, 2012.

Detecting DNA modications from 3rd generation sequencing data by modeling sequence context dependence of polymerase kinetic

Zhixing Feng [1], Xuegong Zhang [1,2,*]

[1]: MOE Key Laboratory of Bioinformatics, Bioinformatics Division and Center for Synthetic and Systems Biology, TNLIST / Department of Automation, Tsinghua University, Beijing 100084, China

[2]: School of Life Sciences, Tsinghua University, Beijing 100084, China

[*]: To whom correspondence should be addressed.

Emails: ZF (zxfeng.thu@gmail.com); XZ (zhangxg@tsinghua.edu.cn)

DNA modifications such as methylation and DNA damage can play critical regulatory roles in biological systems. Single molecule, real time (SMRT) sequencing technology generates DNA sequences as well as DNA polymerase kinetic information that can be used for the direct detection of DNA modifications. We demonstrate that local sequence context has a strong impact on DNA polymerase kinetics in the neighborhood of the incorporation site during the DNA synthesis reaction, allowing for the possibility of estimating the expected kinetic rate of the enzyme at the incorporation site using kinetic rate information collected from existing SMRT sequencing data (historical data) covering the same local sequence contexts of interest.

We develop an Empirical Bayesian hierarchical model for incorporating historical data. Our results show that the model could greatly increase DNA modifications detection accuracy, and reduce requirement of control data coverage [1]. For some DNA modifications that have a strong signal, a control sample is even not needed by using historical data as alternative to control. Thus, sequencing cost can be greatly reduced by using the model. We implemented the model in an R package named seqPatch, which is available at https://github.com/zhixingfeng/seqPatch.

References

1. Zhixing Feng *et al*. Detecting DNA modifications from SMRT sequencing data by modeling sequence context dependence of polymerase kinetic. Plos Comput Bio, In press

An Experimental Evaluation of the Performance of RNA-Seq Mapping Tools

Jingjing Hao [1], Rui Jiang [1,*], Tao Jiang [2,1,*]

[1]: MOE Key Laboratory of Bioinformatics and Bioinformatics Division, TNLIST/Department of Automation, Tsinghua University, Beijing 100084, China

[2]: Department of Computer Science and Engineering, University of California, Riverside, CA 92521, USA

[*]: To whom correspondence should be addressed.

Emails: Rui Jiang (ruijiang@tsinghua.edu.cn); Tao Jiang (jiang@cs.ucr.edu)

Abstract: Transcriptome sequencing (RNA-Seq) has becoming a key technology in the field of transcriptomics for quantifying gene expression, detecting novel transcripts, analysing RNA functions, etc. For organisms with reference genomes, mapping RNA-Seq reads to the genomic sequences is typically the first step to process RNA-Seq data. In the last few years, many algorithms and tools for mapping RNA-Seq reads have been developed. Since the objectives and constraints of these methods are usually different, their performance varies. How to choose the most appropriate mapping tools to analyze a specific RNA-Seq dataset so some particular performance expectations can be met is an important question that bioinformaticians need to address. Here, we provide a systematic experimental evaluation of some state-of-the-art RNA-Seq mapping tools by studying their accuracy in read alignment and junction detection on simulated RNA-Seq data with different sequencing depths, read lengths and rates of substitutions, indels and sequencing error. In addition, the time efficiency and memory usages of the tools are also investigated. We are also interested in the impact of paired-end reads on the performance of the tools and the question of to what degree the tools are able to take advantage of parallel computation. Some real data tests will be used to confirm the simulation results. We hope that our study will provide information useful for choosing suitable mapping tools in RNA-Seq data analysis.

Transposon-derived and satellite-derived repetitive sequences play distinct functional roles in mammalian intron size expansion

Dapeng Wang[*,1], Yao Su[*,1,2], Xumin Wang[1], Hongxing Lei[§,1], and Jun Yu[§,1]

[1]: CAS Key Laboratory of Genome Sciences and Information, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 100029, P. R. China

[2]: Graduate University of Chinese Academy of Sciences, Beijing 100049, P. R. China

[*]: These authors contributed equally to this work.

[§]: To whom correspondence should be addressed.

Emails: DPW (wangdp@big.ac.cn); JY (junyu@big.ac.cn)

Repetitive sequences (RSs) are redundant, complex at times, and often lineage-specific, representing significant "building" materials for genes and genomes. According to their origins, sequence characteristics, and ways of propagation, repetitive sequences are divided into transposable elements (TEs) and satellite sequences (SSs) as well as related subfamilies and subgroups hierarchically. The combined changes attributable to the repetitive sequences alter gene and genome architectures, such as the expansion of exonic, intronic, and intergenic sequences, and most of them propagate in a seemingly random fashion and contribute very significantly to the entire mutation spectrum of mammalian genomes. Our analysis is focused on evolutional features of TEs and SSs in the intronic sequence of twelve selected mammalian genomes. We divided them into four groups—primates, large mammals, rodents, and primary mammals—and used four non-mammalian vertebrate species as the out-group. After classifying intron size variation in an intron-centric way based on RS-dominance (TE-dominant or SS-dominant intron expansions), we observed several distinct profiles in intron length and positioning in different vertebrate lineages, such as retrotransposon-dominance in mammals and DNA transposon-dominance in the lower vertebrates, amphibians and fishes. The RS patterns of mouse and rat genes are most striking, which are not only distinct from those of other mammals but also different from that of the third rodent species analyzed in this study—guinea pig. Looking into the biological functions of relevant genes, we observed a two-dimensional divergence; in particular, genes that possess SS-dominant and/or RS-free introns are enriched in tissue-specific development and transcription regulation in all mammalian lineages. In addition, we found that the tendency of transposons in increasing intron size is much stronger than that of satellites, and the combined effect of both RSs is greater than either one of them alone in a simple arithmetic sum among the mammals and the opposite is found among the four non-mammalian vertebrates. TE- and SS-derived RSs represent major mutational forces shaping the size and composition of vertebrate genes and genomes, and through natural selection they either fine-tune or facilitate changes in size expansion, position variation, and duplication, and thus in functions and evolutionary paths for better survival and fitness. When analyzed globally, not only are such changes significantly diversified but also comprehensible in lineages and biological implications.

Reference

Dapeng Wang, Yao Su, Xumin Wang, Hongxing Lei and Jun Yu. **Transposon-Derived and Satellite-Derived Repetitive Sequences Play Distinct Functional Roles in Mammalian Intron Size Expansion**. *Evolutionary Bioinformatics* 2012:8 301-319.

LCGbase: A Comprehensive Database for Lineage-based Co-regulated Genes

Dapeng Wang[*,1,3], Yubin Zhang[*,1,2,3], Zhonghua Fan[*,1], Guiming Liu[1] and Jun Yu[§,1,2]

[1]: CAS Key Laboratory of Genome Sciences and Information, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 100029, PR China

[2]: Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, PR China

[3]: Graduate University of Chinese Academy of Sciences, Beijing 100049, PR China

*: These authors contributed equally to this work.

[§]: To whom correspondence should be addressed.

Emails: **DPW** (wangdp@big.ac.cn); **JY** (junyu@big.ac.cn)

Animal genes of different lineages, such as vertebrates and arthropods, are well-organized and blended into dynamic chromosomal structures that represent a primary regulatory mechanism for body development and cellular differentiation. The majority of genes in a genome are actually clustered, which are evolutionarily stable to different extents and biologically meaningful when evaluated among genomes within and across lineages. Here, we provide a user-friendly database—LCGbase (a comprehensive database for lineage-based co-regulated genes)—hosting information on evolutionary dynamics of gene clustering and ordering within animal kingdoms in two different lineages: vertebrates and arthropods. Compared to other gene annotation databases with similar purposes, our database has three comprehensible advantages. First, our database is inclusive, including all high-quality genome assemblies of vertebrates and representative arthropod species. Second, it is human-centric since we map all gene clusters from other genomes in an order of lineage-ranks (such as primates, mammals, warm-blooded, and reptiles) onto human genome and start the database from well-defined gene pairs (a minimal cluster where the two adjacent genes are oriented as co-directional, convergent, and divergent pairs) to large gene clusters. Furthermore, users can search for any adjacent genes and their detailed annotations. Third, the database provides flexible parameter definitions, such as the distance of transcription start sites between two adjacent genes, which is extendable to genes that flanking the cluster across species. We also provide useful tools for sequence alignment, gene ontology (GO) annotation, promoter identification, gene expression (co-expression), and evolutionary analysis. This database not only provides a way to define lineage-specific and species-specific gene clusters but also facilitates future studies on gene co-regulation, epigenetic control of gene expression (DNA methylation and histone marks), and chromosomal structures in a context of gene clusters and species evolution. LCGbase is freely available at http://lcgbase.big.ac.cn/LCGbase.

Reference

Dapeng Wang, Yubin Zhang, Zhonghua Fan, Guiming Liu and Jun Yu. **LCGbase: A Comprehensive Database for Lineage-Based Co-regulated Genes**. *Evolutionary Bioinformatics* 2012:8 39-46.

Both size and GC-content of minimal introns are selected in human populations

Dapeng Wang[1,2], and Jun Yu[§,1]

[1]: CAS Key Laboratory of Genome Sciences and Information, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 100029, PR China

[2]: Graduate University of Chinese Academy of Sciences, Beijing 100049, PR China

[§]: To whom correspondence should be addressed.

Emails: **DPW** (wangdp@big.ac.cn); **JY** (junyu@big.ac.cn)

We previously have studied the insertion and deletion polymorphism by sequencing no more than one hundred introns in a mixed human population and found that the minimal introns tended to maintain length at an optimal size. Here we analyzed re-sequenced 179 individual genomes (from African, European, and Asian populations) from the data released by the 1000 Genome Project to study the size dynamics of minimal introns. We not only confirmed that minimal introns in human populations are selected but also found two major effects in minimal intron evolution: (i) Size-effect: minimal introns longer than an optimal size (87 nt) tend to have a higher ratio of deletion to insertion than those that are shorter than the optimal size; (ii) GC-effect: minimal introns with lower GC content tend to be more frequently deleted than those with higher GC content. The GC-effect results in a higher GC content in minimal introns than their flanking exons as opposed to larger introns (>=125 nt) that always have a lower GC content than that of their flanking exons. We also observed that the two effects are distinguishable but not completely separable within and between populations. We validated the unique mutation dynamics of minimal introns in keeping their near-optimal size and GC content, and our observations suggest potentially important functions of human minimal introns in transcript processing and gene regulation.

Reference

Dapeng Wang, and Jun Yu. **Both size and GC-content of minimal introns are selected in human populations**. *PLoS ONE* 2011, 6(3): e17945.

GRiG: A PPV-sensitive method for predicting somatic SNVs from cancer-normal paired sequencing data with greedy rule induction algorithm

Shaoping Ling[1,$], Lili Dong [1,$], Lihua Cao[1], Caiyan Jia[2,3], Xuemei Lu[1*], Chung-I Wu[1,4*]

[1]: Beijing Institute of Genomics, Chinese Academy of Scineces, Beijing, China.

[2]: School of Computer and Information Technology, Beijing Jiaotong University, Beijing 100044, China.

[3]: Department of Bioengineering/Bioinformatics, University of Illinois at Chicago, Chicago, IL 60612, USA.

[4]: Department of Ecology and Evolution, University of Chicago, Chicago, Illinois 60637, USA.

[$]: These authors contributed equally to this work.

[*]: To whom correspondence should be addressed.

Emails: SL (spling@big.ac.cn); LD (donglili@big.ac.cn); LC (caolh@big.ac.cn); CJ (caiyan.jia@gmail.com); XL (luxm@big.ac.cn); CW (wuci@big.ac.cn)

Predicting somatic SNVs from cancer-normal paired sequencing is a key computational issue in high-throughput sequencing-driven cancer genomics. Classic methods based on statistical inference (SI) have been developed and become standard pipeline in human variation detection. However, they can not provide enough high positive prediction value (PPV) for further experimental validation and function analysis. We presented a Greedy Rule Induction alGorithm (GRiG) for predicting somatic SNVs in cancer-normal paired sequencing data, which integrates feature selection and rule inference into a machine learning frame work. We evaluated the performance of GRiG on public datasets which consist of two candidate somatic SNVs datasets from 48 breast exome capture sequencing (ECS) datasets and 4 whole genome sequencing (WGS) datasets for training and testing respectively. GRiG always achieved the better performance in ECS training dataset with 10x cross-validation and WGS testing dataset than both Samtools and GATK and presented comparable performance with four statistical learning algorithms including random forest, Bayesian additive regression tree, support vector machine and logistic regression in ECS training dataset with 10x cross-validation and WGS testing dataset. Especially, it always achieved better PPV than these four classifiers.

References

1. Ding,J. et al. (2012) Feature-based classifiers for somatic mutation detection in tumour-normal paired sequencing data. *Bioinformatics*, 28, 167–175.
2. McKenna,A. et al. (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.*, 20, 1297–1303.
3. Li,H. et al. (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25, 2078–2079.
4. Greenman,C. et al. (2007) Patterns of somatic mutation in human cancer genomes. *Nature.*

psRobot: a web-based plant small RNA meta-analysis toolbox

Hua-Jun Wu[1,2†], Ying-Ke Ma[1,2†], Tong Chen[1,2], Meng Wang[1] and Xiu-Jie Wang[1*]

[1]: The State Key Laboratory of Plant Genomics, Institute of Genetics and Developmental Biology,Chinese Academy of Sciences, Beijing 100101, China

[2]: Graduate University of Chinese Academy of Sciences, Beijing 100101, China

[†]: These authors contributed equally to this work

[*]: To whom correspondence should be addressed. E-mail: xjwang@genetics.ac.cn

Emails: Hua-Jun Wu (hjwu@genetics.ac.cn); Ying-Ke Ma (mayingke@gmail.com); Tong Chen (tchen@genetics.ac.cn) and Meng Wang (mengwang@genetics.ac.cn)

Small RNAs (smRNAs) in plants, mainly microRNAs and small interfering RNAs, play important roles in both transcriptional and post-transcriptional gene regulation. The broad application of high-throughput sequencing technology has made routinely generation of bulk smRNA sequences in laboratories possible, thus has significantly increased the need for batch analysis tools. PsRobot is a web-based easy-to-use tool dedicated to the identification of smRNAs with stem-loop shaped precursors (such as microRNAs and short hairpin RNAs) and their target genes/transcripts. It performs fast analysis to identify smRNAs with stem-loop shaped precursors among batch input data and predicts their targets using a modified Smith–Waterman algorithm. PsRobot integrates the expression data of smRNAs in major plant smRNA biogenesis gene mutants and smRNA-associated protein complexes to give clues to the smRNA generation and functional processes. Besides improved specificity, the reliability of smRNA target prediction results can also be evaluated by mRNA cleavage (degradome) data. The cross species conservation statuses and the multiplicity of smRNA target sites are also provided. PsRobot is freely accessible at http://omicslab.genetics.ac.cn/psRobot/.

References

1. Hua-Jun Wu, Ying-Ke Ma, Tong Chen, Meng Wang, Xiu-Jie Wang. PsRobot: a web-based plant small RNA meta-analysis toolbox. Nucleic Acids Res. 40 (W1): W22-W28, 2012.

Predicting functional DNA elements from histone modifications

Joanna Giemza [1], Bartek Wilczyński [1,*]

[1]: Faculty of Mathematics, Informatics and Mechanics, University of Warsaw, Warsaw, Poland .

[*]: To whom correspondence should be addressed.

Emails: JG (j.giemza@students.mimuw.edu.pl); BW (bartek@mimuw.edu.pl)

Histones, especially their N-terminal tails, are subject to a large number of post-translational modifications. In the last few years, thanks to high-throughput experimental methods, particularly CHIP-chip and CHIP-Seq, remarkable progress has been made in the characterization of histone modifications. The link between histone modifications and transcription has been particularly intensively studied. It has been found ([1], [2]) that some individual modifications can be associated with transcriptional activation or repression. For instance, H3K4me3 is enriched in promoters and H3K36me3 in transcribed regions of active genes. Furthermore, it has been shown, based on genome-wide CHIP-Seq data for 38 histone modifications and one histone variant from human CD4+T-cells, that histone modification levels are predictive of gene expression levels [3]. We analyze the same dataset, but addressing different problem: prediction of functional DNA elements, such as promoters, exons and introns from histone modifications. We compare classifiers based on Bayesian networks and random forests. In the first case, we consider a bipartite Bayesian Network between classification attributes (histone modifications) and predicted classes (binary indicators of functional elements). BNFinder software ([4]) provides the optimal topology of the network, performing feature selection simultaneously. Additionally, we analyze impact of preprocessing on classification quality. Our results indicate that it is possible to accurately predict major functional annotations in their active state, while inactive elements seem to be difficult to distinguish from intergenic regions. While random forest classifiers provide overall better accuracy, the Bayesian models are more useful in selecting the few most informative modifications.

Acknowledgements:

This work was supported by the Foundation for Polish Science within Homing Plus programme co-financed by the European Union - European Regional Development Fund.

References

1. Artem Barski et al. High-resolution profiling of histone methylations in the human genome. *Cell*, 129(4):823–837, 2007.

2. Zhibin Wang et al. Combinatorial patterns of histone acetylations and methylations in the human genome. *Nature Genetics*, 40(7):897–903, 2008.

3. Rosa Karlić et al. Histone modification levels are predictive for gene expression. *PNAS*, 107 (7) 2926-2931, 2010.

4. Bartek Wilczyński, Norbert Dojer. Bnfinder: exact and efficient method for learning Bayesian networks. *Bioinformatics*, 25(2):286–287, 2009.

Supervised learning of enhancer activity from chromatin modifications and sequence motifs

Agnieszka Podsiadło [1] and Bartek Wilczynski [1,*]

[1]: Institute of Informatics, University of Warsaw, Warsaw, Poland.

[*]: To whom correspondence should be addressed.

Emails: a.podsiadlo@students.mimuw.edu.pl; bartek@mimuw.edu.pl

Enhancers are key functional DNA elements leading to diverse gene activity patterns in higher eukaryotes. With recent expansion of high-throughput experimental approaches for enhancer identification, we are quickly gaining knowledge of enhancer location in many model organisms. However, predicting which of the mapped enhancers are going to be active in a given cellular context is often a more complex problem. Chromatin modifications have been shown to be an indicator of gene activity, and recently we have also learned that they are predictive of enhancer activity [1]. Although such methods already provide significant results, the prediction quality could potentially be improved by adding information about the enhancer sequence features.

Our study aims to accurately predict enhancer activity based on both chromatin modifications and sequence motifs for multiple transcription factors. In order to be able to better select important features, we have applied similar analysis of chromatin modifications to an enlarged dataset, yielding models providing more accurate results. We describe how the sequence features are understood as computed TRAP score [2] for the sample sequence and given motifs from the JASPAR database [3], as well as motifs related to the transcriptor factors involved in development of the studied tissue. In addition to more accurate predictions, our idea leads to identification of new patterns among sequence motifs that bring essential information tissue specific activity. In terms of classification methods, we discuss results obtained from different classiffiers, including Bayesian Networks [4] and Random Forest [5].

Acknowledgements: This work was supported by the Foundation for Polish Science within Homing Plus programme co-financed by the European Union - European Regional Development Fund.

References

1. S. Bonn, R. Zinzen, C. Girardot, E. Gustafson, et al., “Tissue-specific analysis of chromatin state identifies temporal signatures of enhancer activity during embryonic development,” Nature genetics, vol. 44, no. 2, pp. 148–156, 2012.
2. H. Roider, A. Kanhere, T. Manke, and M. Vingron, “Predicting transcription factor affinities to dna from a biophysical model,” Bioinformatics, vol. 23, no. 2, pp. 134–141, 2007.
3. A. Sandelin,W. Alkema, P. Engström,W.Wasserman, and B. Lenhard, “Jaspar: an open-access database for eukaryotic transcription factor binding profiles,” Nucleic acids research, vol. 32, no. suppl 1, pp. D91–D94, 2004.
4. B. Wilczy´nski and N. Dojer, “Bnfinder: exact and efficient method for learning bayesian networks,” Bioinformatics, vol. 25, no. 2, p. 286, 2009.
5. L. Breiman, “Random forests,” Machine learning, vol. 45, no. 1, pp. 5–32, 2001.

Modeling chromatin domain boundaries from histone modification data

Paweł Bednarz [1,*], Bartek Wilczyński [1,*]

[1]: Institute of Informatics, Warsaw University.

[*]: To whom correspondence should be addressed.

Emails: Pawel.Bednarz@mimuw.edu.pl, bartek@mimuw.edu.pl

In the process of cell differentiation cells with the same DNA code undergo processes which transform them into multitude of different cell types. We know that the information about cell fate is encoded in the chromatin by the set of factors known as an epigenetic state. Unraveling complex dependencies between the epigenetic state of chromatin and specific transcriptional regulation patterns leading to the terminally differentiated cell seems to be one of the most challenging tasks in the modern molecular biology. Recently, new techniques for investigating the epigenetic state have been developed and their results have been made publicly available. In our work we aim to model the dynamics of interacting loci based on various epigenetic factors from ChIP-on-Chip and ChIP-seq experiments for training, as well as insulator proteins' binding profiles. In this poster we will present the classifier based on Bayesian networks adopted to the task of finding boundaries of regulatory domains in *Drosophila* embryos. The classifier utilizes recent data on chromatin interactions from chromosome conformation capture experiments. It's performance will be compared to a number of unsupervised methods. In addition, we try to figure out key factors correlated with boundary appearance and explain some phenomena observed previously in *Drosophila*.

References

1. B. Wilczynski, N. Dojer. Bnfinder: Exact and efficient method for learning bayesian networks. *Bioinformatics*, 25(2):286–287, January 2009.
2. T. Sexton, E. Yaffe, E. Kenigsberg, F. Bantignies, et al. Three-dimensional folding and functional organization principles of the *Drosophila* genome. *Cell*, 148:458–472, January 2012.
3. G. Filion, J. Van Bemmel, U. Braunschweig, W. Talhout, et al. Systematic protein location mapping reveals five principal chromatin types in *Drosophila* cells. *Cell*, 143(2):212–224, 2010.
4. S. Roy, J. Ernst, P. V. Kharchenko, P. Kheradpour, et al. Identification of functional elements and regulatory circuits by *Drosophila* modENCODE. *Science*, 330(6012):1787–1797, December 2010.
5. J. Ernst and M. Kellis. Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nature Biotechnology*, 28:817–825, July 2010.

Genome-wide analysis of human hotspot intersected genes highlights the roles of meiotic recombination in evolution and disease

Tao Zhou, Zhibin Hu, Zuomin Zhou, Xuejiang Guo* and Jiahao Sha*

State Key Laboratory of Reproductive Medicine, Nanjing Medical University, 140 Hanzhong Road, Nanjing, Jiangsu Province, 210029, P.R. China

*: To whom correspondence should be addressed.

Emails: Tao Zhou (zhoutao@njmu.edu.cn); Xuejiang Guo (guo_xuejiang@njmu.edu.cn); Jiahao Sha (shajh@njmu.edu.cn)

Human meiotic homologous recombination plays important roles in generating genetic diversity and repairing of double strands breaks. Recently, the hotspot regions of human meiosis recombination have been mapped in fine-scale. To systemically analyze the consequences of meiotic recombination and its relationship with human diseases, we first defined meiotic recombination hotspot intersected protein-coding genes as HI genes. Though comparative analysis of HI genes using various pre-defined datasets associated with evolution and disease, we provided interesting and robust results about the double sides of meiotic recombination. First, HI genes as overlapped with meiotic recombination hotspots are evolving. They are prone to locate in membrane and extracellular regions and play role in cell to cell communication. HI genes are important for the evolution of multicellular organisms. As brain and blood specific genes are enriched in HI genes, it indicates that they may be involved in the evolution of human intelligence and the immune system. However, Mendelian heritable disease and cancer associated genes are also enriched in HI genes. We find that HI genes are mostly correlated with chromosomal rearrangement associated disease genes. It indicates the disease susceptibility of hotspot regions and by-product of meiotic recombination. We further listed repeat elements that enriched both in hotspots and chromosomal rearrangement associated disease genes. Our study will enable us to better understand the evolutionary and biological significance of human meiotic recombination.

Phylogenetic Analysis Reveals the Evolution and Diversification of Cyclins in Eukaryotes

Zhaowu Ma, Yuliang Wu, Jun Yan, Hongmei Zhang, Shuzhen Kuang, Mi Zhou, An-Yuan Guo *

Hubei Bioinformatics & Molecular Imaging Key Laboratory, Department of Biomedical Engineering, College of Life Science and Technology, Huazhong University of Science and Technology, Wuhan, Hubei, 430074, PR China.

*: To whom correspondence should be addressed.

Emails: Z. Ma (zhaowuma@yahoo.com.cn); Y. Wu (wylhustcn@foxmail.com); J. Yan (xinsinian2006@163.com); H. Zhang (zhm.2009.happy@163.com); S. Kuang (kuang.hust@163.com); M. Zhou (zhouzhongmi@yahoo.cn); A.-Y. Guo (guoay@mail.hust.edu.cn)

Cyclins are a family of diverse proteins that play fundamental roles in regulating cell cycle progression in Eukaryotes. Cyclins have been identified from protists to higher Eukaryotes, while its evolution remains vague and the findings turn out controversial. Current classification of cyclins is mainly based on their functions, which may not be appropriate for the systematic evolutionary analysis. In this work, we performed comparative and phylogenetic analysis of cyclins to investigate their classification, origin and evolution. Cyclins originated in early Eukaryotes and evolved from protists to plants, fungi and animals. Based on the phylogenetic tree, cyclins can be divided into three major groups designated as the group I, II and III with different functions and features. Group I plays key roles in cell cycle, group II varied in actions are kingdom (plant, fungi and animal) specific, and group III functions in transcription regulation. Our results showed that the dominating cyclins (group I) diverged from protists to plants, fungi and animals, while divergence of the other cyclins (groups II and III) has occurred in protists. We also discussed the evolutionary relationships between cyclins and cyclin-dependent kinases (CDKs) and found that the cyclins have undergone divergence in protists before the divergence of animal CDKs. This reclassification and evolutionary analysis of cyclins might facilitate understanding eukaryotic cell cycle control.

A novel functional beta model for detecting age-related genomewide DNA methylation marks

Qi Shen[1], Chenyang Wang[1], Jinfeng Xu[2] and Hong Zhang[1,*]

[1]: Department of Biostatistics and Computational Biology, School of Life Science, Fudan University, 220 Handan Road, Shanghai 200433, P.R. China.

[2]: Department of Statistics and Applied Probability, National University of Songapore, 6 Science Drive 2, Singapore 117546.

[*]: To whom correspondence should be addressed.

Email: zhanghfd@fudan.edu.cn

DNA methylation (mDNA) has been shown to play an important role in many complex diseases. The rapid development of genome-wide mDNA scan provides great opportunities for genomewide mDNA-disease association studies. The DNA methylation is a dynamic process involving time, and it is quite evident that age contributes to its variation to a large extend. Therefore, in analyzing the genomewide DNA methylation data, it is most relevant to identify the age-related mDNA marks. This helps better understand the underlying biological mechanism and facilitate the early diagnosis and prognosis analysis of complex diseases. We develop a novel functional beta model for analyzing mDNA data and detecting age-related mDNA marks on the whole genome by naturally taken into account sampling scheme and accommodating flexible data dynamics.

We focus on mDNA data obtained through the widely used bisulfite conversion technique which measures the level as a beta value between 0 and 1. A novel beta model is proposed to relate the mDNA level to the age. Adjusting for certain confounders, the functional age effect is left completely unspecified, yielding great flexibility and allowing extra data dynamics. An efficient iterative algorithm is developed for estimating the unknown parameters and function while a test procedure is used to detect age-related mDNA marks. Illustrate with a simulation study and three real data applications, the proposed method exhibits superior performance to existing methods.

Database of human disease and trait related synonymous mutations

Wanjun Gu[1,*], Yihua Zhu[2,3], Xiaofei Wang[1], Chaoqun Zhang[3], Li Liu[2] and Jianming Xie[2]

[1]: Research Center of Learning Sciences, Southeast University, Nanjing, 210096, China

[2]: Department of Biomedical Engineering, Southeast University, Nanjing, 210096, China

[3]: College of Information Technology and Science, Nanjing Agricultural University, Nanjing, 210095, China

*To whom correspondence should be addressed.

Emails: WG (Wanjun.Gu@gmail.com), YZ (zhyh@njau.edu.cn), XW (fei0625@163.com), CZ (zcq0903@gmail.com), LL (liliu1101@163.com), JX (xiejm@seu.edu.cn)

Synonymous mutations are mutations that do not change the encoded amino acids. It is generally assumed synonymous mutations are evolutionary neutral and they have no effects on phenotype. However, recent studies have revealed synonymous mutations are not totally silent in many cases, and some synonymous mutations are related to human diseases. Here we introduced a database of human disease and trait related synonymous mutations. We complied the dataset by integrating SNPs from several sources, including disease related synonymous SNPs identified in genome wide associated studies (GWAS), clinical related synonymous SNPs deposited in Clinical Variation Database (ClinVar), some additional synonymous SNPs in publications, etc. We collected 4,962 synonymous SNPs that are potentially related to human disease or trait in the database. For each SNP, we computed its effects on gene translation efficiency, local RNA secondary structure, RNA splicing, codon usage bias, etc. These measures may have implications in understanding the mechanisms that synonymous mutations cause human diseases. The development of this database will help us annotate functional synonymous SNPs in personal genome sequencing and medical genome sequencing.

New visualization of flow cytometry data to facilitate gating analysis

Peng Qiu

Department of Bioinformatics and Computational Biology, The University of Texas M.D. Anderson Cancer Center.

The flow cytometry technology is able to interrogate protein expressions at the single-cell level. A typical flow cytometry experiment on one biological sample provides measurements of several protein markers on or inside hundreds of thousands of individual cells in the sample. Such data contains information about the cellular heterogeneity underlying the sample, which is of great biological interests. Analysis of such data often aims to identify subpopulations of cells with distinct phenotypes, which is essentially a clustering problem. Despite increasing interests and technological advances, approaches for analyzing such data remain inadequate. Currently, the most widely-used approach is manual gating, a subjective and labor-intensive process on a user-defined sequence of nested biaxial plots. Efforts have been made to automate gating by clustering algorithms. That approach is challenging, because cell counts of different subpopulations are typically quite unbalanced. Here, we propose an alternative strategy. Our goal is to provide novel visualizations to improve the manual analysis, rather than automating it. The basic idea is to view a flow cytometry dataset as a high-dimensional point cloud of cells, extract its skeleton, and unfold the skeleton to obtain a 2D visualization, just like unfolding an origami back to a piece of paper. After that, manual gating can be conveniently performed on one visualization, rather than a sequence of nested 2D plots.

Conditional mutual information for identification of gene-specific methylation threshold

Yihua Liu, Peng Qiu*

Department of Bioinformatics and Computational Biology, The University of Texas M.D. Anderson Cancer Center

*: To whom correspondence should be addressed.

DNA methylation plays an important role in many biological processes by regulating gene expression. Methylation is often described as a binary on-off signal. However, it is typically measured by a beta value derived from either microarray or sequencing technologies, which can take continuous values in [0, 1]. If we believe methylation is binary, a threshold for the beta values should be determined to interpret methylation data. The question we want to ask here is whether or not the appropriate threshold is the same for different genes. Since it is commonly accepted that methylation is associated with silencing of gene expression, we integrate methylation and gene expression data to address the question. The basic idea is: for a methylation controlled gene and its appropriate threshold, its expression should always be low when beta exceeds the threshold, whereas the expression can be either high or low when beta is smaller than the threshold. We used the methylation and gene expression data from The Cancer Genome Atlas (TCGA) project, which contain 997 samples across 7 cancer types. We applied conditional mutual information to examine each gene separately, and identified 798 genes whose expressions are repressed by their methylation. For those methylation controlled genes, we noticed that the appropriate thresholds to binarize their methylation beta values are highly gene-specific.

FastDMR: An Infinium® HumanMethylation450 BeadChip analyzer

Dingming Wu[1], Jin Gu[1,*], Michael Q. Zhang[1,2,*]

[1]: Bioinformatics Division / Center for Synthetic and Systems Biology, Tsinghua National Laboratory for Information Science and Technology (TNLIST); Department of Automation; Tsinghua University; Beijing, China.

[2]: Department of Molecular and Cell Biology, Center for Systems Biology; The university of Texas at Dallas; Dallas, TX USA.

[*]: To whom correspondence should be addressed.

Emails: M.Z. (michael.zhang@utdallas.edu); J.G. (jgu@tsinghua.edu.cn)

DNA methylation is vital for many essential biological processes and human diseases. Illumina Infinium® HumanMethylation450 BeadChip is a recently developed platform studying genome-wide DNA methylation status on more than 480,000 CpG sites and a few CHG sites with high data quality. FastDMR is an analyzer for this chip, which can identify significantly differentially methylated probes, differentially methylated regions (DMRs) in predefined regions, and arbitrary DMRs for both case-control and multigroup studies. Functions other than single probe analysis provide highly interesting result. FastDMR is implemented in C++ and supports multithread computing which makes it very fast and thus suitable for large sample size. FastDMR is freely available at http://fastdmr.sourceforge.net/ and follows the GNU general public license for noncommercial use.

Long range interactions induced by estrogen receptor alpha depend on local open chromatin

Chao He[1], Xiaowo Wang[1,*], Michael Q. Zhang[1,2,*]

[1]: MOE Key Laboratory of Bioinformatics and Bioinformatics Div, Center for Synthetic and System Biology, TNLIST /Department of Automation, Tsinghua University, Beijing 100084, China;

[2]: Department of Molecular and Cell Biology Center for Systems Biology, The University of Texas, Dallas 800 West Campbell Road, RL11 Richardson, TX 75080-3021, USA

[*]: Corresponding authors.

Emails: X.W.W. (xwwang@mail.tsinghua.edu.cn); M.Q.Z. (michaelzhang@tsinghua.edu.cn)

Chromosomes organize into higher-order structure to function. For example, many enhancers regulate their target genes via a long distance interaction. High-throughput experiments like ChIA-PET have been developed to map cell-type specific interactions between regulatory elements. In this study, we integrated multiple types and sources of data, to reveal the general patterns embedded in the ChIA-PET data. We found characteristic distance features among long-range interactions related with promoter-promoter, enhancer-enhancer and insulator-insulator interactions. Our hypothesis is that, although a protein may have many binding sites along the genome, those sites that could share certain open chromatin structure and thus be capable to accommodate relatively larger protein complex containing specific regulatory co-factors and "bridging" partners, should be more likely to associate with long-range interactions. This was at least validated in estrogen receptor alpha (ER) ChIA-PET data. An efficient classifier was provided to help predict ER associated long-range interactions based on ChIP-seq data, to link distal ER-bound enhancers to its target genes. We further applied the classifier to denovo predict thousands of interactions, which was absence in the original ChIA-PET experiments but validated by other experimental data sources. Our work provides an overview of long-range interactions and make up for ChIA-PET experiments, thus point out a way to better understand regulation mechanism and chromosome structures.

Genome-wide noncoding RNA prediction using ENCODE/modENCODE data

Chao Di[1], Long Hu[1], and Zhi John Lu[1] *

[1] MOE Key Laboratory of Bioinformatics, School of Life Sciences, Tsinghua University, Beijing 100084, China

Correspondence should be addressed to Zhi John Lu (zhilu@tsinghua.edu.cn).

Emails: Chao Di (dichao83@gmail.com); Long Hu (hulongptp@gmail.com);

Noncoding RNA as functional genomic elements, played important roles in gene expression and epigenetic regulation in various species. Known ncRNAs could be separate into two groups, one group includes several specific types of shorter ncRNAs, such as miRNA, tRNA, snRNA, snoRNA, etc. which we named canonical ncRNAs, the other group includes only ncRNAs longer than 200nt, which we called lncRNA (long ncRNA). Use the same shuffling method in our previous *incRNA* method, we build machine learning models to predict two types ncRNAs on whole genome in stead of conserved regions. The updated *incRNA* was integrating many types of data in all three organisms, they are expression data by RNA-seq, histone modification data such as H3K4me3, H3K36me3 etc and some TF binding data. All of these data were from ENCODE/modENCODE. We find AUC values for *D. melanogaster* , *C. elegans* and human are all as high as 0.97~0.99 either use canonical ncRNAs or lncRNAs as training set, which prove we could separate known ncRNAs with coding regions or negative control sequences. To further validate the predictions, we carried out RT-PCR experiment in both fly embryos as well as various human tissues, most of the candidates were expressed in at least one condition, which proved our predictions are confident.

Interaction-Based Feature Selection and Classification for High-Dimensional Biological Data

Maggie Haitian Wang[1,2], Shaw-Hwa Lo[3], Tian Zheng[3], and Inchi Hu[1*]

[1]: Department of ISOM, HKUST, Clearwater Bay, Kowloon, Hong Kong

[2]: Division of Biostatistics, School of Public Health and Primary Care, CUHK, Shatin, Hong Kong

[3]: Department of Statistics, Columbia University, New York, USA

[*]: To whom correspondence should be addressed.

Emails: MHW (maggiew@cuhk.edu.hk); SHL (slo@columbia.edu); ZT(tz33@columbia.edu); IH (imichu@ust.hk)

Epistasis or gene-gene interaction has gained increasing attention in studies of complex diseases. Its presence as an ubiquitous component of genetic architecture of common human diseases has been contemplated. However, the detection of gene-gene interaction is difficult due to combinatorial explosion. We present a novel feature selection method incorporating variable interaction. Three gene expression datasets are analyzed to illustrate our method, although it can also be applied to other types of high-dimensional data. The quality of variables selected is evaluated in two ways: first by classification error rates, then by functional relevance assessed using biological knowledge. We show that the classification error rates can be significantly reduced by considering interactions. Secondly, a sizable portion of genes identified by our method for breast cancer metastasis overlaps with those reported in breast cancer database as disease associated and some of them have interesting biological implication. In summary, interaction-based methods may lead to substantial gain in biological insights as well as more accurate prediction.

PiSVM: A new algorithm for predicting piRNA with transposon interaction information and support vector machine

Kai Wang, Fei Li *

Department of Entomology, Plant Protection College, Nanjing Agricultural University

*: To whom correspondence should be addressed.

Emails: Kai Wang (wangkai129000@yahoo.cn); Fei Li (lifei03@tsinghua.org.cn);

Piwi-interacting RNA (piRNA) is a class of small non-coding RNA expressed in germ cells, which can silence transposon at post-transcriptional level. Prediction of piRNA is still a difficult task. Nowadays, only two algorithms, by comparing the genomic piRNA clusters with piRNA sequences (Betel et al., 2007) and piRNApredictor (Zhang et al., 2011), have been reported. With rapid development of next generation sequencing technique, lost of small RNA libraries have been sequenced, which should contain both miRNA and piRNA. However, only miRNA genes were mined from the small RNA library data. Here, we developed a new algorithm named as piSVM to predict piRNA. We matched all known piRNAs of *Drosophila melanogaster* with transposons. The numbers of mismatch were statistically analyzed. To develop a new algorithm, we first matched the small RNA sequences with transposons. The maximum of mismatches was set to seven accordingly. Next, we created pseudo-piRNAs from the exons of mRNAs. We collected 784 true piRNAs and 800 pseudo-piRNA as the training datasets. The 11-dimension features of K-mer strings, the first base and specific heat were analyzed by RNAheat (Vienna RNA Package). These features were used training with libSVM. Then, the trained classifiers were examined with testing dataset containing 196 true piRNAs and 200 pseudo-piRNAs. The results indicated that the accuracy was 95% and the sensitivity was 96%.

Prediction of trans-acting siRNAs in human genome

Xiaoshuang Liu[1], Guangxin Zhang[2], Changqin Zhang[2,*], Jin Wang[1,*]

[1]:The state key laboratory of pharmaceutical biotechnology, School of Life Science, Nanjing University, Nanjing 210093, China 1.

[2]:College of Horticulture, Jinling Institute of Technology, Nanjing 210038, China2.

Emails: Chq Zhang(zcq@jit.edu.cn); J Wang(jwang@nju.edu.cn)

Abstract

Trans-acting siRNAs (ta-siRNAs) are a new class of endogenous RNAs that can repress gene expression through a miRNA-like manner. It was suggested that ta-siRNAs are generated from non-coding transcripts through Argonaute mediated miRNA guided cleavage in plants; and the resulting RNA fragment is further processed by dicer-like enzyme 4 to produce a phased array of 21-nt siRNAs starting at the miRNA cleavage site. We suspect a similar miRNA related gene regulation mechanism in animals. Here， we report the prediction of ta-siRNA-like sequences in human cDNA by computational simulation of ta-siRNA generating process. Human brain sRNAs and human brain degradome data were applied for refine the predictions. Blast, RNAhybrid and bowtie were used to predict the RNA targets. The result shows many new sRNAs from human brain that satisfy the characteristics of ta-siRNAs found in plants, implying that this new small RNAs are likely involved in animals and thus worthwhile for further experimental study.

(This work is supported by the National Science Foundation of China No.J1103512)

Global Discovery of Long Noncoding RNAs in Red Blood Cell Development

Bingbing Yuan*, Wenqian Hu, Juan R. Alvarez-Dominguez, Jiahai Shi, Fran Lewitter and Harvey F. Lodish

Whitehead Institute for Biomedical Research, Cambridge, MA, 02142, USA

*: Bingbing Yuan

Emails: BY (byuan@wi.mit.edu)

Mammalian genomes encode thousands of long noncoding RNAs (lncRNAs). Except for a few dozen lncRNAs with characterized regulatory roles in cell fate and differentiation decisions, the functions of the vast majority of lncRNAs remain unknown

To generate a comprehensive collection of lncRNAs expressed during in vivo erythropoiesis, we employed RNA-Seq profiling of primary mouse fetal liver erythroid progenitors and differentiating erythroblasts. We mapped 462 million 100bp strand-specific paired-end reads to the mouse genome using tophat and assembled transcripts with cufflinks. To retain only reliable transcript models, we applied a read coverage threshold and required transcripts to be multiexonic and >200bp long. To identify high-confidence lncRNAs from this set, we considered only transcripts with no sense overlap with annotated mRNA exons and removed transcripts with high coding potential.

Our erythroid differentiation transcriptome included 655 lncRNA genes. These lncRNAs generally showed greater differentiation-stage specificity than mRNAs in erythroid cells. Knockdown of ten of these lncRNAs revealed their important roles in erythrocyte maturation.

Global Identification and Characterization of Long Non-Coding RNAs in *Arabidopsis*

Jingrui Li [a,d,1], Yang Wu [b,c,d,1], Weilong Guo [e], Michael Q. Zhang [e], Yijun Qi [b,c,d,*]

[a]: College of Biological Sciences, China Agricultural University, Beijing 100193, China

[b]: Tsinghua-Peking Center for Life Sciences, Beijing 100084, China

[c]: School of Life Sciences, Tsinghua University, Beijing 100084, China

[d]: National Institute of Biological Sciences, Zhongguancun Life Science Park, Beijing 102206, China

[e]: Division of Bioinformatics, Center for Synthetic and Systems Biology, TNLIST, Tsinghua University, Beijing 10084, China

[1]: These authors contributed equally to this work

[*]: To whom correspondence should be addressed.

Emails: J.L. (lijingrui@biomed.tsinghua.edu.cn); Y.W. (wuyang.bnu@139.com); Y.Q. (qiyijun@biomed.tsinghua.edu.cn)

Long non-coding RNAs (lncRNAs) have emerged as important regulatory components in major cellular processes in eukaryotes. Here, we developed a stringent selection pipeline for lncRNA identification, combining high-throughput RNA sequencing and computational approaches. Using this pipeline, we annotated 1,353 lncRNAs in Arabidopsis thaliana. Of these, 390 lncRNAs are predominantly polyadenylated and 316 lncRNAs are enriched in the nucleus. Compared to protein-coding genes, lncRNAs have shorter length, fewer exons, lower expression, and lower conservation. The expression of many lncRNAs is developmentally regulated. RNA immunoprecipitation analyses revealed that about one fifth of the lncRNAs are associated with Polycomb repressive complex 2 (PRC2). Some PRC2-associated lncRNAs can repress the expression of their neighboring genes through mediating histone H3 lysine 27 trimethylation. Our catalog of lncRNAs reveals the general properties of lncRNAs in Arabidopsis and paves the way for further functional characterization of these lncRNAs.

Prediction of Disease-causing Nonsynonymous Single Nucleotide Polymorphisms via Integration of Multiple Genomic Data

Jiaxin Wu [1], Rui Jiang [1,*]

[1]: MOE Key Laboratory of Bioinformatics and Bioinformatics Division, TNLIST/Department of Automation, Tsinghua University, Beijing, 100084, China

[*]: To whom correspondence should be addressed.

Emails: JW (wujiaxin0413@gmail.com); RJ (ruijiang@tsinghua.edu.cn)

Detecting associations between human genetic variants and their phenotypic effects is a significant problem in understanding the genetic bases of diseases. We focus on a typical type of genetic variants called nonsynonymous single nucleotide polymorphisms (nsSNPs), which occur may potentially altering structures of proteins, and thereby affecting functions of proteins, and further causing human diseases. Different from some existing methods that formulate the identification of disease-associated nsSNPs as a binary classification problem and give no information about what specific disease the nsSNP is associated with, we formulate the identification of nsSNPs that may be associated with a specific disease from a set of candidate nsSNPs as a prioritization problem. We adopt an approach for predicting novel associations between nsSNPs and diseases that can operate on both diseases with known seed nsSNPs and on novel diseases with no indication information. Specifically, we adopt an adjusted Fisher's method to combine p-values calculated by six popular deleterious prediction scores (SIFT, PolyPhen2, LRT, MutationTaster, GERP and PhyloP), and tailor it to estimate the probability of a nsSNP being disease-causing for a query disease based on integration of multiple genomic data. Then, we compare the predictive powers of six kinds of genomic data (integrated deleterious score, GO analysis, protein protein interaction, protein sequence similarity, functional domains and pathways) in detecting disease-causing nsSNPs from neutral ones. We also demonstrate the effectiveness and performance of our method for both Mendelian diseases and complex disease. Finally, we apply the prediction model to some synthesized proof-of-concept examples with known causal mutations and evaluate the prediction power for distinguishing de novo mutations. Results show that our proposed model is effective and efficient in finding disease-causing nsSNPs for all kinds of diseases with different genetic method of inheritance.

References

1. Joris A. Veltman, Han G. Brunner. De novo mutations in human genetic disease. *Nature Reviews Genetics*, 13:565-575, 2012.
2. Gregory M. Cooper, Jay Shendure. Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. *Nature Reviews Genetics* 12:628-640, 2011.
3. Jiaxin Wu, Rui Jiang. Prediction of Deleterious Nonsynonymous Single Nucleotide Polymorphism for Human Diseases. *The Scientific World Journal*, Volume 2013, Article ID 675851, 2013.
4. James J. Yang. Distribution of Fisher's combination statistic when the tests are dependent. *Journal of Statistical Computation and Simulation*, 80(1):1-12, 2010.

Detecting SNP-SNP Interactions With Piecewise Independence Screening

Seunghak Lee[1], Aurelie Lozano[2], Prabhanjan Kambadur[2], Eric P. Xing[1,*]

[1]: School of Computer Science, Carnegie Mellon University.

[2]: IBM T. J. Watson Research Center.

[*]: To whom correspondence should be addressed.

Emails: S. Lee (seunghak@cs.cmu.edu); A. Lozano (aclozano@us.ibm.com); P. Kambadur (pkambadu@us.ibm.com); E. P. Xing (epxing@cs.cmu.edu)

Interactions between genetic variants are key to understanding the genetic effects on phenotypic traits. However, detecting interaction effects is an ultra-high dimensional problem, and therefore, it is both statistically and computationally challenging. Despite recent breakthroughs in detecting interaction effects, several problems remain including: (1) processing the large number of genetic interactions without compromising for the sake of scalability, (2) solving the multiple testing problem posed by the ultra-high dimensionality of the problem, (3) accounting for strong correlations that exist between SNPs and SNP-SNP pairs, (4) identifying non-linear relationships between genotypes and phenotypes. We present a principled and scalable framework to address these problems in a unified way. Our framework consists of three steps: a screening procedure with piecewise linear model to account for non-linear relationships between genotypes and phenotypes, a procedure for penalized multivariate regression, and a procedure for p-value computation with a correction for multiple testing. The screening procedure is employed to handle the extremely large number of candidate pairs, while the penalized multivariate regression and p-value computation allow the selection of statistically significant SNP-SNP pairs. We demonstrate the effectiveness and scalability of the proposed framework on simulated and real-world datasets. Our results on simulated data show that our framework exhibits higher accuracy vis-a-vis recovering the true associated SNPs and SNP-SNP pairs when compared to existing methods. Those on Alzheimer's data demonstrate that our framework is able to uncover biologically meaningful associations, some as of yet unreported. We also present a high-performance implementation of our screening method, which is highly scalable and is able to screen $O(10^9)$ SNP-SNP pairs in only a few hours.

Using Translational Bioinformatics repertoire to augment understanding of gene polymorphisms implicated in Endometriosis

Roshni Panda [1] and Suresh P.K.[1*]

[1]: School of Biosciences and Technology, VIT University, Vellore, Tamil Nadu India Pin code - 632014

[*]: To whom correspondence should be addressed - **p.k.suresh@vit.ac.in**

Emails: RP (**roshnipanda@vit.ac.in**) and PKS (**p.k.suresh@vit.ac.in**)

Introduction: Endometriosis is a complex gynecological disorder in which endometrial tissue is found in extra –uterine sites leading to severe inflammation and pain. Hundreds of genetic polymorphisms present in several genes have been studied in this disease context. However, none has been assigned with a direct causal link till date.

Methodology: In the present study, we surveyed the Pubmed database and selected 36 Single Nucleotide Polymorphisms from 27 molecular epidemiology studies with case control design. These SNPs are scattered over 12 activation/detoxification genes implicated in pathogenesis of endometriosis – AHR, AHRR, ARNT, CYP1A1, CYP1B1, CYP2C19, CYP2E1, EPHX1, GSTP1, GSTA1, NAT1 and NAT2. Meta Analysis was performed to generate Overall odds ratio (OR) for each SNP across multiple studies using the data given in the epidemiological case-control study and weighted OR were assigned to account for both within- and between-study variations. Additionally, a summary score was generated using prediction scores derived from 4 widely used prediction servers - SIFT, Polyphen, PMut and SNPs3D. We also used a Meta tool – F-SNP to predict the functional significance of each SNP. Consurf server was used to predict the degree of evolutionary conservation of the amino acid residues.

Results: The Spearman's rank correlation coefficient for the summary score and the weighted ORs was $r = 0.569$ ($p < 0.05$). It was found that 22 SNPs were predicted deleterious by F-SNP as they had FS score value between 0.5 to 1.0. Consurf server predicted 15 SNPs to be occurring in evolutionary conserved regions of their respective protein sequences.

Conclusion: To the best of our knowledge, this is the first report of its kind, wherein multiple *in silico* predictive tools for SNP scoring were used to correlate molecular aspects with the clinical phenotype of Endometriosis. Approaches of this kind can aid in the evaluation, development and refinement of predictive approaches for correlative studies of this nature for a better understanding of genes and their involvement in complex diseases like endometriosis.

Hypothesis testing for estimating meiotic recombination rates from population SNP data

Junming Yin [1,*]

[1]: Lane Center for Computational Biology Carnegie Mellon University, Pittsburgh, PA, USA.

[*]: To whom correspondence should be addressed.

Emails: J. Y. (junmingy@cs.cmu.edu)

Crossovers and gene conversions are two known types of meiotic recombination, a central biological process studied in population genetics. It has been observed in the previous studies that when one of these two events is absent in the genealogical model, the population parameter (especially the gene conversion rate) tends to be overestimated by maximum likelihood (or maximum a posterior) point estimation. This is inevitable as the true value lies at the boundary of possible range. Thus, it still remains an open question and an important problem in population genetic studies on how to determine whether a region of a chromosome of interest is subject to crossover or gene conversion only. In this work, we address this problem in a hypothesis testing framework and devise a testing procedure using parametric bootstrap. The likelihood function used in the likelihood ratio test is based on a recent model that explicitly allows overlapping gene conversions in the genealogical process. The performance of our approach is tested on simulated data. The distribution of estimated p-value under the null hypothesis is close to the uniform distribution, demonstrating the correctness of our proposed method; the estimated p-value under the alternative hypothesis shows that the more deviation of the true parameter from zero or the more samples we have, the more likely that the null hypothesis will be rejected by our testing procedure.

Novel modeling of combinatorial miRNA targeting identifies SNP with potential role in bone density

Claudia Coronnello[1,2,*], Ryan Hartmaier[3], Arshi Arora[1], Luai Huleihel[4], Kusum V. Pandit[4], Abha S. Bais[1], Michael Butterworth[5], Naftali Kaminski[4], Gary D. Stormo[6], Steffi Oesterreich[3], Panayiotis V. Benos[1,*]

[1]:Department of Computational and Systems Biology, University of Pittsburgh, Pittsburgh, USA.
[2]: Fondazione Ri.MED, Palermo, Italy.
[3]: Womens Cancer Research Center, University of Pittsburgh, Pittsburgh, USA.
[4]: Dorothy P. and Richard P. Simmons Center for Interstitial Lung Disease, Pittsburgh, USA.
[5]: Department Cell Biology and Physiology, University of Pittsburgh, Pittsburgh, USA.
[6]: Department of Genetics, Washington University School of Medicine, St. Louis, USA.
[*]: To whom correspondence should be addressed.
Emails: CC (clc196@pitt.edu), PVB (benos@pitt.edu)

MicroRNAs (miRNAs) are post-transcriptional regulators that bind to their target mRNAs through base complementarity. Predicting miRNA targets is a challenging task. Until recently, very few algorithms considered the dynamic nature of the interactions, including the effect of less specific interactions, the miRNA expression level, and the effect of combinatorial miRNA binding.

We present a novel thermodynamic model based on the Fermi-Dirac equation that incorporates miRNA expression in the prediction of target occupancy and we show that it improves the performance of two popular single miRNA target finders. Modeling combinatorial miRNA targeting is a natural extension of this model. Two other algorithms show improved prediction efficiency when combinatorial binding models were considered. ComiR (Combinatorial miRNA targeting), a novel algorithm we developed, incorporates the improved predictions of the four target finders into a single probabilistic score using ensemble learning. Combining target scores of multiple miRNAs using ComiR improves predictions over the naïve method for target combination.

ComiR scoring scheme can be used for identification of SNPs affecting miRNA binding. As proof of principle, ComiR identified rs17737058 as disruptive to the miR-488-5p:NCOA1 interaction, which we confirmed *in vitro*. We also found rs17737058 to be significantly associated with decreased bone mineral density (BMD) in two independent cohorts indicating that the miR-488-5p/NCOA1 regulatory axis is likely critical in maintaining BMD in women. With increasing availability of comprehensive high-throughput datasets from patients ComiR is expected to become an essential tool for miRNA-related studies.

The reference part is optional. If your poster is a summary of some papers already published, we strongly recommend you list the published papers in the reference part. If you do not want to include any reference in the abstract, please remove the reference part from the template file.

References

1. Coronnello, Hartmaier, *et al*. Novel modeling of combinatorial miRNA targeting identifies SNP with potential role in bone density. *PLoS Comput Biol*, 8(12): e1002830, 2012.

Discovering and mapping chromatin states using a tree hidden Markov model

Jacob Biesinger[1,3,†] and Yuanfeng Wang[2,†]and Xiaohui Xie[1,3,*]

[1] Department of Computer Science, University of California, Irvine
[2] Department of Physics and Astronomy, University of California, Irvine
[3] Institute for Genomics and Bioinformatics, University of California, Irvine
[*]Corresponding author - xhx@ics.uci.edu
[†]Contributed equally

Abstract. New biological techniques and technological advances in high-throughput sequencing are paving the way for systematic, comprehensive annotation of many genomes, allowing differences between cell types or between disease/normal tissues to be determined with unprecedented breadth. Epigenetic modifications have been shown to exhibit rich diversity between cell types, correlate tightly with cell-type specific gene expression, and changes in epigenetic modifications have been implicated in several diseases. Previous attempts to understand chromatin state have focused on identifying combinations of epigenetic modification, but in cases of multiple cell types, have not considered the lineage of the cells in question.

We present a Bayesian network that uses epigenetic modifications to simultaneously model 1) chromatin mark combinations that give rise to different chromatin states and 2) propensities for transitions between chromatin states through differentiation or disease progression. We apply our model to a recent dataset of histone modifications, covering nine human cell types with nine epigenetic modifications measured for each. Since exact inference in this model is intractable for all the scale of the datasets, we develop several variational approximations and explore their accuracy. Our method exhibits several desirable features including improved accuracy of inferring chromatin states, improved handling of missing data, and linear scaling with dataset size. The source code for our model is available at http://github.com/uci-cbcl/tree-hmm.

A Combinatorial Approach to Characterizing Relationships Between Regulatory Sequences

Christine Lo[1], Boyko Kakaradov[2], Daniel Lokshtanov[1], and Christina Boucher[3]

[1]Department of Computer Science and Engineering, University of California, San Diego
[2]Bioinformatics Graduate Program, University of California, San Diego
[3]Department of Computer Science, Colorado State University

Abstract

RNA splicing is a cellular process driven by the interaction between numerous regulatory sequences and binding sites, however, such interactions have been primarily explored by laboratory methods since computational tools largely ignore the relationship between different splicing elements. Current computational methods identify either splice sites or other regulatory sequences, such as enhancers and silencers. We present an novel approach for characterizing co-occurring relationships between splice site motifs and splicing enhancers. Our approach relies on an efficient algorithm for approximately solving *Consensus Sequence with Outliers*, an NP-complete string clustering problem. In particular, we give an algorithm for this problem that outputs near-optimal solutions in polynomial time. To our knowledge, this is the first formulation and computational attempt for detecting co-occurring sequence elements in RNA sequence data. Further, we demonstrate that SeeSite is capable of showing that certain ESEs are preferentially associated with weaker splice sites, and that there exists a co-occurrence relationship with splice site motifs.

M-NetAligner: A Novel Global Alignment Approach to Identify Functional Orthologs in Multiple Networks

Jialu Hu [1,*], Birte Kehr [1,2], Knut Reinert[1]

[1]: Institut für Informatik, Freie Universität Berlin, Takustr. 9, 14195 Berlin, Germany.

[2]: International Max Planck Research School for Computational Biology and Scientific Computing, Ihnestr. 63-73, 14195 Berlin, Germany.

[*]: To whom correspondence should be addressed.

Emails: J. Hu (jialu.hu@fu-berlin.de); B. Kehr (birte.kehr@fu-berlin.de); K. Reinert (knut.reinert@fu-berlin.de)

Large-scale experimental studies generate genomic data, proteomic data and protein-protein interactions (PPI) at an ever-increasing rate, providing us an opportunity to obtain a deeper understanding of the underlying mechanisms of individual organism at a systematic level. For predicting functionally similar proteins that are shared and conserved by many species, the network alignment approach is an important methodology. Most of the previous network alignment algorithms focus on computing local alignments, which attempt to find conserved protein complexes or partial metabolic pathways. Others compute pairwise global alignment, i.e. focus on finding a best node mapping for two networks. However, there are only a few algorithms for global alignment of multiple networks. For dealing with this problem, here we introduce a fast and effective alignment tool, M-NetAligner. By using an appropriate scoring function, we integrate both topology and sequence information into an alignment score, which turns the alignment problem to an optimization problem. To find a solution, M-Netaligner approximate the highest-scoring alignment by a heuristic method, simulated annealing. To assess its performance, M-NetAligner was applied to real biological networks of four species, C. elegans, D. melanogaster, H. sapiens, and S. cerevisiae. Our results suggest that M-NetAligner outperforms the state-of-the-art approach IsoRank-N in terms of both consistency and speed.

OP-Synthetic: a flux variability analysis based computational framework for synthetic metabolic pathways optimization

Honglei Liu[1], Xiaowo Wang[1,*]

[1]MOE Key Laboratory of Bioinformatics and Bioinformatics Div., Center for Synthetic and System Biology, TNLIST /Department of Automation, Tsinghua University, Beijing 100084, China;
[*]: To whom correspondence should be addressed.

Emails: HL (liuhl3000@163.com); XW (xwwang@mail.tsinghua.edu.cn)

How to optimize microbial strains to produce biochemicals and biofuels has received considerable attention in recent years. Optknock, GDLS, OptForce, OptReg and other computational procedures are designed to find out the reactions to be manipulated for metabolic optimization. However, the previous algorithms sometimes could not work for any topology of metabolic network, especially for the synthetic pathway optimization cases. And they could not take the experimental practice into consideration, such as how to define the number of manipulations. Here, we introduce an computational framework, OP-Synthetic, which could identify all kinds of manipulations (up/down regulation, knock out, substrate addition) using the gradient decreasing method. OP-Synthetic uses flux variability analysis (FVA) to compute the range of flux variability that can be reached under optimal and suboptimal objective states. An optimization procedure step by step is given and users could choose how many steps they want. Moreover, it will distinguish the linked reactions (containing the metabolites that only exist in two reactions) and non-linked reactions when identifying the number of final manipulations. We compared OP-Synthetic with existing methods like Optknock and GDLS on the optimization problems of the Succinate production and N-acetylneuraminic acid synthetic pathway in Escherichia coli. Our method showed a better coincidence with the existing experimental results and gave a reasonable result of synthetic pathway optimization.

Modeling Virus-Host Interactions

N. Sulaimanov[1,2], M. Binder[3], V. Lohmann[3], R. Bartenschlager[3] and L. Kaderali [1,2,*]

[1]: Institute for Medical Informatics and Biometry, Medical School, Technische Universität Dresden, Fetscherstr. 74, 01307 Dresden, Germany.

[2]: ViroQuant Research Group Modeling, BioQuant, University of Heidelberg, Im Neuenheimer Feld 267, 69120 Heidelberg, Germany.

[3]: Department of Infectious Diseases, Molecular Virology, University of Heidelberg, Im Neuenheimer Feld 345, 69120 Heidelberg, Germany.

[*]: To whom correspondence should be addressed.

Email: LK (lars.kaderali@tu-dresden.de)

Hepatitis C virus (HCV) infection is a major global health problem, with 170 million chronically infected individuals worldwide. A main obstacle in treatment is the insidious course of the disease; HCV infection persists in about 80% of patients and is mostly asymptomatic. However, these persons are at high risk to develop liver cirrhosis and hepatocellular carcinoma, making HCV the leading cause of liver transplantation in Europe. There is no vaccine available against HCV, and standard of care treatment with ribavirin, pegylated interferon and protease inhibitors induces long-term responses in approximately 75% of genotype 1 patients, and is associated with severe side effects. While several new anti-HCV compounds are in phase 1 or 2 trials, the development of efficient antiviral drugs is still limited by our poor understanding of the intracellular replication of HCV and its interactions with the host cell.

We use a combination of mathematical modeling, machine learning and bioinformatics approaches to elucidate cellular processes involved in viral infection and viral replication. Based on genome-wide RNA interference experiments, we study host processes exploited by the virus. Machine learning approaches are used to map identified genes to their respective cellular processes. Last but not least, quantitative, dynamic mathematical modeling of the intracellular replication is used to study the replication kinetics, using systems of differential equations. By integrating these approaches, our aim is to gain a deeper understanding of HCV-host interactions, and ultimately to identify new potential drug targets.

We will show results of this approach to understand virus-host interactions, and elucidate the intracellular HCV replication kinetics. Our results specifically show that the participation of host proteins is an essential factor in the formation of intracellular vesicles, the location of RNA genome replication formation. Differences in replication efficiency that is experimentally observed can be explained by differences in the abundance of a host factor involved in this process. Furthermore, our model predicts that HCV fails to replicate successfully without the protection provided by the membranous replication compartment. This is the first time the initial dynamics of HCV replication has been modeled based on time resolved data. Our model may thus provide the basis to better understand the highly dynamic initial steps after infection, and to study in particular the initial race between immune response and successful viral replication.

Network-based gene set perturbation analysis to identify causal or therapeutic miRNAs for cancers

Ting Wang [1], Jin Gu [1], Yanda Li [1,*]

[1]: Bioinformatics Division / Center for Synthetic and Systems Biology, Tsinghua National Laboratory for Information Science and Technology (TNLIST); Department of Automation, Tsinghua University, Beijing, 100084, China.

[*]: To whom correspondence should be addressed.

Emails: TW (w-t09@mails.tsinghua.edu.cn); JG (jgu@tsinghua.edu.cn); YDL (daulyd@tsinghua.edu.cn)

MicroRNA (miRNA) is a class of important post-transcriptional regulator for genes in advanced cells. Nowadays, some miRNAs are detected as oncogenes or tumor suppressors that are causally linked to the emergence and development of cancer, and some are even proposed as potential molecular drugs for cancer therapy. The perturbation of miRNAs may influence not only the expression of their target genes, but also the expression of other secondary targets that are connected or close to those direct targets in molecular interaction networks, and further systematically impact biological processes. Many recent gene set based methods, such as GSEA, only consider the variation of miRNA target genes but ignore the global effect in the whole gene network system. Sometimes target genes are not differentially altered because of the system robustness or feedback mechanism, which makes these methods find few functional miRNAs.

Here we design a network-based gene set perturbation analysis method to identify crucial miRNAs for cancers using normal-disease gene expression data, PPI network and miRNA targets annotation. We apply a Random Walk with Restart (RWR) algorithm to estimate the effect probabilities of all network genes for the perturbation on miRNA target genes, and then integrate the cancer-specific gene expression change information to calculate the global network Effect Score (ES) for the target set perturbation. A permutation-based Sub-GSE algorithm is utilized to test the statistical significance of ES and simultaneously identify the most significantly altered subset, which consists of principal direct and secondary targets.

Using the method we successfully detect experimentally interfered miRNAs from some cellular gene expression change datasets, and then efficiently find out several miRNAs from four cancer datasets respectively, some of which have been verified as vital factors for the cancers, and others with significant ES may also play prominent functions. The identified principal direct and secondary targets are highly enriched in cancer-related gene function annotations, and this may indicate an underlying regulatory mechanism of miRNAs in cancer pathogenesis and progression. Our method is helpful for identifying the causal or therapeutic miRNAs for specific cancers.

ModuleRole: a tool for modulization, role determination and visualization in protein-protein interaction networks

GuiPeng Lee[1], Ming Li[8,9], Rong Li[2,3], Yi Zhao[8],

Roger Guimerà[5,6,7], Michael Q. Zhang[1,4*], Juntao Gao[1*]

[1]: MOE Key Laboratory of Bioinformatics;Bioinformatics Division and Center for Synthetic & Systems Biology, TNLIST; Department of Automation, Tsinghua University, Beijing 100084, China

[2]: Stowers Institute for Medical Research, 1000 East 50th Street, Kansas City, MO 64110

[3]: Department of Molecular and Integrative Physiology, University of Kansas Medical Center, 3901 Rainbow Boulevard, Kansas City, KS 66160

[4]: Department of Molecular and Cell Biology, Center for Systems Biology, The University of Texas at Dallas, Richardson, TX 75080, USA

[5]: ICREA and [6]: Department of Chemical Engineering, Universitat Rovira i Virgili, 43007 Tarragona, Catalonia

[7]: Department of Chemical & Biological Engineering and Northwestern Institute on Complex Systems (NICO), Northwestern University, Evanston, IL 60208

[8]: Bioinformatics Research Group, Key Laboratory of Intelligent Information Processing, Advanced Computer Research Center, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, PR China

[9]: University of Chinese Academy of Sciences, Beijing, PR China.

[*] To whom correspondence should be addressed. E-mail: jtgao@biomed.tsinghua.edu.cn, Mzhang@cshl.edu.

Although protein-protein interaction (PPI) databases such as BioGrid provide powerful online interaction repositories through comprehensive curation efforts, interpretation of these data remains a major challenge. The specialized tools for the analysis of biological networks such as NetworkAnalyzer etc. can calculate a set of topological parameters, but they did not offer valid way to gain insight into the structure, especially modules, of PPI networks.

Using simulated annealing and a method based on the node connectivity, we developed ModuleRole, a user-friendly web server tool which finds modules in PPI network and defines the roles for every node, and produces files for visualization in Pajek. The input of ModuleRole is: (1) a list of proteins provided by user; (2) the species selected by user. The output of MoudleRole is: (1) modules in PPI physical and genetic network; (2) role for every node; (3) files used for visualization in Pajek.

This program can be tested at the website http://www.bioinfo.org/modulerole/index.php, which is free and open to all users and there is no login requirement**,** using demo data provided by "User Guide" in the menu Help. Non-server application of this program is only considered for high-throughput data with protein node number >=200. Users are able to bookmark the web link to the result page and access at a later time. ModuleRole requires no expert knowledge in graph theory on the user side, thus a useful tool for biologist to analyze and visualize PPI networks in databases such as BioGrid.

Protein Ranking Algorithm Improves the Identification of Protein Complexes from the Protein-Protein Interaction Network

Nazar Zaki[1,*], Jose Berengueres[1], Dmitry Efimov[1]

[1]: Bioinformatics Lab, College of Information Technology, United Arab Emirates University

[*]: To whom correspondence should be addressed.

Emails: NZ (nzaki@uaeu.ac.ae); JB (joseb@uaeu.ac.ae); DE (dmitry@uaeu.ac.ae)

Detecting protein complexes from protein-protein interaction (PPI) network is becoming a major focus of researchers in computational biology. There is ample evidence that many disease mechanisms involve protein complexes, and being able to predict these complexes is important to the characterization of the relevant disease for diagnostic and treatment purposes. This highlighted paper introduced a novel method for detecting protein complexes from PPI by using a protein ranking algorithm (ProRank).

ProRank is inspired by Google page rank algorithm which quantifies the importance of each protein based on the interaction structure and the evolutionarily relationships between proteins in the network. A novel way of identifying essential proteins which are known for their critical role in mediating cellular processes and constructing protein complexes was proposed and analyzed. ProRank was evaluated on two PPI networks and two reference sets of protein complexes created from MIPS. The level of the accuracy achieved using ProRank is a strong argument in favor of the highlighted method.

Population dynamics of cancer cells with cell-state conversions between cancer stem cells and non-stem cancer cells

Da Zhou[1], Dingming Wu[1], Zhe Li[2], Minping Qian[3] and Michael Q. Zhang [4,1,*]

[1]: MOE Key Laboratory of Bioinformatics; Bioinformatics Division/Center for Synthetic & Systems Biology, TNLIST; Department of Automation, Tsinghua University, Beijing 100084, China.

[2:] Computational Neuroscience Lab, School of Medicine, Tsinghua University, Beijing 100084, China.

[3:] School of Mathematical Sciences, Peking University, Beijing 100871, China.

[4:] Department of Molecular and Cell Biology, Center for Systems Biology, The University of Texas at Dallas, Richardson, TX 75080, USA

[*]: To whom correspondence should be addressed.

Emails: D. Z (zhouda1112@gmail.com); D.W (wdm2008@gmail.com); Z. L (lizhe.tsinghua@gmail.com); M. Q (qianmp@math.pku.edu.cn); M. Q. Z (michael.zhang@utdallas.edu)

Cancer stem cell (CSC) theory suggests a cell-lineage structure in cancer that CSCs are capable of giving rise to the other non-stem cancer cells (NSCCs) but not vice versa. However, an alternative scenario of bidirectional inter-conversions between CSCs and NSCCs was proposed in [Gupta PB, et al. (2011) Cell 146: 633644]. Here we present computational analyses for further investigating the relation between CSCs and NSCCs through population modeling of cancer cells, where not only can CSCs differentiate into NSCCs by asymmetric cell division, NSCCs can also dedifferentiate into CSCs by cell state conversions. By fitting our model to recent experimental data, it is shown that the conversion from CSCs to NCSSs explains the transient increase in the proportion of CSCs initiated from the purified NSCCs subpopulation. Similar result is also present when generalizing our model to multi-type compartmental cell case, indicating that cell-state conversions in cancer play an important role in effectively keeping the heterogeneity in the population of cancer cells.

Structure Identification for Gene Regulatory Networks via Linearization and Robust State Estimation

Jie Xiong[1,*], Tong Zhou [1,2]

[1]: Department of Automation, Tsinghua University, Beijing, 100084, China.

[2]: Tsinghua National Laboratory for Information Science and Technology, Beijing, 100084, China.

[*]: To whom correspondence should be addressed.

Emails: J. Xiong (xiongj08@mails.tsinghua.edu.cn); T. Zhou (tzhou@mail.tsinghua.edu.cn)

Inferring causal relationships among numerous cellular components is one of the fundamental problems in understanding biological behaviors. The gene regulatory network is widely considered as a nonlinear dynamic stochastic model that consists of the gene measurement equation and the gene regulation equation, in which the extended Kalman filter (EKF) is sometimes used for estimating both the model parameters and the actual value of gene expression levels. However, first-order linearization usually results in modeling errors, but the EKF based method does not take either unmodelled or parametric uncertainty into account. As a result, the estimation performance of the EKF based method may not be satisfactory, such as slow convergence speed and low estimation accuracy. To overcome these problems, a sensitivity penalization based robust state estimator is suggested for reconstructing the structure of a gene regulatory network. The suggested method has been used to identify some parameters of a nonlinear state-space system, and recovery an artificially gene regulatory network. Compared with the widely adopted EKF based method, computation results show that parametric estimation accuracy can be significantly increased and false positive errors can be greatly reduced.

Translating integrated multi-omics study of tumors to improve clinical outcomes

Gang Chen[1*], Yong Hou[1], Zhibo Gao[1], Guoqing Li[1]

[1]: BGI-Shenzhen, Shenzhen, China.

[*]: To whom correspondence should be addressed.

Emails: chengang@genomics.cn

We describe here an on-going tumor study using integrated multi-omics to generate results that will contribute to clinical applications. With the rapid development of high-throughput technologies, massive amounts of data on genomes, epigenomes, transcriptomes, proteomes, interactomes, metagenomes and metabolomes have been generated leading to analyses of tumors at unprecedented resolutions. Although significant advances in tumor research have resulted using these technologies, the complexity of the data impedes our ability to thoroughly understand the underlying mechanisms and translate discoveries into clinical outcomes. To obtain a comprehensive understanding of tumors, we proposed an integrated, multi-omics analysis, in which we systematically integrate various omics data produced by different technologies.

We approach this integrated multi-omics study of tumor via a pilot project for prostate tumors. In this study, transcriptomes and whole-genomes analyses are integrated with interactomes to elucidate drugable genes and the subnetworks which may pose risks to prostate tumor. By integrating multi-omics data, we build weighted gene networks for tumors and normal tissues to find significantly mutated sub-networks. Functional analysis of these sub-networks indicates that these sub-networks are related to some well-known cancer pathways, such as “Axon guidance”. By using network-based drug target and drug target interaction prediction methods, we found some potential drug-targets and drug-target interactions in these tumor-associated sub-networks.

We also discuss implications for future research including development of this framework and novel methods of applying multi-omics studies to clinical applications.

A novel breakpoint based algorithm to detect structural variation in cancer genomes

Hui Zhao and Fangqing Zhao*

Computational Genomics Lab, Beijing Institutes of Life Science, Chinese Academy of Sciences, Beijing 100101

*: To whom correspondence should be addressed.

Email: F.Z.(zhfq@mail.biols.ac.cn)

Extensive studies reveal that cancer genomes are frequently altered in their chromosomal structure by insertion, deletion, translocation and inversion of DNA fragments. Such alterations (termed as structural variation, SV) may produce various phenotypic effects in tumorigenesis and various human genetic disorders. However, current SV detection algorithms are far from being perfect and have limits in terms of the type and size of SVs that they are able to detect. There still lacks a one-stop solution for full range of structural variant detection, especially for both small INDELs and complex forms of SVs. Here we propose a breakpoint-based SV discovery strategy based on decoding abnormally aligned paired end reads from SAM files. This approach can successfully detect both homozygous and heterozygous indels, whose sizes range from as small as several base pairs to as large as several thousand base pairs, with a high accuracy of breakpoint estimation simultaneously. Unlike most other paired-end mapping and depth of coverage based algorithms, our method not only is capable of detecting small indels (10-50 bp) and their breakpoints, but can also distinguish nested SVs within three standard deviations of the insert size. By applying this approach to cancer genome sequencing datasets, we have successfully uncovered a significant amount of novel INDELs that were missed before.

Systematic Identification of Synergistic Drug Pairs Targeting HIV

Xu Tan[1,2], **Long Hu**[3#], Lovelace J. Luquette III[4#], Geng Gao[1,2], **Yifang Liu**[3], Hongjing Qu[1,2], Ruibin Xi[4], **Zhi John Lu**[3] *, Peter J. Park[4] * & Stephen J. Elledge[1,2]

[1]Department of Genetics,

[2]Howard Hughes Medical Institute, Brigham and Women's Hospital, Harvard Medical School, Boston, MA 02115, USA

[3]MOE Key Laboratory of Bioinformatics, School of Life Sciences, Tsinghua University, Beijing 100084, China

[4]Center for Biomedical Informatics, Harvard Medical School, Boston, MA 02115, USA

# These authors contributed equally to the work.

* These authors contributed equally to the work.

Correspondence should be addressed to S.J.E (selledge@genetics.med.harvard.edu).

Speaker Email: hulongptp@gmail.com

Combination drug therapies play important roles in treating diseases such as cancer and AIDS. However, systematic identification of effective drug combinations has been hindered by the large combinatorial search space of interactions. Here we develop a multiplex screening method, *MuSIC* (*Mu*ltiplex *S*creening for *I*nteracting *C*ompounds), which expedites comprehensive assessment of pair-wise interactions for 1000 FDA-approved or clinically tested drugs. In this way we examined ~500,000 drug pairs and identified drugs that synergize to inhibit HIV replication. Multiple drug pairs, notably glucocorticoid and nitazoxanide, synergize by targeting different steps of the HIV life cycle. Our analysis also reveals an enrichment of anti-inflammatory drugs, i.e. glucocorticoids and NSAIDs, in the anti-HIV drug combinations. As inflammation accompanies HIV infection, our findings suggest that HIV may benefit from inflammation and inhibiting inflammation might combat HIV propagation. The *MuSIC* method is robust and can be widely applied to other disease-relevant screens to facilitate drug repurposing.

Efficient methods for identifying mutated driver pathways in cancer

Junfei Zhao [1], Shihua Zhang [1,*], Ling-Yun Wu [1], Xiang-Sun Zhang [1]

[1]: National Center for Mathematics and Interdisciplinary Sciences, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China.

[*]: To whom correspondence should be addressed.

Emails: SZ (zsh@amss.ac.cn)

The first step for clinical diagnostics, prognostics, and targeted therapeutics of cancer is to comprehensively understand its molecular mechanisms. Large-scale cancer genomics projects are providing a large volume of data about genomic, epigenomic, and gene expression aberrations in multiple cancer types. One of the remaining challenges is to identify driver mutations, driver genes and driver pathways promoting cancer proliferation and filter out the unfunctional and passenger ones.

In this study, we propose two methods to solve the so called Maximum Weight Submatrix problem which is designed to de novo identify mutated driver pathways from mutation data in cancer. The first one is an exact method which can be helpful for assessing other approximate or/and heuristic algorithms. The second one is a stochastic and flexible method which can be employed to incorporate other types of information to improve the first method. Particularly, we propose an integrative model to combine mutation and expression data. We first apply our methods onto simulated data to show their efficiency. We further apply the proposed methods onto several real biological data sets such as the mutation profiles of 74 head and neck squamous cell carcinomas samples, 90 glioblastoma tumor samples and 313 ovarian carcinoma samples. The gene expression profiles were also considered for the latter two data. The results show that our integrative model can identify more biologically relevant gene sets. We have implemented all these methods and made a package called MDPFinder (Mutated Driver Pathway Finder) which can be easily used for other researchers.

References

1. Junfei Zhao, Shihua Zhang, Ling-Yun Wu, Xiang-Sun Zhang. Efficient methods for identifying mutated driver pathways in cancer. *Bioinformatics*, 28(22): 2940-2947, 2012.

Applying Co-occurrence Network Construction and Analysis on Human Tongue Coating Microbiome

Lianshuo Li [1], Rui Jiang [1,*]

[1]: Bioinformatics Division, Department of Automation and Tsinghua National Laboratory for Information Science and Technology Tsinghua University, Beijing 100084, China.

[*]: ruijiang@tsinghua.edu.cn

Constructing and analyzing co-occurrence networks is an efficient method which has been used in microbial ecology to describe and study the ecological relationships of microbes living in certain environments. As the method is being improved in last few years, its applications are also extended to microbes living in human bodies. Recently we showed a great interest in this kind of research. We first studied the process of network construction and analysis based on correlation in detail. Then we construct OTU networks based on 16S rRNA gene datasets for analysis. We treat each OTU as a node, then corrections in abundance between each pair of OTU may reflect its underlying relationships such as symbiosis. We collected microbial data from three kinds of human tongue dorsum which are diagnosed as normal, hot syndrome or cold syndrome by Chinese traditional medicine. Networks were constructed for the three datasets respectively. Differences on network topology found here gave us a great passion on finding detailed structural and biological differences of microbiome living on tongue dorsum with different status. Along with this study, we also find an interesting thing that different methods in constructing networks leading to very different network topologies. Combining different methods in network construction are turned to be effective and robust. So we will improve our work quickly.

Gene prioritization for attention deficit hyperactivity disorder by integrating multi-evidence score system and random walk interactome

Suhua Chang [1], Jing Wang [1*]

[1]: Key Laboratory of Mental Health, Institute of Psychology, Chinese Academy of Sciences, 16 Lincui Rd., Chaoyang District, Beijing 100101, China.

[*]: To whom correspondence should be addressed.

Emails: SHC (changsh@psych.ac.cn); JW (wangjing@psych.ac.cn)

Attention deficit hyperactivity disorder (ADHD) is a common, highly heritable psychiatric disorder with heritability estimates 75%-91%. Previously, we have developed the first genetic database for ADHD (ADHDgene) to provide researchers a comprehensive ADHD genetic resource [1]. The abundant genetic data provides novel candidates for ADHD genetic study, but it also brings new challenge for selecting promising candidates for replication and verification research. Gene prioritization is an effective method to deal with this problem. However, most of gene prioritization tools available need training genes as input and by now there are no definitive ADHD genes as training data. In this study, we developed an integrated method to combine multi-evidence score system and random walk interactome to prioritize ADHD candidate genes. 3,589 ADHD candidate genes from ADHDgene were included for the analysis. Different gene sources in ADHDgene (including literature-origin, mapped by SNP, mapped by LD-proxy, mapped by copy number variation or region or pathways), gene expression profile in animal model and ADHD related brain regions were taken as multiple evidences to build score system. To get the best weights, we used the middle result of score system to generate training data and test data to do gene prioritization using random walk interactome method. The best weight should make more test genes with high ranking in score system have higher ranking in the gene prioritization result based on random walk interactome. The final ranking result was evaluated by using ADHD genome-wide association study (GWAS) data. Finally, the integrated gene prioritization method selected 51 prioritized genes. When using 25 of them as training genes and other 26 genes together with other candidate genes as test genes to do random walk interactome gene prioritization analysis, 13 of the top 26 genes were overlapped with the 26 high ranking test genes, with 50% identity. Analysis of the distribution of prioritized genes and all candidate genes in two ADHD GWAS data showed the *P*-values of the prioritized genes were significantly smaller than all candidate genes. Functional analysis for the prioritized genes showed they were mainly involved in neurotransmitter systems and nervous system development pathways. By integrating multiple data sets, the prioritized genes provided in this study will provide reliable candidates for further genetic and functional analysis for ADHD, especially the genes supported by interactome.

References

1. Zhang, L., S. Chang, Z. Li, K. Zhang, Y. Du, J. Ott and J. Wang. ADHDgene: a genetic database for attention deficit hyperactivity disorder. *Nucleic Acids Res* 40(Database issue): D1003-1009, 2012.

AMBIENT: active modules for bipartite networks

William A Bryant [1], Michael JE Sternberg [1], John W Pinney [1,2]

[1]: CISBIO, Imperial College London, South Kensington Campus, SW7 2AZ, UK.

[2]: Division of Molecular Biosciences, Imperial College London, South Kensington Campus, London SW7 2AZ, UK.

[*]: To whom correspondence should be addressed.

Email: Pinney, JW (j.pinney@imperial.ac.uk)

With the continued proliferation of high-throughput biological data there is a pressing need for tools to integrate these data and come to biologically meaningful conclusions with them. The increased number of available metabolic models for organisms analysed in high-throughput data-producing experiments gives the opportunity to use the properties of these metabolic networks as-is to analyse these data in system-wide, but highly specific ways. We have developed AMBIENT[1] (Active Modules for Bipartite Networks), a tool that uses simulated annealing to discover connected metabolic subnetworks (modules) that are affected by some genetic or environmental change. These modules are found by taking advantage of metabolic network topology combined with highthroughput data to build areas of the network representing coordinated changes in metabolism.

This approach returns biologically meaningful metabolic modules across the entire metabolic network and adds insight into metabolic processes through specific identification of directly affected sets of connected reactions in a particular environment. AMBIENT represents a significant advancement in the analysis of high-throughput data in a metabolic context. The active modules approach to coordinated-pathway finding fits between the individual gene transcription tables which are common in current microarray studies and general functional classifications such as GSEA and pathway enrichment analysis. This approach can be used in any system in which reactions (or metabolites) can be assigned a score based on some biological observation without the limitation of predefined pathways, and with the ability to come to highly specific conclusions about metabolism in that biological context, making AMBIENT widely applicable across the spectrum of highthroughput experimentation.

References

1. William A Bryant, Michael JE Sternberg, John W Pinney. AMBIENT: Active Modules for Bipartite Networks. Using high-throughput transcriptomic data to dissect metabolic response. *In submission.*

A Comparative Study of Reverse Engineering of Biological Network Using Prior Networks: Local and Global Methods

Yang Xiang [1,*], Florian Martin [1], Joe Whittaker [2]

[1]: Philip Morris Research and Development, CH-2000 Neuchâtel, Switzerland

[2]: Department of Mathematics and Statistics, Lancaster University, UK

[*]: To whom correspondence should be addressed.

Emails: YX (Yang.Xiang@pmi.com); FM (Florian.Martin@pmi.com); JW (joe.whittaker@lancaster.ac.uk)

Identification of the interactions between molecular entities within cells is the key to understanding the biological processes involved. Although numerous methods have been developed for inferring gene/protein regulatory networks from expression data, reliable network inference remains an unsolved problem. There are some published methods which use prior network information to improve the inferred network. In our work this approach is extended with the development of a new method, called "contracting neighborhood" (CN), and the application of graphical Lasso (GLASSO) to include prior network information, that is called REBNR (reverse engineering of biological network by regularization). A comprehensive framework, valuable for comparing available alternative methods in a fair and systematic way, is developed. It starts with a Gaussian assumption, and generates inversed covariance matrix and realistically sized experimental data sets over an enormous combination of parameters. A ten times 5-fold cross validation is performed automatically to refine the parameters of every algorithm/method according to the simulated data set. Several reverse engineering methods, including ARACNE, CN, RN (relevance network), MRNET, CLR, PC, GLASSO, ZPC (Zhang) and REBNR are evaluated, with and without prior network knowledge. Our results confirm that the inferred network can be improved but the percentage of improvement depends on the quality of prior network and the way of incorporating prior network. Our CN performs comparably with current best algorithms. The top ranking methods are applied to a real protein dataset to illustrate the biological insight gained from the built network.

Adding uncertainty to biological networks improves clustering results

Benoît Robisson [1,3,*], Alain Guénoche [2,3], Christine Brun [1,3]

[1]: TAGC, Inserm.

[2]: IML, CNRS.

[3]: Aix-Marseille Université, Marseille, FRANCE.

[*]: To whom correspondence should be addressed.

Email: BR (benoit.robisson@inserm.fr)

The inherent uncertainty in biological data is a recurrent problem when analyzing biological —such as protein protein interaction (PPI)—networks. Here, inspired by classical methods in phylogeny, we define bootstrap clustering[1] adding uncertainty to the graph. Our method iteratively alters the network, adding some edges and weighting them using the Dice index; a potential partition is then computed for each altered network; and a consensus of partitions is made from the profile of potential partitions.

We have evaluated our approach by generating a reference partition and its corresponding network. According to the Rand index, the consensus partition of altered networks (ANs) is closer to the reference partition than that of unaltered networks (Uns).

We have also applied our method to PPI networks and have developed a functional homogeneity score to assess the biological relevance of the clusters obtained before and after the introduction of uncertainty. Strikingly, partitions of the ANs again scored better than partitions of the Uns.

Our analysis shows that for sparse—such as PPI—networks, reasonable levels of alteration improve clustering results and lead to more meaningful biological clusters. These findings call for a reconsideration of the effect of uncertainty in protein interaction networks: what has long been considered a drawback could well improve the results of clustering algorithms.

References

1. Philippe Gambette & Alain Guénoche. Bootstrap clustering for graph partitioning. *RAIRO – Operations Research*, 45: 339-352, 2012.

Gene signatures of proliferating B cells predict response to influenza vaccination

Yan Tan[a,b], Pablo Tamayo[a], Helder Nakaya[c], Bali Pulendran[c], Jill Mesirov[a,b,*], W. Nicholas Haining[b,d,e,*].

[a]Bioinformatics Program, Boston University, Boston, Massachusetts, United States of America, [b]Broad Institute of Harvard and MIT, Cambridge, Massachusetts, United States of America, [c]Emory Vaccine Center, Yerkes National Primate Center, Department of Pathology, 954 Gatewood Road, Atlanta, GA 30329, USA, [d]Department of Pediatric Oncology, Dana-Farber Cancer Institute, Harvard Medical School, Boston, Massachusetts, USA, [e]Division of Hematology/Oncology, Children's Hospital, Harvard Medical School, Boston, Massachusetts, USA

Vaccines are very effective at preventing infectious disease but not all recipients of a vaccine mount a protective immune response to vaccination. Recently, measurements of gene expression profiles from peripheral blood mononuclear cell samples in vaccinated individuals have been used to develop models to predict the development of protective immunity. However, there are two major barriers to the development of gene expression-based predictors of vaccine response. First, the magnitude of change in gene expression profile that separates vaccine responders and non-responders is likely to be small and distributed across networks of genes, making the selection of individual predictor genes difficult. Second, selecting biologically important predictive genes is difficult because the mechanistic relevance of individual predictive genes is often obscure. Here we apply a new approach to developing gene expression predictors of vaccine response based on detecting coordinate up-regulation of sets of biologically informative genes in post vaccination gene expression profiles. We found that enrichment of gene sets related to proliferation and immunoglobulin genes accurately segregated subjects with high antibody response to influenza vaccination from low responders (AUC 0.94) and predicted the antibody response in a blind validation set with an accuracy of 88%. The enrichment of these gene sets was highly correlated with the frequency of plasmablasts in post-vaccination blood samples. However, many of the genes in these predictive gene sets would not have been identified using conventional, single-gene approaches because of their small fold change in responders. Our results demonstrate that gene expression predictors based on gene set enrichment can capture subtle transcriptional changes such as those caused by an increase in a rare population of cells in post vaccine samples. Methods that use sets of biologically informative genes as predictive features may be a generally useful approach for developing and interpreting predictive models of the human immune response.

Prioritizing disease candidate genes by integrating multiple biological networks from diverse databases

Yuanhua Huang, Peng He, Rui Jiang
MOE Key Laboratory of Bioinfomatics and Bioinfomatics Division, TNLIST/Department of Automation, Tsinghua University, Beijing 100084, China

E-mail: Yuanhua Huang – huangyh09@mails.tsinghau.edu.cn ,
Peng He - hep10@mails.tsinghua.edu.cn , Rui Jiang - ruijiang@tsinghua.edu.cn

Abstract | It is widely accepted that no matter the Mendelian diseases or the complex disorders are related to the genetic inheritance. Thus, finding out the association between inherited phenotype and genotype may have a great significance in the diagnosis and treatment of these disorders, though it is still a challenging problem. Our goal was to develop some sophisticated machine learning tools to integrate diverse biological database for prioritizing and predicting these diseases related genes based on the hypothesis of Guilty-by-Association. There are two main parts in our project. In the first part, we established some disease networks and analyzed them; meanwhile we proposed methods to calculate the similarity of genes from multiple biological networks, such as protein interaction networks and gene regulation networks. Then we obtained the features from these two types of network at both disease level and gene level. In the second part, we developed the suitable machine learning methods to integrate the multiple features. We formulated an ensemble machine learning method combining with the thoughts of boosting to solve the data missing when integrating features. Finally, with the effective analysis on both disease and gene networks, as well as the appropriate machine learning methods, our methodology shows a great power in the prioritization and prediction of the candidate genes.

Random walking on a tissue specific protein-protein interaction network for the discovery of disease-related protein-complexes

Thibault Jacquemin[1], Rui Jiang[1,*]

[1]: MOE Key Laboratory of Bioinformatics and Bioinformatics Div, TNList, Department of Automation, Tsinghua University, BEIJING

[*]: To whom correspondence should be addressed.

Emails: T.J. (thibault.jacquemin@centrale-marseille.fr); R.J. (ruijiang@tsinghua.edu.cn);

In the study of human genetic disorders, scientists try to identify the genes and the biological mechanisms which come at stake in the apparition of the diseases. In this regard, we focused on protein complexes, which are groups of multiple gene products aggregating to perform cellular functions, and developed a method to sort the protein-complexes according to their closeness to the input disease.

This method is based on a three level integrated network composed of three types of nodes (phenotypes, proteins, and protein complexes) and four types of interactions (literature based phenotype similarity (weighted) , known protein-disease associations (non-weighted), tissue specific protein-protein interactions (weighted), and protein-complex membership (non-weighted)). We used 60 different tissue specific protein-protein interaction networks based on gene expression analysis. Wherein, the disease tissue association is literature based. Finally, for each input disease, we run a random walk through the all network to assess the connections between the disease and each protein complex.

Our results show that we can correctly identify many disease-related complexes (first-ranked), except when their subunits belong to too many complexes. Besides, the contribution of the tissue specificity does not seem to be that significant.

References

1. Börnigen Daniela et al. An Unbiased Evaluation of Gene Prioritization Tools. *Bioinformatics*: 1-8, 2012
2. Lage Kasper et al. A Large-scale Analysis of Tissue-specific Pathology and Gene Expression of Human Disease Genes and Complexes. *Proceedings of the National Academy of Sciences of the United States of America* 105, no. 52: 20870–5, 2008
3. Lage Kasper et al. A Human Phenome-interactome Network of Protein Complexes Implicated in Genetic Disorders. *Nature Biotechnology* 25, no. 3: 309–16., 2007
4. Macropol Kathy et al. Repeated Random Walks on Genome-scale Protein Networks for Local Cluster Discovery. *BMC Bioinformatics* 10, no. Mc: 283, 2009
5. Magger Oded et al. Enhancing the Prioritization of Disease-Causing Genes Through Tissue Specific Protein Interaction Networks. *PLoS Computational Biology* 8, no. 9: e1002690, 2009
6. Yang Peng et al. Inferring Gene-phenotype Associations via Global Protein Complex Network Propagation. *PloS One* 6, no. 7: e21502, 2011

Differential methylation analysis for the identification of epigenomic factors in HBV vaccination responses

Youtao Lu[1], Yi Chen[2], Weili Yan[2*], Christine Nardini[1*]

[1]: Group of Clinical Genomic Networks, Key Laboratory of Computational Biology, CAS-MPG Partner Institute for Computational Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai, PR China.

[2]: Department of Clinical Epidemiology, Children's Hospital of Fudan University, Shanghai, PR China.

[*]: To whom correspondence should be addressed.

Emails: WY (yanwl@fudan.edu.cn); CN (christine@picb.ac.cn).

Hepatitis B virus (HBV) is the pathogen responsible for hepatitis B, which has a worldwide prevalence and is a potential contributor to liver cancer. Vaccination against HBV is an effective way to prevent the disease, however the responses vary among people. In order to search for the epigenomic factors underlying the differential responses, 25 infants were grouped as good/poor responders, and their whole-genome DNA methylation levels were sampled using the Illumina HumanMethylation® 450K microarray. Despite methylation array technology being used now for some time, the peculiar distribution of methylation data (almost bimodal) cannot take advantage directly of the popular differential analysis tools based on normal distribution, developed for other omic data array. Therefore, we developed a custom pipeline for the data analysis: after basic quality filtering, we applied dispersion filtering to narrow down the candidate loci, and we then performed differential methylation analysis based on a combination of the average β difference and a series of statistical tests, including Wilcoxon rank-sum test and Fisher's exact test (after data stratification), to warrant stability in the loci selection. We finally identified 146 differentially methylated loci, and complied them into a *core* (stringent) and *extended* (relaxed) list of loci, according to the degree of differential evidence and methylation reliability. This double categorization (core/extended) allows more flexibility in the definition of further validation candidates. In particular, we were able to identify several loci corresponding to RNF39, a gene that is within the MHC class I region and thought to be related to immunity, found to be hypo-methylated.

Differential regulation enrichment analysis method (DREAM): a novel gene set analysis method based on the gene regulatory network

Shining Ma[1], Tao Jiang[2,1], Rui Jiang[1,*]

[1]: MOE Key Laboratory of Bioinformatics and Bioinformatics Division, TNLIST/Department of Automation, Tsinghua University, Beijing 100084, China

[2]: Department of Computer Science and Engineering, University of California, Riverside, CA 92521, USA

[*]: To whom correspondence should be addressed. Address: FIT 1-107, Tsinghua University, Beijing 100084, China.

Email: Rui Jiang (ruijiang@tsinghua.edu.cn)

Although many gene set enrichment analysis methods have been proposed to explore associations between a phenotype and a group of genes sharing common biological functions or involved in the same biological process, the underlying biological mechanisms of identified gene sets are typically unexplained. To overcome this limitation, we propose a method called DREAM (Differential Regulation Enrichment Analysis Method) to identify gene sets in which a significant proportion of genes have their transcriptional regulatory patterns changed in a perturbed phenotype. We conduct comprehensive simulation studies to demonstrate the capability of our method in identifying differentially regulated gene sets. We further apply our method to two independent human microarray expression data sets, both with hormone-treated and control samples. Our results indicate that DREAM's ability to rank hormone-associated gene sets among the most enriched gene sets is significantly superior to that of three existing methods. We conclude that the proposed differential regulation enrichment analysis complements the existing gene set enrichment analysis methods and provides a promising new direction for the interpretation of gene expression data.

Characterization of regulatory features of housekeeping and tissue-specific genes with tissue regulatory networks

Pengping Li, Xu Hua, Zhen Zhang, Jie Li*, Jin Wang *

The State Key Laboratory of Pharmaceutical Biotechnology, Jiangsu Engineering Research Center for MicroRNA Biology and Biotechnology, School of Life Science, Nanjing University, China

*Corresponding authors: Jie Li & Jin Wang, Tel: 025-83686785,

Email: jwang@nju.edu.cn

Abstract

Transcription factors (TFs) and miRNAs are essential for the regulation of gene expression; however, the global view of human regulatory networks remains poorly understood. We analyzed the network properties of housekeeping and tissue-specific genes in gene regulatory networks of seven human tissues. The results showed that different classes of genes behaved quite differently in the networks. The housekeeping TFs tended to have higher cluster coefficients compared to other genes that are neither housekeeping nor tissue-specific, indicating that housekeeping TFs tend to regulate their targets synergistically. The tissue-specific TFs had a more significant topological bias, indicating that tissue-specific TFs transfer information from upstream pathways to downstream pathways more quickly than other TFs. Tissue-specific miRNAs showed a higher average number of targets while their targets had a lower indegree, which indicates that this class of miRNAs regulates a greater number of targets and that this regulation is independent of other regulators. Several topological properties of disease-associated miRNAs and genes were found to be significantly different from those of non-disease associated miRNAs and genes. Determining the network properties of these regulatory factors will help define the basic principles of human gene regulation and the molecular mechanisms of disease.

Repeat-Enriched Proteins Are Related to Host Cell Invasion and Immune Evasion in Parasitic Protozoa

T.A.O. Mendes [1], F.P. Lobo [1], T.S. Rodrigues [2], G.F. Rodrigues-Luiz [1], W.D. daRocha [3], R.T. Fujiwara [1], S.M.R. Teixeira [4], and D.C. Bartholomeu [*,1]

[1]: Departamento de Parasitologia, Universidade Federal de Minas Gerais, Belo Horizonte, Minas Gerais, Brazil

[2]: Departamento de Computaccao, Centro Federal de Educacao Tecnologica de Minas Gerais, Belo Horizonte, Minas Gerais, Brazil

[3]: Departamento de Bioquımica e Biologia Molecular, Universidade Federal do Parana, Parana, Brazil

[4]: Departamento de Bioquımica e Imunologia, Universidade Federal de Minas Gerais, Belo Horizonte, Minas Gerais, Brazil

*: To whom correspondence should be addressed.

Emails: DCB (daniella@icb.ufmg.br)

Proteins containing repetitive amino acid domains are widespread in all life forms. In parasitic organisms, proteins containing repeats play important roles such as cell adhesion and invasion and immune evasion. Therefore, extracellular and intracellular parasites are expected to be under different selective pressures regarding the repetitive content in their genomes. Here, we investigated whether there is a bias in the repetitive content found in the predicted proteomes of 6 exclusively extracellular and 17 obligate intracellular protozoan parasites, as well as 4 free-living protists. We also attempted to correlate the results with the distinct ecological niches they occupy and with distinct protein functions. We found that intracellular parasites have higher repetitive content in their proteomes than do extracellular parasites and free-living protists. In intracellular parasites, these repetitive proteins are locatedmainly at the parasite surface or are secreted and are enriched in amino acids known to be part of N- and O-glycosylation sites. Furthermore, in intracellular parasites, the developmental stages that are able to invade host cells express a higher proportion of proteins with perfect repeats relative to other life cycle stages, and these proteins have molecular functions associated with cell invasion. In contrast, in extracellular parasites, degenerate repetitive motifs are enriched in proteins that are likely to play roles in evading host immune response. Altogether, our results support the hypothesis that both the ability to invade host cells and to escape the host immune response may have shaped the expansion and maintenance of perfect and degenerate repeats in the genomes of intra- and extracellular parasites.

References

1. Mendes, T. A. O., F. P. Lobo, T. S. Rodrigues, G. Luiz-Rodrigues, W. D. Rocha, R. T. Fujiwara, S. M. R. Teixeira, and D. C. Bartholomeu. "Repeat-enriched proteins are related to host cell invasion and immune evasion in parasitic protozoa." *Molecular biology and evolution* (2013).

Development of large scale machine learning methods for Alzheimer's disease classification using imaging and genetic data

Ho Jang [1], Hyunju Lee [1,*]

[1]: School of Information and Communications, Gwangju Institute of Science and Technology, Gwangju, Republic of Korea.

[*]: Corresponding author.

Emails: HJ (walker83@gist.ac.kr); HL (hyunjulee@gist.ac.kr)

Reliable classification of disease patients is important because the correct identification of disease can increase a chance to cure patients with the proper treatment. Genetic data has been studied to identify disease related genes and pathways and to classify diseases. Especially, genome-wide association studies (GWAS) is useful in the identification of genetic variance that influence on the development of complex diseases. However, the challenge is that the number of SNPs is much larger than samples. To solve these issues, combining MRI imaging data can be useful for improving the statistical powers. But researches to combine these two types of data are still at the early stage.

The purpose of this study is to develop a machine learning algorithm to classify Alzheimer's diseases using a large scale medical data such as imaging and genomic data. Several heterogeneous features were used for machine learning methods including SVM and LASSO to predict patient's status. First, hippocampal volume is highly related to Alzheimer's disease. SNPs associated to proportional volume of hippocampal region were found by association-wide association study and these SNPs were used as features. Second, data from MRI imaging such as hippocampal volume and entorhinal cortex thickness were used as features. Third, demographic data such as age and sex were another type of features. The proposed approach was applied to classify Alzheimer's disease using data from The Alzheimer's Disease Neuroimaging Initiative (ADNI).

"This research was supported by the MKE(The Ministry of Knowledge Economy), Korea and Microsoft Research, under IT/SW Creative research program supervised by the NIPA (National IT Industry Promotion Agency) (NIPA-2012-H0503-12-101)"

A novel link prediction algorithm for reconstructing protein–protein interaction networks by topological similarity

Chengwei Lei
University of Texas at San Antonio
Email: clei@cs.utsa.edu

Motivation: Recent advances in technology have dramatically increased the availability of protein–protein interaction (PPI) data and stimulated the development of many methods for improving the systems level understanding the cell. However, those efforts have been significantly hindered by the high level of noise, sparseness and highly skewed degree distribution of PPI networks. Here, we present a novel algorithm to reduce the noise present in PPI networks. The key idea of our algorithm is that two proteins sharing some higher-order topological similarities, measured by a novel random walk-based procedure, are likely interacting with each other and may belong to the same protein complex.

Results: Applying our algorithm to a yeast PPI network, we found that the edges in the reconstructed network have higher biological relevance than in the original network, assessed by multiple types of information, including gene ontology, gene expression, essentiality, conservation between species and known protein complexes. Comparison with existing methods shows that the network reconstructed by our method has the highest quality. Using two independent graph clustering algorithms, we found that the reconstructed network has resulted in significantly improved prediction accuracy of protein complexes. Furthermore, our method is applicable to PPI networks obtained with different experimental systems, such as affinity purification, yeast two-hybrid (Y2H) and protein-fragment complementation assay (PCA), and evidence shows that the predicted edges are likely bona fide physical interactions. Finally, an application to a human PPI network increased the coverage of the network by at least 100%.

Viscosity of Biomolecular Transport along Wavy-rough Interfaces

Kwang-Hua Chu

School of Mathematics, Physics and Biological Engineering, Inner Mongolia University of Science and Technology, Baotou 014010, PR China

Email: chukha49@gmail.com

There are many types of very rapid flows of biomolecules in nature. One example is the fast transport nature of the protein channel aquaporin-1. Viscosity of fast transport of complex biomolecular fluids along a membrane composed of wavy-rough nanotubes was calculated by using the verified Eyring's transition-state approach (cf. *J. Phys. Chem. B* 112, 3019-3023 (2008)), together with a boundary perturbation approach, which has been successfully adopted to study the selective transport of polymeric matter in confined nanodomains.

Our results can also make the membrane composed of aligned wavy-rough nanotubes a promising mimic of protein channels for transdermal drug delivery and selective chemical sensing. *Acknowledgments.* The only author would like to thank the partial support of 2013--Inner Mongolia University of Science and Technology-- Starting Funds for Scientific Researcher.

Protein Inference and Protein Quantification: Two Sides of the Same Coin

Ting Huang, Peijun Zhu, Zengyou He*

School of Software, Dalian University of Technology, Dalian, China.

*: To whom correspondence should be addressed.

Emails: T.H.(thuang0703@gmail.com); P.Z.(zhupeijun@outlook.com); Z.H.(zyhe@dlut.edu.cn)

In MS-based shotgun proteomics, protein quantification and protein identification are two major computational problems. To quantify the protein abundance, a list of proteins must be firstly inferred. Until now, researchers have been dealing with these two processes separately. Then, one interesting question is if we regard the protein inference problem as a special protein quantification problem, is it possible to achieve better protein inference performance?

In this paper, we investigate the feasibility of using protein quantification methods to solve the protein inference problem. Protein inference is to determine whether each candidate protein is present in the sample or not. Protein quantification is to calculate the abundance of each protein. Naturally, the absent proteins should have zero abundances. Thus, we argue that the protein inference problem can be viewed as a special case of protein quantification problem: present proteins are those proteins with non-zero abundances. Based on this idea, our paper tries to use three simple protein quantification methods to solve the protein inference problem effectively. The experimental results show that these three methods are competitive with previous protein inference algorithms.

A new distributed algorithm for side-chain repacking in protein-protein association

Mohammad Moghadasi [1], Dima Kozakov [2], Fuzhuo Huang [1], Pirooz Vakili [1,3], Sandor Vajda [2], and Ioannis Ch. Paschalidis [1,4,*]

[1]: Center for Information & Systems Eng. and Division of Systems Eng., Boston University.

[2]: Dept. of Biomedical Eng., Boston University.

[3]: Dept. of Mechanical Eng., Boston University.

[4]: Dept. of Electrical & Computer Eng., Boston University.

[*]: To whom correspondence should be addressed.

Emails: MM (mohamad@bu.edu); DK (midas@bu.edu); FH (huangfz@bu.edu); PV (vakili@bu.edu); SV (vajda@bu.edu); IP (yannisp@bu.edu)

Side-chain repacking (SCR) is an important component of computational protein docking methods. Existing SCR methods and available software have been designed for protein folding applications where side-chain positioning is also important. As a result they do not take into account significant special structure that SCR for docking exhibits. We propose a new algorithm which poses SCR as a Maximum Weighted Independent Set (MWIS) problem on an appropriately constructed graph. We develop an approach which solves a relaxation of the MWIS and then rounds the solution to obtain a high-quality feasible solution to the problem. The algorithm is fully distributed and can be executed on a large network of processing nodes requiring only local information and message-passing between neighboring nodes. Motivated by the special structure in docking, we establish optimality guarantees for a certain class of graphs. Our results on a benchmark set of enzyme-inhibitor protein complexes show that our predictions are close to the native structure. We find that the inclusion of the unbound side-chain structures in the set of most probable conformations significantly improves prediction quality. We also establish that the use of our SCR algorithm produces superior docking results.

Prediction of Obligate Protein Interactions Using Short Linear Motifs

Manish Pandit [1], Luis Rueda [1,*]

[1]: School of Computer Science, University of Windsor. 401 Sunset Ave, Windsor, ON, N9B3P4, Canada.

[*]: To whom correspondence should be addressed.

Emails: MP (panditm@uwindsor.ca); LR (lrueda@uwindsor.ca)

Protein-protein interactions (PPIs) play important roles in many biological processes and functions in living cells. Prediction of PPIs has gained much interest in recent years with over 20 different proposed methods. Obligate interactions are usually considered as permanent, while non-obligate interactions can be either permanent or transient. Non-obligate and transient interactions are more difficult to study and understand due to their instability and short life, while obligate and permanent interactions last for a longer period of time, and hence are more stable. The most successful approaches for prediction of PPIs use mainly structural information of protein complexes as descriptors or features. These models, however, depend on structural information, which is available for fewer complexes (approximately 80,000) in the Protein Data Bank, compared to millions of protein sequences available in other databases. On the other hand, motifs are patterns widespread over a group of related protein sequences usually having strong biological relationships. In particular, we focus on a special case of motifs, 3-10 amino acids long, and which are called short, linear motifs (SLiMs) or minimotifs.

We propose a model that uses SLiMs to predict obligate and non-obligate protein-protein interaction types. To find SLiMs, we use Multiple EM for Motif Elicitation (MEME) [1]. For classification, we use these SLiMs and apply a leave-one-out validation approach in which the sequences of the complex to be classified (target sequences) are partitioned in a set of overlapping *l*-mers of different lengths (3-10 amino acids). The *l*-mers are then ranked based on information content given all SLiMs in the training dataset, and the top 20 ranked scores are fed into a feature vector used for classification. We have tested a few state-of-the-art classifiers, including *k*-nearest neighbor, linear dimensionality reduction and support vector machines. We demonstrate the power of SLiMs as predictive properties for obligate and non-obligate complexes. Our prediction results on well-known datasets, namely the Zhu et al. [3] and Mintseris et al. [2] datasets show an impressive accuracy in prediction of more than 99%, which imply a significant increase from previous approaches, even better than structure-based methods, while using only sequence information.

References

1. Timothy Bailey *et al.* MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Research*, 37:W202-W208 (2009).
2. Julian Mintseris, Zhiping Weng. Structure, function, and evolution of transient and obligate protein--protein interactions. *Proceedings of the National Academy of Sciences*, 102(31), 10930 (2005).
3. Hongbo Zhu *et al.* NOXclass: prediction of protein-protein interaction types. *BMC Bioinformatics*, 7(1), 27 (2006).

Mining Abnormal Groups in Biological Data

Yun Xiong [1], Yangyong Zhu [1], Jian Pei [2] , Philip S. Yu [3,*]

[1]: Research Center for Dataology and DataScience, Fudan University, China.

[2]: School of Computing Science, Simon Fraser University, Canada.

[3]: Department of Computer Science, University of Illinois at Chicago, USA.

[*]: To whom correspondence.

Emails: YX (yunx@fudan.edu.cn); YYZ (yyzhu@fudan.edu.cn); JP (jpei@cs.sfu.ca) ; PSY (psyu@uic.edu)

With the increasing of large-scale biological data, in some biological applications, such as biological functional prediction, protein complexes prediction and transcriptional regulation mechanism analysis, it is desirable to find compact clusters formed by a small portion of objects in a large biological data set. In general it is still a clustering problem, but it cannot be served well by the conventional clustering methods since most of those methods try to assign most of the data objects into clusters and always suffer from a large number of false positives. The objects of such clusters are minority in population; therefore, these clusters are regarded as abnormal in the whole data set. In this paper, we model this novel and application-inspired task as the problem of mining abnormal groups in biological data, which is a non-trivial variation of clustering. We propose a general framework and a principled approach to tackle the problem. Unlike conventional clustering methods, our algorithm can mine abnormal groups flexibly without the need for a predefined similarity threshold. The experimental results on real biological data sets verify the effectiveness and efficiency of our proposed approach.

References

1. V. Chandola, A. Banerjee, V. Kumar. Anomaly Detection: A Survey. *ACM Computing Surveys*, 41(3): 1-58, 2009.
2. A. Corral, Y. Manolopoulos, Y. Theodoridis, and *et al*. Algorithms for Processing K-closest-pair Queries in Spatial Databases. *Data & Knowledge Engineering*, 67-104, 2004.
3. M. Dettling, P. Buhlmann. Supervised Clustering of Genes. *Genome Biology*. 3(12), 2002.
4. G. Gupta, J. Ghosh. Bregman Bubble Clustering: A Robust Framework for Mining Dense Clusters. *ACM Transactions on Knowledge Discovery from Data*, 2(2): 1-49, 2008.
5. A. K. Jain. Data clustering: 50 Years Beyond K-Means. *Pattern Recognition Letters*, 31(8): 651-666, 2010.
6. Y. Xiong, Y. Y. Zhu. Mining Peculiarity Groups in Day-by-Day Behavioral Datasets. In *Proc. of ICDM*, 578-587, 2009.
7. G. Y. Zheng, K. Tu, Q. Yang, and *et al*. ITFP: an Integrated Platform of Mammalian Transcription Factors. *Bioinformatics*, 24(20): 2416-2417, 2008.

Prediction of tyrosine phosphatase substrates based on sequence features

Zheng Wu [1], Tingting Li [1,]*

[1]: Department of Biomedical Informatics, School of Basic Medical Sciences, Peking University Health Science Center, Beijing 100191, China.

[*]: To whom correspondence should be addressed.

Email: litt@hsc.pku.edu.cn

Tel:+86 10 8280 1585; Fax:+86 10 8280 1001

Abstract

Tyrosine phosphorylation plays crucial roles in numerous physiological processes. The level of phosphorylation state depends on the combined action of protein tyrosine kinases (PTKs) and protein tyrosine phosphatases (PTPs). Detection of possible substrate sites and corresponding enzymes can provide useful information to the functional study of relevant proteins. There have been several studies focused on the identification of PTK substrates, most of which predict phosphorylation sites based on primary sequences around the phosphorylation sites. However, compared to PTKs, prediction of PTP substrates involved in the balance of protein phosphorylation level falls behind. In this study, we took advantage of the k-nearest neighbor (k-NN) algorithm to predict putative substrates of three PTPs, which are PTP1B, PTP1C and PTP1D. Based on manually collected dephosphorylation sites of these three PTPs, we first analyzed the characterization of PTP1B, PTP1C and PTP1D substrates, and then developed a method to predict relevant dephosphorylation proteins or sites. Next, we utilized tests to evaluate the performance of this method, and got results which verified the effectiveness of prediction. Finally, we applied the method on a set of known tyrosine phosphorylation sites to search for candidate substrates of the above-mentioned PTPs.

Key words: *tyrosine phosphorylation, dephosphorylation, protein tyrosine phosphatases (PTPs), prediction*

A novel hierarchical gene clustering method

Dan Wei[1,2], Qingshan jiang[2,*], Yanjie Wei[2,*] Shengrui Wang[3]

[1]:Cognitive Science Department & Fujian Key Laboratory of the Brain-like Intelligent Systems, Xiamen University, Xiamen, China
[2]: Shenzhen Key Lab for High Performance Data Mining, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China.
[3]: Department of Computer Sciences, University of Sherbrooke, Sherbrooke, QC, Canada.
[*]: To whom correspondence should be addressed.

Emails: DW (dan.wei@siat.ac.cn); QJ(qs.jiang@siat.ac.cn); YW (yj.wei@siat.ac.cn); SW(shengrui.wang@usherbrooke.ca)

In bioinformatics and computational biology, clustering gene sequences into functional groups plays an important role for the discovery of novel gene structure and function, and the understanding of evolutionary relationships between genes. In this paper, we propose a modified bisecting K-means clustering algorithm, mBKM, based on a new alignment-free distance measure, DMk, for clustering gene sequences. DMk takes into account the occurrence, location and order relation of k-tuples within a gene sequence and extract numeric features from sequence. mBKM chooses the initial centroids by the maximum and minimum principle and selects the cluster to split based on the compactness of clusters. It can produce either a hierarchical clustering or a partition clustering result. The proposed method are evaluated by clustering functionally related genes and by phylogenetic analysis. DMk shows better performance than the k-tuple distance, which considers only the k-tuples frequencies. mBKM outperforms three hierarchical clustering methods including single-linkage clustering, complete-linkage clustering, average-linkage clustering, and the partitioning approaches such as K-means and bisecting K-means when tested on public gene sequence datasets. Furthermore, the proposed method, mBKM with DMk, also outperforms alignment-based methods such as BlastClust and CD-HIT-EST when clustering functionally related genes, and it performs better than alignment-based methods including UPGMA with CLUSTALW and ML with CLUSTALW when building the phylogenetic trees for β-globin genes of 10 species and 60 H1N1 viruses.

References

1. Dan Wei, Qingshang Jiang, Yanjie Wei, Shengrui Wang. A novel hierarchical clustering algorithm for gene sequences. BMC Bioinformatics, 13: 174, 2012.

Ligand Binding Site Prediction Using Ligand Interacting and Binding Site-Enriched Protein Triangles

Zhong-Ru Xie [1] and Ming-Jing Hwang [1]*

[1]: Institute of Biomedical Sciences, Academia Sinica, Taipei 115, Taiwan

*: To whom correspondence should be addressed.

Email address: amianxie@gmail.com (ZRX), mjhwang@ibms.sinica.edu.tw (MJH)

Knowledge about the site at which a ligand binds provides an important clue for predicting the function of a protein and is also often a prerequisite for performing docking computations in virtual drug design and screening. We have previously shown that certain ligand-interacting triangles of protein atoms, called protein triangles, tend to occur more frequently at ligand-binding sites than at other parts of the protein. In this work, we describe a new ligand-binding site prediction method that was developed based on binding site-enriched protein triangles. The new method was tested on two benchmark datasets and on 19 targets from two recent community-based studies of such predictions, and excellent results were obtained. Where comparisons were made, the success rates for the new method for the first predicted site were significantly better than methods that are not a meta-predictor. Further examination showed that, for most of the unsuccessful predictions, the pocket of the ligand-binding site was identified, but not the site itself, whereas for some others, the failure was not due to the method itself but due to the use of an incorrect biological unit in the structure examined, although using correct biological units would not necessarily improve the prediction success rates. These results suggest that the new method is a valuable new addition to a suite of existing structure-based bioinformatics tools for studies of molecular recognition and related functions of proteins in post-genomics research.

References

Xie, Z.R. and Hwang, M.J. Ligand Binding Site Prediction Using Ligand Interacting and Binding Site-Enriched Protein Triangles. *Bioinformatics*, 28, 1579-1585, 2012.

Super-resolution imaging of protein localization to detect polarity establishment in budding yeast

Pengyun Lv[1], Zhen Zhao[2], Michael Q. Zhang[2]*, Juntao Gao[2]*

1: Center for Synthetic & Systems Biology, Tsinghua University, Beijing 100084, China
2: MOE Key Laboratory of Bioinformatics; Bioinformatics Division and Center for Synthetic & Systems Biology, TNLIST, Department of Automation, Tsinghua University, Beijing 100084, China
*: To whom correspondence should be addressed. E-mail: jtgao@biomed.tsinghua.edu.cn, Mzhang@cshl.edu
Emails: Pengyun Lv (pengyunlv@mail.tsinghua.edu.cn) ; Zhen Zhao （zhaozhen205@gmail.com）; Michael Q. Zhang (Mzhang@cshl.edu); Juntao Gao (jtgao@biomed.tsinghua.edu.cn)
Keywords: photoconvertible protein, PALM/STORM, super-resolution microscope, mEos3.1/3.2, budding yeast

Cell polarity is universal phenomena and conserved from budding yeast to human being. However, how is cell polarity formed and established, is still unknown. With the advance of super-resolution imaging techniques such as PALM/STORM and STED, cellular ultrastructure observation reached an unprecedented level at the resolution of ~10nm. We plan to study the mechanism of cell polarity formation using super-resolution imaging and mathematical model.

We first asked if super-resolution imaging could be applied in budding yeast. We tested three proteins Utp13, Rpb7 and Fba1 at first. Photoconvertible protein mEos3.1 and mEos3.2 derived from Eos FP recently have been considered to be the best for PALM super-resolution imaging. GFP site of plasmid pFA6a-GFP(S65T)-His3MX6 was replaced with mEos3.1 or mEos 3.2 (mEos3.1/3.2) ORF using molecular biology techniques. Then with homologous recombination, mEos3.1/3.2 was inserted into the downstream of Utp13, Rpb7 and Fba1, respectively. The colonies of Utp13-mEos3 and Rpb7-mEos3 that could grow on His Minus Medium have been confirmed using laser scan confocal microscope. Among three tested proteins, Utp13-mEos3 and Rpb7-mEos3 recombinant strains gave low emission light both before and after photoconversion, and could be easily bleached during the excitation. However, Fba1-mEos3 exhibited high emission light both in the green and red state and can't be easily bleached. We concluded that protein abundance is an important factor to label specific proteins with mEos3 for super-resolution imaging in budding yeast.

With this technique in hand, we are inserting mEos3.1/3.2 downstream of key polarity proteins such as Cdc42, Rho1, Rho3, LifeAct and others to get super-resolution localization of these key proteins in a systematic way. Based on these info and others, we will set up the math model to describe how polarity is formed (and maintained) in budding yeast.

Microarray Inspector: simple tissue mixtures detection software for raw microarray data

Piotr Stępniak[1,*], Matthew Maycock[1], Konrad Wojdan[1,2], Monika Nesteruk[3], Serhiy Perun[1,4], Aashish Srivastava[5], Lucjan S. Wyrwicz[5] and Konrad Świrski[1,2]

[1]: Transition Technologies S.A., Pawia 55, 01-030 Warsaw, Poland.
[2]: Institute of Heat Engineering, Warsaw University of Technology, Nowowiejska 21/25 00-665 Warszawa, Poland.
[3]: Department of Gastroenterology and Hepatology, Medical Center for Postgraduate Education, Marymoncka 99/103, 01-813 Warsaw, Poland.
[4]: Institute of Physics PAS, Al. Lotników 32/46, 02-668 Warsaw, Poland.
[5]: Laboratory of Bioinformatics and Biostatistics, Maria Sklodowska-Curie Memorial Cancer Center and Institute of Oncology, WK Roentgena 5, 02-781 Warsaw, Poland.
* To whom correspondence should be addressed.

Emails: PS (P.Stepniak@tt.com.pl); MM (mhm2159@columbia.edu), KW (K.Wojdan@tt.com.pl), MN (monika.nesteruk@gmail.com), SP (S.Perun@tt.com.pl), AS (aashish.srivastava1302@gmail.com), LSW (lucjanwyrwicz@gmail.com), KŚ (K.Swirski@tt.com.pl)

Microarray data repositories of experiment results with various conditions and samples serve the scientific community as a precious resource for further studies as reference and comparison. Even though data bases are more often manually curated it is still important to validate sample quality, specifically, that in many cases, datasets lack information concerning pre-experimental quality assessment. The risk of tissue cross contamination is especially high in oncological studies, where it is often difficult to extract the sample. Moreover there are usually no technical replicates. Identifying data coming from mixed tissues before including it in further studies has large impact on quality of future results. Another difficulty in selecting data for analysis is screening large repositories for samples representing desired minimum expression of certain genes. We present MicroArray Inspector: a customizable, user-friendly software system that enables easy screening of microarray data for samples containing mixed tissue types, and representing other desired expression patterns. The algorithm uses raw expression data files and analyzes each array independently. We also present several examples of examination of data from public repositories. MicroArray Inspector is available for many platforms and is provided free of charge for nonprofit research institutions.

Translational Systems Biology: understanding the limits of animal models as predictors of human biology

C. Poussin[1], L.G. Alexopoulos[2], V. Belcastro[1], E. Bilhal[3], C. Mathis[1], P. Meyer[3], R. Norel[3], Y.Xiang[1,*], J.J. Rice[3], G. Stolovitzky[3], J.Hoeng[1], M. C. Peitsch[1]

[1]: Philip Morris Research and Development, CH-2000 Neuchâtel, Switzerland

[2]: Protatonce Ltd, Glyfada 16675, Greece

[3]: IBM Thomas J. Watson Research Center, NY 10598, USA

[*]: To whom correspondence should be addressed.

Emails: YX (Yang.Xiang@pmi.com)

Inferring how humans respond to external cues such as drugs, chemicals, viruses or hormones is an essential question in biomedicine. Very often, however, this question cannot be addressed due to the impossibility to perform experiments in humans. A reasonable alternative consists of generating responses in animal models and "translating" the results to humans. The limitations of such translation, however, are far from clear, and systematic assessments of its actual potential are badly needed.

We have designed a series of challenges in the context of the 'sbv IMPROVER' project (Industrial Methodology for Process Verification in Research; http://sbvimprover.com/) to address the issue of translatability between humans and rodents. Our main aim is to understand the limits and opportunities of species to species translatability at different levels of biological organization: signalling, transcriptional, and release of secreted factors (such as cytokines, chemokines or growth factors).

The sbv IMPROVER project are part of a collaborative project designed to enable scientists to learn about and contribute to the development of a new crowd sourcing method for verification of scientific data and results.

Hierarchical Clustering and Similarity Maps for Annotating FT-IR Spectral Images

Qiaoyong Zhong [1,2], Chen Yang [1], Frederik Großerüschkamp [2], Angela Kallenbach [2], Peter Serocka [1,2], Klaus Gerwert [1,2], Axel Mosig [2,*]

[1]: Department of Biophysics, CAS-MPG Partner Institute and Key Laboratory for Computational Biology, Shanghai Institutes for Biological Sciences, Shanghai, China.

[2]: Department of Biophysics, Ruhr University Bochum, Bochum, Germany.

[*]: To whom correspondence should be addressed.

Emails: Qiaoyong Zhong (zhongqiaoyong@picb.ac.cn); Axel Mosig (axel.mosig@bph.rub.de)

Unsupervised segmentation of multi-spectral images plays an important role in annotating infrared microscopic images. In this context, we introduce two contributions. First, we provide a validation scheme for hierarchical clustering, which is applied to clustering schemes commonly used in infrared microscopy and compared to other validation schemes. Second, we introduce the application of so-called *interactive similarity maps* to infrared microscopy.

The validation scheme introduced in the first part allows a quantitative comparison between different clustering approaches for image annotation, in contrast to previous merely qualitative comparisons. Based on so-called tree assignments, our validation scheme shows that performance of hierarchical two-means is comparable to the traditionally used Ward's clustering. As the former is much more efficient in time and memory, our results suggest that it is a viable alternative, in particular for handling large spectral images. Furthermore, by comparing our tree-assignment approach with two previous tree-segmenting approaches, we demonstrate that tree assignment performs best in terms of classification accuracy.

Finally, we introduce and validate the concept of *similarity maps* as a novel interactive approach implemented in our *Lasagne* software for infrared image segmentation. Using our validation scheme, we demonstrate that similarity maps are capable of producing more accurate segmentations than commonly used hierarchical clustering.

Implementation of efficient haplotype matching using suffix array based methods

Tomislav Ilicic [1,2*], Richard Durbin [1]

[1]: Wellcome Trust Sanger Institute

[2]: University of Cambridge

[*]: To whom correspondence should be addressed.

Emails: TI (ti1@sanger.ac.uk); RD (rd@sanger.ac.uk);

Haplotype phasing and imputation of genotypes are important in a variety of genetic data analyses. Current phasing and imputation algorithms using probabilistic hidden Markov models are accurate but computationally demanding, making them insufficient for large data sets. With the tremendous increase of sequencing data it becomes crucial to be able to handle data sets containing hundreds of thousands of samples.

Here we describe an implementation of a haplotype matching algorithm based on suffix arrays, widely used for DNA sequence read matching and assembly, to build a foundation for fast haplotype phasing and imputation, with the aim to scale to much larger data sets than those currently handled by genotype algorithms. We show that given M sequences and N bi-allelic variable sites our algorithm can derive a representative data structure based on positional prefix arrays in O(NM). Using this representation we can find maximal matches between a new sequence and the set in O (N), i.e. independent of the number of sequences.

Noise-Resistant Bicluster Recognition

Huan Sun[1,*], Gengxin Miao[2], Yu S. Huang[3], Xifeng Yan[1]

[1]: University of California, Santa Barbara, USA.

[2]: Google, Mountain View, USA.

[3]: Center for Neurobehavioral Genetics, Semel Institute, UCLA, USA.

[*]: To whom correspondence should be addressed.

Emails: H. S. (huansun@cs.ucsb.edu); G. M. (budaoweng@gmail.com); Y. S. H. (polyactis@gmail.com); X. Y. (xyan@cs.ucsb.edu)

Biclustering is an important tool for analyzing gene expression data. By simultaneously grouping genes and conditions, it can reveal a group of genes that are regulated under a subset of conditions. Over the past decade, various biclustering algorithms have been developed and improved constantly afterwards. In this paper, we discuss the common issues existing in these algorithms and propose a noise-resistant neural network model, named *AutoDecoder* (AD), to find biclusters from expression data. To suppress severe noises present in gene expression data, we introduce a non-uniform signal recovery principle: Instead of reconstructing the whole input data to capture the bicluster patterns, AD weighs the bicluster and non-bicluster parts of the input data differently. A fast numerical solution is developed in AD to optimize the objective function corresponding to this principle. We compared the results of our approach on real-life expression datasets with four state-of-the-art biclustering algorithms. In three out of the four datasets, our approach significantly outperforms the others. For controlled synthetic datasets, AD performs the best when the noise ratio is beyond 15%.

References

1. C. Huttenhower, K.T. Mutungu, N. Indik, W. Yang, M. Schroeder, J.J. Forman, O.G. Troyanskaya, and H.A. Coller. Detailing regulatory networks through large scale data integration. *Bioinformatics*, 25(24):3267–3274, 2009.
2. G. Li, Q. Ma, H. Tang, A.H. Paterson, and Y. Xu. Qubic: a qualitative biclustering algorithm for analyses of gene expression data. *Nucleic acids research*, 37(15):e101–e101, 2009.
3. S. Hochreiter, U. Bodenhofer, M. Heusel, A. Mayr, A. Mitterecker, A. Kasim, T. Khamiakova, S. Van Sanden, D. Lin, W. Talloen, et al. Fabia: factor analysis for bicluster acquisition. *Bioinformatics*, 26(12):1520–1527, 2010.
4. M. Sill, S. Kaiser, A. Benner, and A. Kopp-Schneider. Robust biclustering by sparse singular value decomposition incorporating stability selection. *Bioinformatics*, 27(15):2089–2097, 2011.
5. G.E. Hinton and R.R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504, 2006.

Identification of Cell Compartments by Label-free Raman Imaging

Sascha Krauß [1], Dennis Petersen [1], Inka Fricke [1], Samir F. El-Mashtoly [1], Klaus Gerwert [1], Axel Mosig [1*]

[1]: Department of Biophysics, Ruhr University Bochum, Bochum, Germany.

[*]: To whom correspondence should be addressed.

Emails: SK (sascha.krauss@bph.rub.de); DP (dennis-petersen@bph.rub.de); IF (inka.fricke@bph.rub.de); SE (samir.elmashtoly@bph.rub.de); KG (klaus.gerwert@bph.rub.de); AM (axel.mosig@bph.rub.de)

A major promise of Raman microscopy is the label-free detailed recognition of cellular and subcellular structures. To this end, identifying colocalization patterns between Raman spectral images and fluorescence microscopic images is a key step to “annotate” subcellular components in Raman spectroscopic images.

While existing approaches to resolve subcellular structures are based on fluorescent labeling, we propose a combination of a colocalization scheme with subsequent training of a supervised classifier that allows label-free resolution of cellular compartments. Our colocalization scheme unveils statistically significant overlapping regions by identifying correlation between the fluorescent color channels and clusters from unsupervised machine learning methods like hierarchical cluster analysis. The colocalization scheme is used as a preselection to gather appropriate spectra as training data. These spectra are used in the second part as training data to establish a supervised Random Forest classifier. We demonstrate that compared to a recent mutual-information based approach, our method is robust against inevitable inaccuracies in overlaying the spectral image with the fluorescent image. We validate our approach by examining Raman spectral images overlayed with fluorescent labelings of different cellular compartments, indicating that many components may indeed be identified label-free in the spectral image.

Gradients of replication-associated mutational asymmetry drive the evolution of human genome composition

Chun-Long Chen [1,*], Benjamin Audit [2], Yves d'Aubenton-Carafa [1], Olivier Hyrien [3], Alain Arneodo [2] and Claude Thermes [1,*]

[1]: Centre de Génétique Moléculaire, CNRS UPR 3404, 91198 Gif-sur-Yvette, France.

[2]: Laboratoire de Physique, ENS Lyon, CNRS, 69364 Lyon, France.

[3]: ENS Paris, CNRS UMR 8197, INSERM U1024, 75005 Paris, France.

[*]: To whom correspondence should be addressed.

Emails: C.L.C. (chen@cgm.cnrs-gif.fr); B.A. (benjamin.audit@ens-lyon.fr); Y.D.C. (daubenton@cgm.cnrs-gif.fr); O.H. (hyrien@biologie.ens.fr); A.A. (alain.arneodo@ens-lyon.fr); C.T. (thermes@cgm.cnrs-gif.fr)

Human genome studies have shown that replication induces different mutation rates on the leading and lagging strands (Chen et al. MBE 2011). This generates during evolution strong asymmetries of nucleotide composition. Analysis of this asymmetry, S=(G-C)/(G+C)+(T-A)/(T+A), have revealed large ~1Mb domains exhibiting a characteristic N-shaped pattern covering more than 1/3 of the genome (Huvet et al. Genome Res. 2007). We showed that these mutational asymmetries decrease from maximum values at left ends of N-domains to zero at centers, and to opposed values at right ends, generating over evolutionary times the N-shaped pattern. This indicates a progressive inversion in replication fork polarity from one end to the other. We propose that replication first initiates at N-domain extremities and secondary origins fire coordinately from borders to centers mediated by gradients of open chromatin conformation. Computational simulations of this model generate linear gradients of replication fork polarity and N-shaped skew profile. N-domains are observed in all studied mammals, birds and reptiles but not in amphibians and fishes. It seems that this replication program has been conserved since amniota divergence. This indicates that the specific spatio-temporal replication program associated with gradients of chromatin structure is a major determinant of genome evolution.

References

1. CHEN C.L., Duquenne L., Audit B., Guilbaud G., Rappailles A., Baker A., Huvet M., d'Aubenton-Carafa Y., Hyrien O., Arneodo A. and Thermes C. Replication-associated mutational asymmetry in the human genome. *Mol. Bio. Evol.*, 28(8): 2327-2337, 2011.
2. Huvet M., Nicolay S., Touchon M., Audit B., d'Aubenton-Carafa Y., Arneodo A. and Thermes C. Human gene organization driven by the coordination of replication and transcription. *Genome Res.*, 17(9): 1278-1285, 2007

TxT: A Tool for Reconciliation of Non-binary Trees

Yu Zheng[1], Louxin Zhang[1,2,*]

[1]: Department of Mathematics, National University of Singapore (NUS), Singapore 119076
[2]: NUS Graduate School for Integrative Sciences and Engineering, NUS, Singapore 117456
[*]: To whom correspondence should be addressed.

Emails: YZ (yu_zheng@nus.edu.sg); LXZ (matzlx@nus.edu.sg)

The gene tree of a gene family is sometimes discordant with the corresponding species tree due to gene duplication, horizontal gene transfer, or hybridization. Hence, gene tree and species tree reconciliation is used for inferring gene duplication histories and for reconstructing phylogenetic trees. In the past decade, the tree reconciliation problem has been intensively investigated for binary gene and species trees. Motivated by the fact that reference species trees and real gene trees are often non-binary, we investigate the reconciliation of two non-binary trees in a binary refinement model. A binary tree is a binary refinement of a non-binary tree if every cluster of the latter is found in the former. Under the binary refinement framework, the reconciliation problem is, given a species tree S, a set of gene trees G_i ($1 \le i \le k$) and a reconciliation cost $c(\ ,\ )$, to find a binary refinement S' of S and binary refinements G_i' of G_i that maximizes $\sum c(G_i', S')$. Clearly, the reconciliation problem is a natural generalization of the standard reconciliation problem for binary trees and also includes the species tree reconstruction problem as a special case. The reconciliation problem is NP-hard for non-binary species trees. Nevertheless, we develop different exact and heuristic methods for reconciling two non-binary trees by using a novel data structure. The proposed methods have been implemented in Golang and tested using both random and real datasets. Our tool named TxT is ready to serve the bioinformatics community for gene duplication inference, gene orthology identification and phylogeny reconstruction.

SCANER: Sequential Chaining and Analysis of New Elementary Repeats

Nathan Figueroa [1], John Karro [1,2,3*]

[1]: Miami University, Departments of Computer Science, [2]Microbiology, and [3]Statistics

[*]: To whom correspondence should be addressed.

Emails: N.F. (figuernd@miamioh.edu); J.K. (karroje@miamioh.edu)

In this poster we present a computational method for the discovery of genomic repeats without a need for reference data. De novo repeat discovery is an important problem, made increasingly difficult by the volume of new sequence data. Library-based programs such as RepeatMasker[1] effectively expand known families of repeats, but discovering new families within a genome is difficult when dealing with inexact copies. Tools relying on self-alignment (e.g RECON[2]), become prohibitively time-consuming with large sequences, while text-indexing methods, such as the Suffix Array or FM Index, are poorly suited for the wildcard searches needed to account for single base mismatches. We present a tool that uses spaced seeds in the spirit of PatternHunter[3] to identify inexact repeats with wildcard matching in linear time, followed by the decomposition of the identified repeats into maximal repeat elements and the mapping of the mosaic structures formed by these elements. SCANER's speed allows extensive parameter tuning, processing Human Chromosome 22 in approximately 8 minutes (as compared to the multi-hour runs for RepeatModeler[4] or RECON). SCANER also shows great promise in terms of sensitivity, with initial testing resulting in a five-fold increase in the number of bases identified as compared to RECON.

References

1. Arian Smit, Robert Hubley. RepeatMasker Open-3.0. *Institute for Systems Biology*, 1996-2010. Accessed February 26, 2013, http://www.repeatmasker.org/
2. Zhirong Bao, Sean R. Eddy. Automated de novo identification of repeat sequence families in sequenced genomes. *Genome Research*, 12(8): 1269-1276, 2002.
3. Bin Ma, John Tromp, Ming Li. PatternHunter: faster and more sensitive homology search. *Bioinformatics*, 18(3) 440-445, 2002.
4. Arian Smit, Robert Hubley. RepeatModeler Open-1.0. *Institute for Systems Biology*, 1996-2010. Accessed February 26, 2013, http://www.repeatmasker.org/RepeatModeler.html

A Generic Data Decomposition Tool for Parallel Environments

Ahmad Salah [1,2,3], Kenli Li [1,3,*]

[1]: College of Information Science and Engineering, Hunan University, Changsha, Hunan, China.

[2]: Computer Science Dept., College of Computers and Informatics, Zagazig University, Zagazig, Egypt.

[3]: The National Supercomputing Center in Changsha, Hunan University, Changsha, Hunan, China.

[*]: To whom correspondence should be addressed.

Emails: A.S. (ahmad@hnu.edu.cn); K.L. (lkl@hnu.edu.cn);

The steady increase of the biological data encourages computer scientists to develop data-analytical methods in order to study the biological systems. Most of these methods are firstly presented in the sequential form and then it is converted to a parallel version due to the constant expanding of data size. The sequential method is a data or functional decomposition. In this work, we focus on the category of the data decomposition methods, which means several similar tasks handle different data. Protein structure comparison is an example of the problems that can be handled by the proposed tool. The tool can be used to decompose the data into n portions, according to the user's needs. Secondly, it starts the sequential method n times, locally or remotely. Finally, collect and merge the results into the final form. Using this tool, the time gap, between presenting the sequential and the parallel versions, is eliminated. The proposed tool is a generic, regarding the operating system and the level of parallelism, from multi-core standalone PC to a grid of nodes. The experimental results show a linear speed up, which is close to optimal, without any accuracy loose.

References

1. Xiangyuan Zhu, Kenli Li, Ahmad Salah. A data parallel strategy for aligning multiple biological sequences on multi-core computers. *Computers in Biology and Medicine,* in press, 2013.

Detecting pico-inversions in primate genomes based on multi-species alignment

Minmei Hou [1,*]

[1]: Department of Computer Science, Northern Illinois University, DeKalb, IL USA.

[*]: To whom correspondence should be addressed.

Emails: M. Hou (mhou@cs.niu.edu)

Inversion is a type of genomic mutation where a piece of DNA is replaced by its reverse complement. Characterization of inversions has been useful in many aspects of biomedical research. In the past, inversions between different species are mostly discovered by local alignment. Therefore, the size of detectable inversions is determined by the significance threshold of local alignment. The study of very small inversions has been very limited due to computational restrictions. We previously implemented a probabilistic approach to detect very small inversions (which we call *pico-inversions*) between a pair of very closely related species and discovered thousands of such inversions between human and chimpanzee [1]. However, the accuracy of this approach decreases when the divergence of the pair of sequences increases.

In this study, we use multi-species alignment to detect pico-inversions among multiple primate genomes. The sequence information from multiple species can help to improve the reliability of detecting pico-inversions. But the conflicting inversion relationship among multiple species also complicates the detection and verification of inversions. In the framework we developed, potential pico-inversions are first obtained for each pair of species. The inversion relationship among multiple species is then reconciled based on the species tree. Potential inversions that cannot be reconciled are determined to be false positive. In the process of the reconciliation, the inversions are also located to branches of the species tree. For each genomic region (of multiple species) with a potential inversion, we reconstruct two sets of ancestral sequences with the maximum probabilities assuming there is no inversion (null hypothesis) and there is an inversion (alternative hypothesis) respectively. We then use Bayes factor to test the hypothesis based on the phylogenetic information [2]. The ones with significant evidence are reported as true pico-inversions.

We applied this framework to detect pico-inversions among human, chimpanzee, gorilla, and orangutan. Hundreds of pico-inversions are found in each lineage and ancestor, most of which were not reported before. The computation pipeline can be applied to more species. However, simulation shows that this framework has high specificity and relatively low sensitivity. We still work on to improve the sensitivity of the pipeline.

References

1. Minmei Hou, Ping Yao, Angela Antonou, Mitrick A. Johns. Pico-inplace-inversions between human and chimpanzee. *Bioinformatics*, 27(23):3266-3275, 2011.
2. Minmei Hou, Ping Yao, Mitrick A. Johns. Computation verification of a potential pico-inversion with multi-species comparison. To be appeared in *the Proceedings of BiCOB-2013*.

A novel parallel algorithm of biclustering based on the association rules

Yun Xue [1,*], Tiechen Li [1], Xiaohui Hu [1]

[1]:School of Physics and Telecommunication Engineering, South China Normal University, Guangzhou 510006,China

Emails: Y. X. (xueyun@scnu.edu.cn); T. C. L. (ltch2013@gmail.com); X. H. H (huxh@scnu.edu.cn)

Because of the ability of simultaneously capturing correlations among subsets of attributes (columns) and records (rows), biclustering is widely used in data mining applications such as biological data analysis, financial forecasting, and text mining, etc.

However, the biclusters with coherent evolutions generally require a strictly monotonic variation. In this paper, we proposed a new non-strictly monotonic variation mode (i.e. a mixture of equal and ascending modes), which expands the scope of bicluster patterns.

Furthermore, since biclustering is known to be a NP-hard problem, biclusters are identified through heuristic approaches in most algorithms whose results are non-deterministic. A new algorithm based on the association rules is proposed in this paper which is deterministic and enables exhaustive discovery of coherent evolution biclusters.

Finally, the algorithm is parallelized to save the time for running. The experimental results show that the improved parallel algorithm achieves nearly linear speedups and has a better extension.

Correction of Fluorophore Crosstalk in second generation sequencer(SGS)

Musheng Li [1], Xueying Xie [2], Zuhong Lu [1,*]

[1,2,3,*] :State Key Laboratory of Bioelectronics, School of Biological Science and Medical Engineering, Southeast University

Emails: Musheng Li (truth1988@gmail.com); Xueying Xie (xxying.rcls@seu.edu.cn); Zuhong Lu (zhlu@seu.edu.cn)

High-throughput sequencing technologies have noticeably reduced the cost of DNA sequencing, which has been proved central to the study of molecular Biology, and significantly increase the 'read' production rate. that is ,these platforms execute billions of simultaneous sequencing reactions, produce tons of digital images captured the information of nuclear acid probes .and use a so called 'base calling' pipeline to translate these raw data to 'reads"—strings of A,C,G, or T's .Even so,there still are some challenges associated with reducing the error rate, For example these platforms suffer from similar signal distortion factors: Fluorophore Crosstalk, which results from the overlap of the emission spectra of the fluorophores. A 4*4 matrix G can be established to describe this overlap .Here , we present a improved model for the determination of matrix G. In this model we use the density of plots in the 4-dimension fluorescence intensity space to determine the elements of matrix G. this model was applied on AG100,a second generation high-throughput sequencer which use cyclic sequencing-by-ligation chemistry reactions. This model proved itself faster and robust .

References

1. Lei Li, Terence P. Speed, An estimate of the crosstalk matrix in four-dye fluorescence-based DNA sequencing *Electrophoresis* , 20, 1433–1442,1999.
2. Whiteford N, Skelly T, Curtis C, et al. Swift: primary data analysis for the Illumina Solexa sequencing platform[J]. *Bioinformatics* , 25(17): 2194-2199,2009.

Novel prostate cancer specific transcripts identified using RNA-seq

Antti Ylipää[1,*], Kati K Waltering[1], Matti J Annala[1], Kimmo Kartasalo[1], Leena Latonen[2], Simo-Pekka Leppänen[1], Mauro Scaravilli[2], Wei Zhang[3], Tapio Visakorpi[2], Matti Nykter[2,*]

[1]: Tampere University of Technology, Tampere, Finland.

[2]: University of Tampere, Tampere, Finland.

[3]: The University of Texas, MD Anderson Cancer Center, Houston, Texas, USA.

[*]: To whom correspondence should be addressed.

Emails: AY (antti.ylipaa@tut.fi); KKW (kati.waltering@tut.fi); MJA (matti.annala@tut.fi); KK (kimmo.kartasalo@tut.fi); LL (leena.latonen@uta.fi); S-PL (simo-pekka.leppanen@tut.fi); MS (mauro.scaravilli@uta.fi); WZ (wzhang@mdanderson.org); TV (tapio.visakorpi@uta.fi); MN (matti.nykter@uta.fi)

The poster abstract should include the Prostate cancer is the third most common cause of male cancer deaths in developed countries, with castration resistance being the most challenging clinical problem. Here we report an investigation into novel transcripts in primary prostate cancers (PCs), and castration resistant prostate cancers (CRPCs) in particular. We characterized 28 PCs, 13 CRPCs, and 12 benign prostatic hyperplasias (BPH) using deep transcriptome sequencing (RNA-seq). Reference-based transcriptome assembly uncovered 145 previously unannotated intergenic PC associated transcripts or isoforms. The expression patterns of the transcripts were confirmed in two previously published independent cohorts of primary tumors (n=30 and n=34), 21 PC cell lines, and 25 normal tissues or cell lines. By integrating publicly available ChIP-sequencing data and transcription factor (TF)-transcript expression correlations, we identified a transcript that positively correlated with ERG expression and exhibited an ERG binding event in a PC cell line coinciding with the canonical ETS-family TF binding motif at its proximal promoter region. Enrichment of histone modification H3K4me3 and PolII at the promoter of the transcript in the cell line provided further evidence of open chromatin and active transcription. We downregulated the expression of the transcript with siRNAs in the cell line and observed a decrease in cell growth and reduced migration, invasion and colony formation. Annexin V assay indicated increased rate of apoptosis in the cells. Pathway analysis indicated that cell cycle, mitosis and apoptosis were the most extensively affected cellular processes. These results suggested that the transcript significantly affects tumor growth in ETS-positive prostate cancers.

Maximum Likelihood Scaffold Assembly

Anton Akhi [1,*], Alexey Sergushichev [1], Fedor Tsarev [1]

[1]: Genome Assembly Algorithms Laboratory, St. Petersburg National Research University of IT, Mechanics and Optics, St. Petersburg, Russia.

[*]: To whom correspondence should be addressed.

Emails: anton.akhi@gmail.com; alsergbox@gmail.com; fedor.tsarev@gmail.com

Genome assembly is usually performed in two steps: contig (continuous genome subsequences) assembly and scaffold (ordered contig set) assembly. Mate-pair reads with typical insert size of several thousand base-pairs are used for scaffold assembly. Scaffold assembly usually consists of two stages: estimation of the distance between contigs and finding the order of contigs in scaffolds.

For the distance estimation reads are aligned to contigs. For reads alignment we use Bowtie. The proposed algorithm uses a common assumption of mate-pair insert size normal distribution. The likelihood function for each pair of contigs connected by mate-pair reads takes into consideration not only reads that connect a pair of contigs but also contigs lengths and the number of mate-pair reads that do not connect the pair of contigs.

For the contig ordering graph is constructed where nodes represent contigs and edges represent mate-pair reads. Nodes having a degree above some threshold are removed to simplify the graph. The remaining graph looks like a number of chains that still have a couple of complex parts. To order contigs in these parts maximum likelihood approach is used. For every possible contigs ordering most likely distances between them are found using gradient descent and the order that has a maximum likelihood is taken as the result.

The algorithm has been tested on Illumina reads of *E. Coli* bacteria. Contigs were built from pair-end reads. All the data as well as reference genome sequence are available on the internet: http://genome.ifmo.ru/node/17. The distance estimation results were compared with SOPRA (http://www.biomedcentral.com/1471-2105/11/345) that uses average distance as estimation. The proposed algorithm is closer to the actual value than SOPRA in 67% of cases and in other 15% results are the same. The mean error is -2 and -99 for the proposed algorithm and SOPRA respectively while the error standard deviation is 489 and 1881. The scaffolding results were compared with OPERA (http://www.ncbi.nlm.nih.gov/pubmed/21929371). Total of 13 non-singleton scaffolds were built against 17 for OPERA. These scaffolds cover 81% of reference genome sequence against 95% for OPERA and N50 is 416405 against 474009. Total of 5% and 3% erroneous connections were found in the resulting scaffolds of the proposed algorithm and OPERA respectively.

Acknowledgements

Research is supported by the federal program “Research and scientific-pedagogical personnel of innovative Russia in 2009-2013” (contract №16.740.11.0495, agreement №14.B37.21.0562).

Differential alternative splicing identification in kinome and phosphatome in prostate cancer from RNA-Seq data

Huijuan Feng [1], Tingting Li [2], Xuegong Zhang *[1,3]

[1]: MOE Key Laboratory of Bioinformatics, Bioinformatics Division and Center for Synthetic and Systems Biology, TNLIST / Department of Automation, Tsinghua University, Beijing 100084, China.

[2]: Department of Biomedical Informatics and Institute of Systems Biomedicine, Peking University Health

Science Center, Beijing 100191, China

[3]: School of Life Sciences, Tsinghua University, Beijing 100084, China.

*: To whom correspondence should be addressed.

Emails: HF(fhj11@tsinghua.edu.cn); TL (li-tt03@mails.tsinghua.edu.cn); XZ (zhangxg@tsinghua.edu.cn)

Kinases and phosphatases are large protein families and play key roles in various biological processes like cell cycle and proliferation. Recent advances have revealed that some of the kinases and phosphatases may act as oncoproteins in human cancers and alternative splicing can have a significantly impact on the function of kinases and phosphatases and thus to some extent plays a role in oncogenic process in tumor progression. RNA-Seq has become a promising tool to identify differential alternative splicing between samples. To have a better understanding on what kinase and phosphatase genes would have alternative splicing events and how they would function during prostate cancer progression, we collected 518 kinases and 157 phosphatases, and performed differential alternative splicing analysis from two published RNA-Seq datasets of prostate cancer, with an exon-centric method we developed for detecting differential splicing named DSGSeq [1] and estimated isoform ratio of differential splicing genes with NURD [2]. We detected differential splicing genes with possible splicing loci and genes that undergo isoform switching events that may play role in prostate cancer progress. Further functional enrichment analysis of differential splicing genes was found to be related to prostate-cancer-specific pathways. An example gene CDK5 is found to be carried out major isoform switching event and may be functional in cancer cell migration.

References

1. Weichen Wang, Zhiyi Qin, Zhixing Feng, Xi Wang, Xuegong Zhang. Identifying differentially spliced genes from two groups of RNA-seq samples. *Gene*. 2012.11.045
2. Xinyun Ma, Xuegong Zhang. NURD: A New Tool for Estimating Isoform Expression from Non-Uniform RNA-Seq Data. *Under review*

Web-QUAST: Quality Evaluation of Genome Assemblies

Alexey Gurevich [1,*], Vladislav Saveliev [1], Nikolay Vyahhi [1], Glenn Tesler [2]

[1]: Algorithmic Biology Laboratory, St. Petersburg Academic University, Russian Academy of Sciences, St. Petersburg, Russia, 194021.

[2]: Department of Mathematics, University of California, San Diego, La Jolla, CA 92093-0112, USA.

[*]: To whom correspondence should be addressed.

Email: AG (gurevich@bioinf.spbau.ru)

De novo genome assembly represents a formidable challenge, for which dozens of genome assembly programs have been developed based on different algorithmic approaches. Results may vary significantly, which leads to the question of how to assess assembly quality and compare assemblies to each other. Several attempts to address this problem were recently published ([1], [2], [3]). However, one should be well-prepared to produce benchmarking experiments on other data sets and assemblies.

We introduce a web tool for fast and convenient quality evaluation and comparison of assemblies. The tool is based on QUAST [4]. It uses a number of well-known metrics, including contig accuracy, number of genes discovered, N50, and others, as well as introducing new ones. Users can upload multiple assemblies and choose a corresponding reference genome and gene annotation from our database or upload their own. For research on previously unsequenced species, one may evaluate assemblies without a reference.

Web-QUAST performs a comprehensive analysis and produces interactive web-based reports with summary tables and colorful plots. This helps users to examine characteristics of assemblies and choose the most appropriate ones for their research.

Web-QUAST is available at http://quast.bioinf.spbau.ru

References

1. Earl, D. *et al*. Assemblathon 1: A competitive assessment of de novo short read assembly methods. *Genome Res*, 21(12): 2224-2241, 2011.
2. Salzberg, S. *et al*. GAGE: A critical evaluation of genome assemblies and assembly algorithms. *Genome Res*, 22(3): 557-567, 2011.
3. Bradnam, K., Fass, J. *et al*. Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species. *GigaScience*: submitted, 2013.
4. Gurevich, A. *et al*. QUAST: Quality Assessment Tool for Genome Assemblies. *Bioinformatics*: to appear, 2013.

High-throughput sequencing of fecal microflora of autistic and control children

Beili Sun[1*], Dongrui Zhou[2], Qinyu Ge[2], Jing Tu[1], Zuhong Lu[1]

[1]: State Key Laboratory of Bioelectronics, Southeast University, Nanjing, 210096, China

[2]: Key Laboratory of Child Development and Learning Scienceof Ministry of Education of China, Southeast University, Nanjing, 210096, China

[*]: To whom correspondence should be addressed.

Emails: Beili Sun (sunbl@seu.edu.cn); Dongrui Zhou (junbai1013@seu.edu.cn); Qinyu Ge (geqinyu@seu.edu.cn); Jing Tu (jtu@seu.edu.cn); Zuhong Lu (zhlu@seu.edu.cn)

Autism is a complex disorder resulting in profound behavioural and emotional problems. Gastrointestinal (GI) disorders and associated symptoms are commonly reported in autistic patients, suggesting the potential role of the abnormal GI microbiota in autistic children. It is generally appreciated that onset of autism is known to occur often following antimicrobial therapy. Additionally, antimicrobials are known to have a significant effect in both the relapse and continuation of the autistic condition. It has been reported that the count of *clostridial* species in stools of autistic children was higher than in control children. Pyrosequencing study of the fecal microflora of autism children showed the significant difference between groups of varying severities of autism at the phylum level. We analyzed the fecal microbiota of two autism pedigrees and two healthy children using the SOLiD high-throughput sequencing technology, in order to find the specific bacterial in autism subjects. More than 700 bp of the 16S rRNA gene were amplified, randomly sequenced and aligned with RDP II database, resulting in 763 genera assigned. At the phylum level, *Firmicutes* showed the most abundance in all the subjects, while *Actinobacteria* appeared to be the second abundant in autism children rather than *Proteobacteria* in healthy children. There were 9 genera discovered only in autistic subjects: *Methylococcus, Caenispirillum, Azoarcus, Cryobacterium, Nesterenkonia, Gordonia, Herbidospora, Brachybacterium, Leptolinea*. The UniFrac results showed that no significant difference occurred between autism and healthy children.

Genomic dissection of a tumor in an iPS mouse at single cell resolution

Xuexia Miao[1], Rongrong Le[2], Guojing Liu[1], Shuangli Mi[1], Shaorong Gao[2], and Jun Cai[1*]

[1]: Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing, China.

[2]: National Institute of Biological Sciences, Beijing, China.

[*]: Email: juncai@big.ac.cn

Induced pluripotent stem (iPS) cell totipotency was demonstrated by the fact that iPS cells can produce viable mice through tetraploid complementation. But some studies improved that mice generated from tetraploid complementation had the problem of possible tumor formation. Here we dissect a subcutaneous tumor in a one-month-old mouse derived from iPS cells in order to track the multistep process of mutation in tumorigenesis. We utilized the strategy of inducing several single cells from different parts of the mouse tumor into pluripotent stem cell lines and then sequencing the exomes of the iPS cell lines. The high-throughput sequencing data and further Sequenom technology revealed and validated the single nucleotide variants and copy number variations specific in each iPS cell line. These mutations represent the same genomic background of every single tumor cell from which the corresponding iPS cell line was cultured. Therefore we use an indirect way to analyze the genomic composition of the tumor at single-cell resolution. A tree of cell lineage among 13 single tumor cells was constructed which supports the conclusion that, in the scale of single cell, homogeneous cell cluster and heterogeneous cells co-exist in the same tumor cell population during tumorigenesis.

EBSeq: An empirical Bayes hierarchical model for inference in RNA-seq

Ning Leng [1,*], John Dawson [1], James Thomson [2], Victor Ruotti [2], Anna Rissman [3], Bart Smits [3], Jill Haag [3], Michael Gould [3], Ron Stewart [2], Christina Kendziorski [4]

[1]: Department of Statistics, University of Wisconsin, Madison, WI

[2]: Morgridge Institute for Research, Madison, WI

[3]: McArdle Laboratory for Cancer Research, Department of Oncology, University of Wisconsin, Madison, WI

[4]: Department of Biostatistics and Medical Informatics, University of Wisconsin, Madison, WI

[*]: To whom correspondence should be addressed.

Emails: NL (nleng@wisc.edu); JD (jadawson@wisc.edu); JT (jthomson@morgridgeinstitute.org); VR (ruotti@wisc.edu); AR (rissman@wisc.edu); BS (bsmits@wisc.edu); JH (jhaag@facstaff.wisc.edu); MG (gould@oncology.wisc.edu); RS (rstewart@morgridgeinstutute.org); CK(kendzior@biostat.wisc.edu)

Motivation: Messenger RNA expression is important in normal development and differentiation, as well as in manifestation of disease. RNA-seq experiments allow for the identification of differentially expressed (DE) genes and their corresponding isoforms on a genome-wide scale. However, statistical methods are required to ensure that accurate identifications are made. A number of methods exist for identifying DE genes, but far fewer are available for identifying DE isoforms. When isoform DE is of interest, investigators often apply gene-level (count-based) methods directly to estimates of isoform counts. Doing so is not recommended. In short, estimating isoform expression is relatively straightforward for some groups of isoforms, but more challenging for others. This results in estimation uncertainty that varies across isoform groups. Count-based methods were not designed to accommodate this varying uncertainty and consequently application of them for isoform inference results in reduced power for some classes of isoforms and increased false discoveries for others.
Results: Taking advantage of the merits of empirical Bayesian methods, we have developed EBSeq for identifying DE isoforms in an RNA-seq experiment comparing two or more biological conditions. Results demonstrate substantially improved power and performance of EBSeq for identifying DE isoforms. EBSeq also proves to be a robust approach for identifying DE genes.
Availability: An R package containing examples and sample data sets is available at http://www.biostat.wisc.edu/~kendzior/EBSEQ/

References

N. Leng, J.A. Dawson, J.A. Thomson, V. Ruotti, A.I. Rissman, B.M.G. Smits, J.D. Haag, M.N. Gould, R.M. Stewart, and C. Kendziorski. EBSeq: An empirical Bayes hierarchical model for inference in RNA-seq experiments. *Bioinformatics*, 2013

NGS assembly of highly polymorphic diploid genomes

Yana Safonova[1,*], Alexey Kondrashov[2], Maria Logacheva[2],
Aleksey Penin[2], Anton Bankevich[1,*]

[1]: Algorithmic Biology Laboratory, St. Petersburg Academic University, St. Petersburg, Russia
[2]: Department of Genetics, Moscow State University, Moscow, Russia
[*]: To whom correspondence should be addressed.
Emails: YS(safonova.yana@gmail.com); AB(anton.bankevich@gmail.com)

Assembly of highly polymorphic diploid genomes (with polymorphism level 1-10%) is a computationally hard problem. Several approaches have already been proposed for assembling such datasets (Vinson et al 2005 [1]; N.Donmez, M. Brudno 2011 [2]), but all known methods are based on overlap-layout-consensus and cannot be applied to NGS (Next Generation Sequencing) data. Existing NGS assemblers are inefficient too: they do not use information about genome diploidy which, as we show in this work, can help improve assembly. We present an algorithm for assembling highly polymorphic diploid genomes for NGS data, that is able to utilize information about diploidy. Our approach combines de Bruijn graph based methods for computing draft contigs with an application of the overlap graph constructed from draft contigs; the overlap graph is used to improve assembly and search for polymorphism. Due to high polymorphism level, sequences of overlapping contigs may differ significantly, which further complicates the construction of an overlap graph. We present an algorithm for the search of contig overlaps based on polymorphism masking done with a modification of the de Bruijn graph. This lets us find overlaps by analyzing draft contigs mapping to the modified de Bruijn graph. The algorithm has been tested on a dataset corresponding to the mix of two libraries obtained by sequencing of two genomes with 5% difference. Our experiments demonstrate that the resulting assembly considerably improves even upon the results of assembly where each individual dataset was processed separately.

References

[1] Jade P. Vinson, David B. Jaffe, Keith O'Neill, Elinor K. Karlsson, Nicole Stange-Thomann, Scott Anderson, Jill P. Mesirov, Nori Satoh, Yutaka Satou, Chad Nusbaum, Bruce Birren, James E. Galagan, Eric S. Lander. Assembly of polymorphic genomes: algorithms and application to Ciona savignyi. *Genome research*, 15(8): 1127–1135, 2005.

[2] Nigun Donmez and Mikhael Brudno. Hapsembler: An assembler for highly polymorphic genomes. *Proceedings of RECOMB 2011*: 38-52, 2011.

Effective computational tools for next generation microbiome sequence analysis

Weizhong Li [1, *], Sitao Wu[1], Limin Fu[1]

[1]: University of California San Diego, Center for Research in Biological Systems, La Jolla California, USA

[*]: To whom correspondence should be addressed.

Emails: WL (liwz@sdsc.edu); SW (siw006@ucsd.edu); LF (l2fu@ucsd.edu)

Complex and dynamic microbial communities play a profound role in shaping the environment they inhabit. Recent advances in metagenomics approaches and high-throughput Next Generation Sequencing (NGS) technologies have enabled comprehensive study of the microbes from many diverse environments, such as the human microbiota that are thought to deeply influence human health. Vast amounts of sequences being generated impose extreme challenges in computational analyses such as sequence error correction, assembly, mapping, gene prediction, and function analysis. To address these challenges, we build a set of unique NGS-specific tools using fast clustering and alignment algorithms combined with statistical methods and visualization interface[1-11]. These tools not only allow orders of magnitude faster computational analysis but also offer inimitable way to investigate novel and complex data. We have analyzed several Terabytes of sequence for hundreds of human microbiome samples from healthy people and patients with diseases and obtained comprehensive results from our analysis.

References

1. Zhu Z, Niu B, Chen J, Wu S, Sun S, Li W: MGAviewer: A desktop visualization tool for analysis of metagenomics alignment data. *Bioinformatics* 2012.
2. Li W, Fu L, Niu B, Wu S, Wooley J: Ultrafast clustering algorithms for metagenomic sequence analysis. *Brief Bioinform* 2012, 13(6):656-668.
3. Fu L, Niu B, Zhu Z, Wu S, Li W: CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* 2012, 28(23):3150-3152.
4. Wu S, Zhu Z, Fu L, Niu B, Li W: WebMGA: a customizable web server for fast metagenomic sequence analysis. *BMC Genomics* 2011, 12:444.
5. Sun S, Chen J, Li W, Altintas I, Lin A, Peltier S, Stocks K, Allen EE, Ellisman M, Grethe J *et al*: Community cyberinfrastructure for Advanced Microbial Ecology Research and Analysis: the CAMERA resource. *Nucleic Acids Res* 2011, 39(Database issue):D546-551.
6. Niu B, Zhu Z, Fu L, Wu S, Li W: FR-HIT, a very fast program to recruit metagenomic reads to homologous reference genomes. *Bioinformatics* 2011, 27(12):1704-1705.
7. Niu B, Fu L, Sun S, Li W: Artificial and natural duplicates in pyrosequencing reads of metagenomic data. *BMC Bioinformatics* 2010, 11:187.
8. Huang Y, Niu B, Gao Y, Fu L, Li W: CD-HIT Suite: a web server for clustering and comparing biological sequences. *Bioinformatics* 2010, 26(5):680-682.
9. Huang Y, Gilna P, Li WZ: Identification of ribosomal RNA genes in metagenomic fragments. *Bioinformatics* 2009, 25(10):1338-1340.
10. Li W, Wooley JC, Godzik A: Probing metagenomics by rapid cluster analysis of very large datasets. *PLoS ONE* 2008, 3(10):e3375.
11. Li W: Analysis and comparison of very large metagenomes with fast clustering and functional annotation. *BMC Bioinformatics* 2009, 10:359.

DSGseq: A software for detecting differential splicing genes between two groups of samples

Zhiyi Qin[1], Weichen Wang [1,2], Xuegong Zhang[1,3,*]

[1]: MOE Key Laboratory of Bioinformatics, Bioinformatics Division and Center for Synthetic and Systems Biology, TNLIST / Department of Automation, Tsinghua University, Beijing 100084, China

[2]: Department of Operational Research and Financial Engineering, Princeton University, NJ 08544, USA

[3]: School of Life Sciences, Tsinghua University, Beijing 100084, China

[*]: To whom correspondence should be addressed.

Emails: QZ (qzy06@mails.tsinghua.edu.cn); WW (weichenw@princeton.edu); ZX (zhangxg@tsinghua.edu.cn)

Recent study revealed that most human genes have alternative splicing through RNA-seq. As differences in the relative abundance of the isoforms of a gene can have significant biological consequences, identifying genes that are differentially spliced between two groups of samples is having been an important task in the study of transcriptomes with next-generation sequencing technology. There have been several methods aimed to detect differential splicing genes, but most of them were designed for comparing two individual samples. Recently we studied the question of identifying genes that are differentially spliced between two groups of samples and proposed one exon-based method used an NB-statistic with the negative binomial model to detect differential splicing [1]. This was a new route to study alternative splicing quantitatively in an exon-centric manner.

We implemented this method named DSGseq, which can detect differentially spliced genes between two groups of samples. It is designed for comparing two groups of RNA-seq samples and does not need to infer isoform structure or to estimate isoform expression. With counting exons' covered reads and then applying the software DSGseq, one can identify differentially spliced genes between two groups of samples, as well as the exons that contribute the most to the differential splicing. Simulation experiments showed that the proposed method performs well on both detecting differentially spliced genes and identifying the alternative exons. Experiments on real RNA-seq data of human kidney and liver samples illustrated the method's good performance and applicability. DSGseq is written in R and can run on all major computer platforms running Windows or Unix/Linux. The software tool is available at: http://bioinfo.au.tsinghua.edu.cn/software/DSGseq for free academic use.

References

1. Weichen Wang, Zhiyi Qin, Zhixing Feng, Xi Wang and Xuegong Zhang, 2012. Identifying differentially spliced genes from two groups of RNA-seq samples. *Gene*, http://dx.doi.org/10.1016/j.gene.2012.11.045.

Use of uneven read coverage depth in bacterial single-cell repeat resolution

Dmitry Antipov [1], Ksenia Krasheninnikova [1,*], Pavel A. Pevzner [1,2]

[1]: Algorithmic Biology Laboratory, St. Petersburg Academic University, Russian Academy of Sciences, St. Petersburg, Russia, 194021.

[2]: Department of Computer Science & Engineering, University of California, San Diego, La Jolla, USA, CA 92093-0404

[*]: To whom correspondence should be addressed.

Emails: DA (dmitrij.antipov@gmail.com); KK (krasheninnikova@gmail.com); PP (ppevzner@ucsd.edu)

One of the main single-cell sequencing data characteristics is highly uneven read coverage depth. However, coverage is likely to be uniform locally, so that closely located genomic regions have similar read coverage. This observation can be used to resolve repeats in single-cell genome assembly. In this work we propose an algorithm that uses uneven coverage for repeat resolution and is designed for compressed de Bruijn graph. The algorithm consists of two steps: detection and resolving of repetitive elements.

Firstly, we use compressed de Bruijn graph structure and paired-end reads (if available) to detect edges that are likely to appear in genome more than once (repetitive edges). We define repetitive element as a group of such adjacent repetitive edges. We consider these elements to appear in genome several times and edges that surround these repetitive elements to represent unique genomic sequence.

The second step aims to match the incoming edges to outgoing, which are adjacent to the same repetitive element. To do so we analyze coverage values of incoming and outgoing edges. Since we assume that local coverage is uniform, we choose one incoming edge and one outgoing edge with similar coverage and thus likely to be located in the same genomic region. In order not to introduce misassemblies we ignore cases when more than a pair of incoming and outgoing edges have similar coverage.

The proposed approach was tested with SPAdes genome assembler [1] (http://bioinf.spbau.ru/spades) and demonstrates a notable assembly quality improvement.

References

1. Anton Bankevich, *et al*. SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *J Comput Biol*, 19(5): 455-477, 2012.
2. Toshiaki Namiki, Tsuyoshi Hachiya, Hideaki Tanaka, Yasubumi Sakakibara. MetaVelvet: an extension of Velvet assembler to *de novo* metagenome assembly from short sequence reads. *Nucleic Acids Res*, 40(20): e155, 2012.

Pathobuster: a web server for estimating bacterial pathogenicity

Salvatore Cosentino[1*], Mette Voldby Larsen [1], Ole Lund [1]

[1]: Center for Biological Sequence Analysis (CBS), Technical University of Denmark, Kemitorvet, Building. 208 , 2800 Kongens Lyngby, Denmark.

[*]: To whom correspondence should be addressed.

Emails: SC (salvocos@cbs.dtu.dk); MVL (metteb@ cbs.dtu.dk); OL (lund@ cbs.dtu.dk)

Currently almost 4000 complete genomes are available (http://www.genomesonline.org/). This massive amount of data gives us the chance to try to answer many questions associated to bacteria.

Every month thousands of people get infected and many die because of pathogen bacteria. Pathobuster is a webserver that allows also people with a low knowledge of computers to upload a genome (complete or draft) or raw reads from clinical samples and have an estimation of how dangerous is the unknown organism. Understanding how dangerous it is could help in detecting possible bacterial outbreaks and intervene in time to save human lives.

Pathobuster can analyse all kind of bacteria in less than 10 minutes by identifying genes associated to pathogenicity. In case of raw reads the webserver will first perform a denovo assembly and then use the obtained draft genome for the analysis.

Pathobuster will be available for everybody and free of charge, and will be one of the services offered by the *Center for Genomic Epidemiology*, which aim is to detect and control bacterial outbreaks (http://www.genomicepidemiology.org/).

Accuracy of Next generation sequencing data for EGFR mutation detection in non-small-cell lung cancer

Yu-Cheng Li [1], Hsuan-Yu Chen [1], Shin-Sheng Yuan [1], Yi-Chiung Hsu [1], Ker-Chau Li [1]

[1]: Institute of Statistical Science, Academia Sinica, Taiwan

Emails: Y.C.L. (yuchengli@stat.sinica.edu.tw); H.Y.C. (hychen@stat.sinica.edu.tw); S.S.Y. (syuan@stat.sinica.edu.tw); Y.C. H. (syic@stat.sinica.edu.tw); K.C.L. (kcli@stat.sinica.edu.tw)

The accuracy of next-generation sequencing (NGS) has greatly improved in recent years owing in large part to the ability of generating longer reads. However the sequencing errors still persist at a non-ignorable rate and the validation by alternative techniques is necessary. We are interested in the cancer genomics application of NGS. Our web-lab collaborators has earlier established a protocol for matrix-assisted laser desorption ionization–time of flight mass spectrometry (MALDI-TOF MS) that allows for the detection of the cancer-driving EGFR activating mutations (L858R, Del19) and TKI drug-resistant mutation (T790M) at low frequencies in no-small-cell lung cancer (Su et al, 2012, Journal of Clinical Oncology, page 433- 440). Here we analyzed the read data from Illumina Exon Capture and compared with the mass spectrometry results. We observed a high degree of similarity between the two methods in general. We further investigated the inconsistent cases and addressed the important issue of the read coverage of NGS in discovering new low frequency mutations. This presentation is based on joint work with Drs S.L. Yu, K.Y. Su, P.C. Yang and G.C Chang.

An ensemble algorithm for incorporating results of multiple de novo genome assembly methods

Kui Xu [1], Ke Chen [1,*]

[1]: BinShui Road #399, XiQin District, School of Computer Science and Software Engineering, Tianjin Polytechnic University, Tianjin, China 300387

[*]: To whom correspondence should be addressed.

Emails: KX (kuixu.tj@gmail.com); KC (kchen1.tj@gmail.com)

Abstract: The emergence of next-generation sequencing platforms led to resurgence of research in whole-genome shotgun assembly algorithms and software. DNA sequencing data from recent platforms typically present shorter read lengths, higher coverage, and different error profiles compared with Sanger sequencing data. Since 2005, several assembly software packages have been created or revised for de novo assembly of next-generation sequencing data, including Velvet, ABySS, AllPaths and SOAPdenovo. Although it is widely recognized that assembly results generated by different methods are complementary, no algorithm was presented to merge the results generated by multiple methods. In this study, we propose a graph-based algorithm that incorporates the contigs generated by multiple assembly algorithms. The algorithm first identifies common segments for each pair of contigs. Then contigs with significant common segments are assigned to one cluster. For each cluster, a directed-graph is established to imply the relation between different contigs. By employing our algorithm, the longest contig is extended for up to 50%. We also demonstrate that the assembly results could be further improved when more assembly algorithms are included.

A simple method for detecting cross-sample contamination in deep exome sequencing data

Xueya Zhou[1*], Suying Bao[2], Youqiang Song[2], and Xuegong Zhang[13*]

[1]: MOE Key Laboratory of Bioinformatics, Bioinformatics Division and Center for Synthetic and Systems Biology, TNLIST / Department of Automation, Tsinghua University, Beijing 100084, China

[2]: Department of Biochemistry, Li Ka-Shing Faculty of Medicine, The University of Hong Kong, Hong Kong, China

[3]: School of Life Sciences, Tsinghua University, Beijing 100084, China

[*]: To whom correspondence should be addressed.

Emails: X.Y.Z (zhou-xy02@mails.tsinghua.edu.cn); S.B (sybao@hku.hk); Y.S (songy@hku.hk); X.G.Z (zhangxg@tsinghua.edu.cn)

Cross-sample contamination is a potential issue in exome sequencing which will cause genotype misclassification, create false positive variant calls, and confound the detection of somatic mutations. Whereas high level contamination or sample mixing can be easily discovered from unusual heterozygosity values; low level contamination poses a more difficult statistical problem. Motivated by our observations in ongoing exome projects, we found that the major effect of low level cross-sample contamination was to create false heterozygous variant calls with lower than expected proportion of reference alleles in supporting reads. The resulting allele balance (AB) histogram of heterozygous SNP calls deviates from the expected normal distribution with mean 0.5 to become a mixture distribution. We propose to test the mixture proportion in the AB histogram as a simple method to quantify the level of contamination. When applied to the in silico contaminated exome data, the proposed method showed similar performance to other more sophisticated methods.

High Throughput Mutation Screening of the TP53 Gene in Lung Cancer Using Single Molecule Real Time (SMRT) Sequencing

Jin Jen, JinSung Jang, Karl Oles, Ana Robles*, Jaime Davila, Bruce Eckloff, Curt Harris*, and Eric Wieben.

The Medical Genome Facility, Center for Individualized Medicine, Mayo Clinic, Rochester, Minnesota, and * the Laboratory of Human Carcinogenesis, Center for Cancer Research, National Cancer Institute, Bethesda, Maryland, United States.

TP53 is one of the most commonly mutated genes in human cancer. In wild type form, TP53 functions as a tumor suppressor gene by negatively regulating cell cycle and inducing apoptosis to preserve genomic stability. In non-small cell lung cancer (NSCLC), mutations of TP53 occur in nearly 90% of squamous cell carcinomas and roughly 50% of adenocarcinomas. Clinical studies suggest that NSCLCs with TP53 alterations have less favorable prognosis and may confer tumor resistant to chemotherapy and radiation. Therefore, it is highly desirable to develop robust and efficient ways to determine the genetic status of TP53 or other highly targetable genes. Utilizing the Next Generation Sequencing (NGS) based high throughput technologies, we explored the use of the PacBio Single Molecule SMRT sequencer for rapid and high-throughput, targeted gene sequencing for the entire coding region of the TP53 gene. Using a multiplex PCR strategy, the entire 11 coding exons of the TP53 gene were amplified in a single well and indexed for each sample. SMRTbell libraries were then generated after pooling up to 8 samples. After bead-purification of the PCR products, single molecule sequencing was then carried out to generate up to 75,000 circular consensuses reads (CCR) per SMRT cell. Our results show that the CCRs generated high quality sequences that were accurate and detected mutations present at less than 10% of the alleles in the sample. Our experience sequencing 50 lung adenomas with and without TP53 gene mutations demonstrate that PacBio single molecule sequencing can provide a highly robust, reliable and cost effective method for rapid identification of genetic changes commonly associated with cancer development and progression. This approach is also easily applicable for the rapid analysis of any other candidate genes of biological and clinical interest.

Meta-Mesh: Metagenome Database and Data Analysis System

Xiaoquan Su[1], Baoxing Song[1], Jian Xu[1], Kang Ning[1,*]

[1]:Qingdao Institute of Bioenergy and Bioprocess Technology, Chinese Academy of Sciences

[*]: To whom correspondence should be addressed.

Emails: XS(suxq@qibebt.ac.cn); BS(songbx@qibebt.ac.cn); JX(xujian@qibebt.ac.cn); KN(ningkang@qibebt.ac.cn)

With the current accumulation of metagenome data, it is possible to build a database of rigorously selected metagenomic samples (also referred as "metagenomic communities" here) of interests. Any metagenomic samples could then be searched against this database to find the most similar sample(s). However, on one hand, current databases with a large number of metagenomic samples mostly serve as data repositories but not well annotated database, and they only offer few functionalities for analysis. On the other hand, the few available methods to measure the similarity of metagenomic data could only compare a few pre-defined set of metagenome. It has long been intriguing scientists to effectively find similar microbial communities from a large repository, to know the meta-information of these samples and to examine how similar these samples are.

In this study, we have proposed a novel system, Meta-Mesh (http://www.meta-mesh.org/), which includes a database and its companion analysis system that could systematically and efficiently search similar metagenomic samples. In the database part, we have collected more than 7, 000 high quality and well annotated metagenomic samples from the public domain and in-house facilities. The analysis part includes a list of tools which could accept metagenomic samples, build taxonomical annotations, and then search for similar samples against its carefully selected, well-organized and annotated database by a fast scoring function. It has a multi-thread submission portal and a well-designed data management client for easy submission of large and complex data sets and integrates a variety of viewers to provide a visualization solution for result analysis. Users can also use Meta-Mesh to compare their samples and get a similar score matrix. In the Meta-Mesh online service work, user access is protected to ensure data privacy.

Meta-Mesh would serve as a database and data analysis system to quickly parse and identify similar metagenomic samples from a large pool of well annotated samples.

Whole genome comparative study of eleven *Mycobacterium tuberculosis* strains isolated in China

Qi Wang[1,§], Yu Pang[2,§], Peijin Zhang[3,5§], Hui Guo[1,4], Yang Zhou[2], Huanqin Dai[1], Kaixia Mi[1], Lixin Zhang[1*], Gil Alterovitz[3*] and Yanlin Zhao[2*]

[1]: Key Laboratory of Pathogenic Microbiology and Immunology, Institute of Microbiology, Chinese Academy of Sciences Beijing, 100190, China;

[2]: National Center for Tuberculosis Control and Prevention, Chinese Center for Disease Control and Prevention, Beijing 102206, China;

[3]: Department of Electrical Engineering and Computer Science, MIT, Cambridge, MA 02115, USA;

[4]: Graduate Universities of Chinese Academy of Sciences, Beijing, 100049, China;

[5]: MIT PRIMES Program, Cambridge, MA 02115, USA;

[*]: To whom correspondence should be addressed.

[§]: These authors contributed equally to this work.

Emails: LXZ (zhanglixin@im.ac.cn); YLZ (zhaoyanlin@chinatb.org); GA (gil@mit.edu)

Beijing genotype strains are one of the most predominant strains in China. Though a significant association between the Beijing genotype strains and resistance was observed, the molecular mechanism is still unknown. In order to determine whether Beijing genotype strains have an increased intrinsic resistance to anti-TB drugs or an enhanced capacity to gain resistance against these drugs, we performed a whole genome shotgun sequencing of eleven *Mycobacterium tuberculosis* strains (including seven Beijing genotype strains and four non-Beijing genotype strains) isolated in China. Together with nine Beijing genotype strains and twenty-five non-Beijing genotype strains publicly available, we were able to identify 903 SNPs common in Beijing genotype strains, including several mutations significantly enriched in Beijing genotype strains, such as katG(S315T), gyrA(A90V, D94G), gidB(E92D). Regions of difference (RDs) derived from whole-genome sequencing also indicated five regions differ significantly between Beijing and non-Beijing genotype strains, including a ~1kb region in lipoprotein A(lppA) which is found only in pathogenic mycobacteria. The SNPs and RDs presented in this study will facilitate future molecular epidemiological and phylogenetic studies on Beijing strains.

References

1. Ioerger TR, *et al.* Genome analysis of multi- and extensively-drug-resistant tuberculosis from KwaZulu-Natal, South Africa. PLoS One 2009;4(11):e7778.
2. Ioerger TR, *et al*. 2010. The non-clonality of drug resistance in Beijing-genotype isolates of Mycobacterium tuberculosis from the Western Cape of South Africa. BMC Genomics 11:670.
3. Schürch AC, *et al*. SNP/RD typing of *Mycobacterium tuberculosis* Beijing strains reveals local and worldwide disseminated clonal complexes. PLoS One. 2011;6(12):e28365.

Comparison of *D. melanogaster* and *C. elegans* Developmental Stages by modENCODE RNA-Seq data

Jingyi Jessica Li[1], Haiyan Huang[1], Peter J. Bickel[1], Steven E. Brenner [2*]

[1]: Department of Statistics, University of California, Berkeley, CA 94720, USA

[2]: Department of Plant & Microbial Biology, University of California, Berkeley, CA 94720, USA

[*]: To whom correspondence should be addressed.

Email addresses: SEB (brenner@compbio.berkeley.edu); PJB (bickel@stat.berkeley.edu); HH (hhuang@stat.berkeley.edu)

Drosophila melanogaster and *Caenorhabditis elegans* are two well-studied model organisms in developmental biology. Their morphological development differ greatly, yet we postulated that there may nonetheless be underlying shared developmental programs employing orthologous genes. The availability of modENCODE RNA-Seq data for developmental stages of the two organisms enables a transcriptome-wide comparison study to address this question. We undertook a comparison of their developmental time courses, seeking commonalities in orthologous gene expression. Our approach centers on using stage-associated orthologous genes to link the two organisms. For every stage in each organism, we select stage-associated genes which are defined as relatively highly expressed at that stage compared with others. To test the dependence of a pair of *D. melanogaster* and *C. elegans* stages in terms of orthologous gene expression, we used an overlap statistic—the number of orthologous gene pairs associated with both stages. Under the null hypothesis that the two stages have independent gene expression profiles, a p-value can be calculated for the overlap statistic.

We first carried out the test on pairs of stages within *D. melanogaster* and *C. elegans* respectively, and we found that temporally adjacent stages in both species exhibit high dependence in gene expression, supporting the validity of this approach. We also found other connections, such as female fly adults having similar stage-associated genes as fly embryos. Most important, when comparing fly with worm, we observed a strong colinearity of their developmental time courses from early embryos to late larvae. Another parallel collinear pattern is found between fly white prepupae through adults and worm late embryos through adults. Small p-values are also observed between fly early embryos and worm adults, between fly female adults and worm early embryos, and between fly female adults and worm adults. Our results are the first findings regarding the comparison between *D. melanogaster* and *C. elegans* time courses. Investigating stage-associated genes overlapped between stages shows that many-to-one fly-worm orthologs are key factors leading to the two collinear patterns. Some orthologs having known biological functions have been verified to play similar roles in both organisms, and their mapping in this study may help inform their functions in the development of *D. melanogaster* and *C. elegans*.

Inferring HIV Quasispecies from Paired-end Reads

Serghei Mangul[3], Nicholas Wu[2], Nicholas Mancuso[3], Alex Zelikovsky[3], Ren Sun[2], Eleazar Eskin[1]

[1]Computer Science, Human Genetics, University of California, Los Angeles
[2]Molecular & Medical Pharmacology, University of California, Los Angeles
[3] Computer Science, Georgia State University

Error rates in Next Generation Sequencing(NGS) data make it infeasible to apply this technology to solve quasispecies spectrum reconstruction problem. Here, we propose a novel sequencing library preparation technique that allow correcting error in NGS reads. Further, we introduce a method for inferring HIV quasispecies from high accurate paired-end reads. It consists of three key steps: (a) build consensus from reads, (b) map pe reads to consensus, (c) Build conflict graph and assemble quasispecies.

De novo Assembly of RNA-seq Data Based on Exact Match

Chao Deng [1], Qingfeng Xing [1], Yu Lu [1], Yuming Kuang [1], Quan Wang [2], Jie Ren [1], Ruibin Xi [1,3], Minping Qian [1,2], Minghua Deng [1,2,3*]

[1]: Affiliation 1. School of Mathematical Sciences, Peking University, China

[2]: Affiliation 2. Center for Quantitative Biology, Peking University, China

[3]: Affiliation 3. Center for Statistical Sciences, Peking University, China

[*]: To whom correspondence should be addressed.

Emails: DMH (dengmh@math.pku.edu.cn)

High-throughput RNA sequencing (RNA-seq) data provides a powerful protocol for the analysis of transcriptomes. Currently, a number of assemblers, including reference-based and de novo assemblers, have been developed to analyze RNA-seq data. De novo assemblers are advantageous in their capability to detect novel transcripts and provide an opportunity to find new isoforms and to estimate their abundance. However, because currently available de novo assemblers cannot correctly distinguish sequencing errors and biological variations implied in the data, they cannot achieve both good sensitivity and specificity at the same time. In addition, gene expression level can vary significantly across genes and the read abundance will accordingly vary vastly. This large variation makes it difficult for currently available assemblers to distinguish errors in highly expressed transcripts from truly low expressed transcripts.

In this paper, we propose a new algorithm, Alignment Free Assembly Method (AFAM), that can achieve good sensitivity and specificity simultaneously. This assembler is based on the de Bruijn graph, which takes exact match for local connection and combines with artificial mate (pairs of k- mers in the same read) to distinguish different branches in the de Bruijn graph. Furthermore, we set multiple thresholds for k-mer occurrence in reads to address the fact that there are more errors in high abundance regions than low abundance regions. Comparison with popular de novo assemblers shows that AFAM performs better in both sensitivity and specificity than other assemblers at a variety of sequencing depth. We also compared the performance of AFAM and other assemblers on a single cell sequencing data of mouse blastomere. We found that AFAM predicted similar number of new exon-exon junctions but with the lowest error rate.

An efficient random overlapping pool design for next generation sequencing based rare variant identification

Chang-Chang Cao, Xiao Sun*

State Key Laboratory of Bioelectronics, School of Biological Science and Medical Engineering, Southeast University, Nanjing, China.

*: To whom correspondence should be addressed.

Emails: Chang-Chang Cao (caochch@gmail.com); Xiao Sun (xsun@seu.edu.cn);

Motivation: Identification of rare variants through large scale resequencing is important for understanding complex disease. Population studies that involve sequencing thousands of individual genomes for charactering rare variants are still unaffordable until now, regardless of the drop in the price of next generation sequencing (NGS). Nevertheless group testing (GT) based overlapping pool design which greatly reduces sequencing pools to identify all the individuals helps to solve this problem.

Results: Here, taking advantage of quantitative information in sequencing results, we propose a random overlapping pool design algorithm that enables efficient recovery of variant carriers in groups of individuals with less cost. First of all, the optimal depths of coverage for pooled sequencing are computed based on a mathematic model. Random k-set pool design is used with appropriate selected parameters to guarantee the efficiency. Utilizing the information of reads number, we design a fast heuristic probability decoding algorithm to classify variant carriers. The results of simulation experiments with DNA preparation bias and sequencing errors indicate that our method performs similar or better compared with other previous algorithms in all aspects, including the amount of DNA libraries, data requirement and mixing procedure.

References

1. Shental, N., Amir, A. & Zuk, O. Identification of rare alleles and their carriers using compressed se (que) nsing. *Nucleic Acids Research* **38**, e179-e179, 2010.
2. Erlich, Y., Gordon, A., Brand, M., Hannon, G. J. & Mitra, P. P. Compressed genotyping. *Information Theory, IEEE Transactions on* **56**, 706-723, 2010.
3. Bruno, W. J. *et al.* Efficient pooling designs for library screening. *Genomics* **26**, 21-30, 1995.
4. Hwang, F. Random k-set pool designs with distinct columns. *Probability in the Engineering and Informational Sciences* **14**, 49-56, 2000.
5. Das, S. Binary Solutions for Overdetermined Systems of Linear Equations. *arXiv preprint arXiv:1101.3056* , 2011.

Bringing next generation sequencing to the clinic: Analytical validation and initial deployment of a comprehensive cancer genomic profiling test

Kai Wang[1], Garrett M. Frampton[1], Alex Fichtenholtz[1], Sean Downing[1], Jie He[1], Frank Juhn[1], Tina Brennan[1], Geoff Otto[1], Alex Parker[1], Vincent A. Miller[1], Jeffrey S Ross[1,2], John Curran[1], Philip J. Stephens[1], Doron Lipson[1], Roman Yelensky[1,*]

[1]Foundation Medicine, Cambridge, Massachusetts,

[2]Albany Medical College, Albany, NY

Emails: K Wang(kwang@foundationmedicine.com);R Yelensky(ryelensky@foundationmedicine.com)

Background: As the number of clinically relevant cancer genes increases, next-generation sequencing (NGS) is becoming an attractive diagnostic tool, as it can detect most genomic alterations in a single assay on limited tissue. However, rigorous analytical validation and comparison to current tests is required before clinical use.

Methods: We have developed an NGS-based test to characterize all classes of genomic alteration across 4,604 exons of 287 cancer-related genes from routine FFPE clinical specimens, including needle biopsies. To validate the test, we created reference samples reflecting key drivers of detection accuracy for somatic alterations across the targeted regions: For base substitutions, we mixed 2 pools of 10 normal cell-lines, testing 1,650 variants at allele frequencies (AF) 5%-100%. For indels, 28 tumor cell lines with 47 alterations 1-40bp long were used to generate 41 pools, testing 161 events at AF≥10%. For copy number alterations (CNAs), 7 tumor cell-lines with 19 gene amplifications and 9 homozygous gene deletions were pooled with their matched normal in 5 ratios with tumor content 20-75%. We confirmed accuracy in 308 FFPE tumors characterized for 95 alterations in 12 genes (e.g., EGFR, BRAF, HER2) by other assays, including PCR, mass-spec, FISH, and IHC. Precision was established on two control FFPE specimens processed a total of 150 times. We then assessed the potential clinical impact of comprehensive NGS by examining the nature and prevalence of genomic alterations revealed by the validated test in >2,000 consecutive patient specimens.

Results: On reference samples, sensitivity reached >99% (1,649/1,650) for base substitutions (AF≥5%), 98% (157/161) for indels (AF≥10%), >99% (56/56) for gene amplifications at CN≥8 and 97% (35/36) for homozygous deletions (tumor purity ≥30%), all with high specificity (PPV>99%). Robust performance translated to FFPE: concordance averaged 97% across subs/indels (109/113) and CNAs (41/42) relative to prior calls. All known alterations were called in all replicates of precision control specimens, including a base substitution present at only 4%. 2112/2221 (95%) of clinical specimens were successfully profiled (mean coverage 1134X), with 76% containing at least one alteration directly linked to a clinically available targeted treatment option or a mechanism-driven clinical trial. 963 unique such "actionable" alterations were reported, less than 1/3rd of which would have been identified by standard clinical tests, highlighting the advantages of a comprehensive NGS-based approach.

Conclusions: We present rigorous validation of a comprehensive NGS-based diagnostic test optimized for use in oncologic care and advocate that such genomic profiling be used to drive the appropriate use of targeted therapy and expand treatment choices for cancer patients.

Bellerophon: a hybrid method for detecting interchromosomal rearrangements at base pair resolution using next-generation sequencing data

Matthew Hayes and Jing Li

Department of Electrical Engineering and Computer Science, Case Western Reserve University, 10900 Euclid Ave., Cleveland, OH,
USA

Email: Matthew Hayes - matthew.hayes@case.edu; Jing Li - jingli@case.edu;

Somatically-acquired translocations may serve as important markers for assessing the cause and nature of diseases like cancer. Algorithms to locate translocations may use next-generation sequencing (NGS) platform data. However, paired-end strategies do not accurately predict precise translocation breakpoints, and "split-read" methods may lose sensitivity if a translocation boundary is not captured by many sequenced reads. To address these challenges, we have developed "Bellerophon", a method that uses discordant read pairs to identify potential translocations, and subsequently uses "soft-clipped" reads to predict the location of the precise breakpoints. Furthermore, for each chimeric breakpoint, our method attempts to classify it as a participant in an unbalanced translocation, balanced translocation, or interchromosomal insertion. We compared Bellerophon to four previously published algorithms for detecting structural variation (SV). Using two simulated datasets and two prostate cancer datasets, Bellerophon had overall better performance. Our method accurately predicted the presence of the interchromosomal insertions placed in our simulated dataset, which is an ability that the other SV prediction programs lack. The combined use of paired reads and soft-clipped reads allows Bellerophon to detect interchromosomal breakpoints with high sensitivity, while also mitigating losses in specificity. This trend is seen across all datasets examined. Because it does not perform assembly on soft-clipped subreads, Bellerophon may be limited in experiments where sequence read lengths are short.

A mixture model for expression deconvolution from RNA-seq in heterogeneous tissues

Yi Li [1], Xiaohui Xie [*1,2,3]

[1]: Department of Computer Science, University of California, Irvine, USA.
[2]: Institute for Genomics and Bioinformatics, University of California, Irvine, USA.
[3]: Center for Machine Learning and Intelligent Systems, University of California, Irvine, USA.
[*]: To whom correspondence should be addressed.
Emails: YL (yil8@uci.edu); XX (xhx@ics.uci.edu)

Background: RNA-seq, a next-generation sequencing based method for transcriptome analysis, is rapidly emerging as the method of choice for comprehensive transcript abundance estimation. The accuracy of RNA-seq can be highly impacted by the purity of samples. A prominent, outstanding problem in RNA-seq is how to estimate transcript abundances in heterogeneous tissues, where a sample is composed of more than one cell type and the inhomo-geneity can substantially confound the transcript abundance estimation of each individual cell type. Although experimental methods have been proposed to dissect multiple distinct cell types, computationally "deconvoluting" heterogeneous tissues provides an attractive alternative, since it keeps the tissue sample as well as the subsequent molecular content yield intact.

Results: Here we propose a probabilistic model-based approach, Transcript Estimation from Mixed Tissue samples (TEMT), to estimate the transcript abundances of each cell type of interest from RNA-seq data of heterogeneous tissue samples. TEMT incorporates positional and sequence-specific biases, and its online EM algorithm only requires a runtime proportional to the data size and a small constant memory. We test the proposed method on both simulation data and recently released ENCODE data, and show that TEMT significantly outperforms current state-of-the-art methods that do not take tissue heterogeneity into account. Currently, TEMT only resolves the tissue heterogeneity resulting from two cell types, but it can be extended to handle tissue heterogeneity resulting from multi cell types. TEMT is written in python, and is freely available at https://github.com/xhxielab/TEMT.

Conclusions: The probabilistic model-based approach proposed here provides a new method for analyzing RNA-seq data from heterogeneous tissue samples. By applying the method to both simulation data and ENCODE data, we show that explicitly accounting for tissue heterogeneity can significantly improve the accuracy of transcript abundance estimation.

Keywords: RNA-seq, Tissue Heterogeneity, Mixture Model, Online Expectation-Maximization, Positional Bias, Sequence-specific Bias, ENCODE

Nonsynonymous substitution rate (Ka) is a relatively consistent parameter for defining fast-evolving and slow-evolving protein-coding genes

Dapeng Wang[*,1,2], Fei Liu[*,1,2], Lei Wang[1,2], Shi Huang[3] and Jun Yu[§,1]

[1]: CAS Key Laboratory of Genome Sciences and Information, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 100029, PR China

[2]: Graduate University of Chinese Academy of Sciences, Beijing 100049, PR China

[3]: State Key Laboratory of Medical Genetics, Xiangya Medical School, Central South University, Changsha, Hunan 410078, PR China

*: These authors contributed equally to this work.

[§]: To whom correspondence should be addressed.

Emails: **DPW** (wangdp@big.ac.cn); **JY** (junyu@big.ac.cn)

Mammalian genome sequence data are being acquired in large quantities and at enormous speeds. We now have a tremendous opportunity to better understand which genes are the most variable or conserved, and what their particular functions and evolutionary dynamics are, through comparative genomics. We chose human and eleven other high-coverage mammalian genome data – as well as an avian genome as an outgroup – to analyze orthologous protein-coding genes using nonsynonymous (Ka) and synonymous (Ks) substitution rates. After evaluating eight commonly-used methods of Ka and Ks calculation, we observed that these methods yielded a nearly uniform result when estimating Ka, but not Ks (or Ka/Ks). When sorting genes based on Ka, we noticed that fast-evolving and slow-evolving genes often belonged to different functional classes, with respect to species-specificity and lineage-specificity. In particular, we identified two functional classes of genes in the acquired immune system. Fast-evolving genes coded for signal-transducing proteins, such as receptors, ligands, cytokines, and CDs (cluster of differentiation, mostly surface proteins), whereas the slow-evolving genes were for function-modulating proteins, such as kinases and adaptor proteins. In addition, among slow-evolving genes that had functions related to the central nervous system, neurodegenerative disease-related pathways were enriched significantly in most mammalian species. We also confirmed that gene expression was negatively correlated with evolution rate, i.e. slow-evolving genes were expressed at higher levels than fast-evolving genes. Our results indicated that the functional specializations of the three major mammalian clades were: sensory perception and oncogenesis in primates, reproduction and hormone regulation in large mammals, and immunity and angiotensin in rodents. Our study suggests that Ka calculation, which is less biased compared to Ks and Ka/Ks, can be used as a parameter to sort gene s by evolution rate and can also provide a way to categorize common protein functions and define their interaction networks, either pair-wise or in defined lineage s or subgroups. Evaluating gene evolution based on Ka and Ks calculations can be done with large datasets, such as mammalian genomes.

Reference

Dapeng Wang, Fei Liu, Lei Wang, Shi Huang and Jun Yu. **Nonsynonymous substitution rate (Ka) is a relatively consistent parameter for defining fast-evolving and slow-evolving protein-coding genes**. *Biology Direct* 2011, 6:13.

The Rice Genome Knowledgebase (RGKbase): an annotation database for rice comparative genomics and evolutionary biology

Dapeng Wang[*,1], Yan Xia[*,1,2], Xinna Li[1], Lixia Hou[1] and Jun Yu[§,1]

[1]: CAS Key Laboratory of Genome Sciences and Information, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 100029, PR China

[2]: Graduate University of Chinese Academy of Sciences, Beijing 100049, PR China

*: These authors contributed equally to this work.

[§]: To whom correspondence should be addressed.

Emails: **DPW** (wangdp@big.ac.cn); **JY** (junyu@big.ac.cn)

Over the past 10 years, genomes of cultivated rice cultivars and their wild counterparts have been sequenced although most efforts are focused on genome assembly and annotation of two major cultivated rice (*Oryza sativa* L.) subspecies, 93-11 (*indica*) and Nipponbare (*japonica*). To integrate in-formation from genome assemblies and annotations for better analysis and application, we now intro-duce a comparative rice genome database, the Rice Genome Knowledgebase (RGKbase, http://rgkbase.big.ac.cn/RGKbase/). RGKbase is built to have three major components: (i) integrated data curation for rice genomics and molecular biology, which includes genome sequence assemblies, transcriptomic and epigenomic data, genetic variations, quantitative trait loci (QTLs) and the relevant literature; (ii) User-friendly viewers, such as Gbrowse, GeneBrowse and Circos, for genome annotations and evolutionary dynamics and (iii) Bioinformatic tools for compositional and synteny analyses, gene family classifications, gene ontology terms and pathways and gene co-expression networks. RGKbase current includes data from five rice cultivars and species: Nipponbare (*japonica*), 93-11 (*indica*), PA64s (*indica*), the African rice (*Oryza glaberrima*) and a wild rice species (*Oryza brachyantha*). We are also constantly introducing new datasets from variety of public efforts, such as two recent releases—sequence data from 1000 rice varieties, which are mapped into the reference genome, yielding ample high-quality single-nucleotide polymorphisms and insertions–deletions.

Reference

Dapeng Wang, Yan Xia, Xinna Li, Lixia Hou and Jun Yu. **The Rice Genome Knowledgebase (RGKbase): an annotation database for rice comparative genomics and evolutionary biology**. *Nucleic Acids Researc*h. 2013; 41: D1199-D1205.

To detect mechanism of gene expression regulation using cell cycle synchronization and long-range DNA interaction

Yue Zhao[1], Yanjian Li[1], Yang Chen[1], Juntao Gao[1], Michael Q. Zhang[1,2*]

[1]: MOE Key Laboratory of Bioinformatics and Bioinformatics Division, Center for Synthetic and Systems Biology, TNLIST, Tsinghua University, Beijing 100084, China;

[2]: Department of Molecular and Cell Biology, Center for Systems Biology, The University of Texas at Dallas, Richardson, TX 75080, USA

[*]: Correspondence author.

Emails: Yue Zhao (zhaoyue_0512@163.com); Yanjian Li (liyanjian@gmail.com); Yang Chen (ebox.yc@gmail.com); Juntao Gao (jtgao@biomed.tsinghua.edu.cn); Michael Q. Zhang (mzhang@cshl.edu)

Regulation of gene expression includes a wide range of mechanisms used by cells to increase or decrease the production of specific gene products, but the mechanism of regulation of gene transcription remains elusive. Promoters and distal elements such as enhancers are involved in looping interactions and spatial proximity important for gene regulation. Yeast cells have provided an excellent model system to investigate long range interaction in eukaryotic cells at different stages of one cell cycle.

Here, more than 90% of the cell population was arrested in G1/S by the block-and-release method: First we treated cells with 5μg/ml α-factor for 3h to stop cells at G1/S transition point, with cell enlarged, unbudded and having “schmoo” morphology; then cells were released in YPD medium prewarmed at 30°C to pass START point. With flow cytometry, at G1/S transition point (i.e. START point), we observed that most cells (>75% of cell population) showed a higher peak, indicating 1N DNA content. A lower peak behind showed cells (<10% of cell population) with 2N DNA content. Cells (~15% of cell population) between two peaks were in S phase. As cells passed START point, more and more cells were getting larger and started budding.

Next we did ChIA-PET (chromatin interaction analysis by paired-end tag sequencing) to provide a genome-wide map of chromatin interactome bound by RNA Polymerase II at base-pair resolution. Superresolution microscope (STORM/PALM) imaging will help us visualize and confirm these long-range interactions. Promoters and distal elements are engaged in multiple long-range interactions to form complex networks. We will next set up a math model to describe the complex relationship among these regulatory elements based on ChIA-PET and ChIP-seq data, to understand the mechanism of regulation of gene transcription in eukaryotic cells in a better way.

Keywords: Long-range DNA interaction, budding yeast, synchronization, G1/S transition point, ChIA-PET

A (1.408+ε)-Approximation Algorithm for Sorting Unsigned Genomes by Reciprocal Translocations

Haitao Jiang [1,*] Lusheng Wang [2], Binhai Zhu [3,*] Daming Zhu [1]

[1]: School of Computer Science and Technology, Shandong University, Jinan, China.
[2]: Department of Computer Science, City University of Hong Kong, Kowloon, Hong Kong.
[3]: Department of Computer Science, Montana State University, Bozeman, MT 59717, USA.

[*]: To whom correspondence should be addressed.

Emails: HJ (htjiang@mail.sdu.edu.cn); LW (cswangl@cityu.edu.hk);

BZ (bhz@cs.montana.edu); DZ (dmzhu@sdu.edu.cn)

Sorting genomes by translocations is a classic combinatorial problem in genome rearrangements. The translocation distance for signed genomes can be computed exactly in polynomial time, but for unsigned genomes the problem becomes NP-hard and the current best approximation ratio is 1.5+ε. In this paper, we investigate the problem of sorting unsigned genomes by translocations by designing a better approximation algorithm. Firstly, we propose a tighter lower bound of the optimal solution by analyzing some special sub-permutations in the corresponding breakpoint graph; then, by exploiting the two well-known algorithms for approximating the maximum independent set on graphs with bounded degrees and for maximum set packing with sets of bounded size, we try to find (approximately) as many as 2-cycles and 3-cycles in the corresponding breakpoint graph. The numbers of these 2-cycles and 3-cycles can help us bound the approximation ratio. Lastly, we convert the corresponding problem instance into sorting signed genomes with translocations which can be solved in polynomial time. Putting all these together, we devise a new polynomial-time approximation algorithm, improving the approximation ratio to 1.408+ε, where ε=O(1/log n).

Hidden Markov Model based approaches to find function motifs of chromatin states

Tianying Zeng [1], Xiaowo Wang[1,*]

[1]:MOE Key Laboratory of Bioinformatics and Bioinformatics Div, Center for Synthetic and System Biology, TNLIST /Department of Automation, Tsinghua University, Beijing 100084, China.

[*]: To whom correspondence should be addressed.

Emails: TYZ (zengaaquila@gmail.com); XWW (xwwang@mail.tsinghua.edu.cn)

Chromatin states which are marked by different histone modifications and histone variants are key factors that regulate gene expression. Recently, as the ChIP-Seq technology has been widely applied to derive the histone modification profiles in different cell types, a mass of high throughput epigenetic data can be derived which enable researchers to discover the principles of epigenetic code at genome scale. It has been reported that the chromatin statuses are highly correlated with the underlying chromatin function, and histone modifications profiles could be used to predict where the functional chromatin domains like promoters or enhancers are located. But despite existing works, there is still large room for improvement in prediction accuracy and discovery of functional patterns in un-annotated genomic regions.

Here we proposed a new computational approach to find functional chromatin state motifs patterns. We reasoned that the function of a nucleosome is related to its neighbors, and functional chromatin regions could be defined as chromatin state motifs. First, we used a hidden markov model to compress the high-dimension histone modification profiles into one dimension markov state chain. Then we introduced supervised and unsupervised machine learning methods to see whether there are motifs significantly related to known genome functions or enriched in unmarked positions for further analysis. Finally we got key patterns of state which could be used to predict functional chromatin regions in annotated and un-annotated genomic areas.

Key words: epigenetic; hidden markov model; chromatin state motif discovery;

Gene order of an ancestral polyploid inferred from fractionated descendant genomes by sorting consolidated intervals, and sorting within intervals

Chunfang Zheng, David Sankoff *

Department of Mathematics and Statistics, University of Ottawa, Ottawa, Canada .

*: To whom correspondence should be addressed.

Emails: cz (czhen033@uottawa.ca); DS (sankoff@uottawa.ca)

With the fixation of a whole genome duplication (WGD) in a species, two very different types of mechanism may operate simultaneously to scramble gene order on the chromosomes. One consists of rearrangement events, notably inversion, reciprocal translocation, and chromosome fusions and fissions. The other is duplicate gene loss on a massive scale, affecting both members of each pair of homeologous chromosomes or regions, a process called fractionation.

The algorithmic study of rearrangements in genomes containing single copies of each gene has been extended to allow multiple gene copies and gene insertion and deletion. However, it is misleading to analyze the results of fractionation in terms of parsimonious combinations of rearrangement, insertion and deletion events, because these yield systematically biased results.

We show how to analyze fractionated genomes in three steps. The first is through a consolidation algorithm, with finds common (but incomplete) intervals in all the descendant genomes which descend from identical in- tervals in the original polyploid. The entire genomes are partitioned into such intervals. The second step is to reconstruct an ancestor using any of the standard techniques but where the genes are replaced by consolidated intervals. Finally each interval is sorted internally.

A simulation study shows that taking account of fractionation, as distinct from an insertion/deletion approach, produces a more accurate account of gene order evolution.

We apply our methods to reconstruct the gene order of a number of ancestral flowering plants, specifically the cereals, and the core eudicots.

References

1. Katharina Jahn, Chunfang Zheng, Jakub Kováč, David Sankoff. A consolidation algorithm for genomes fractionated after higher order polyploidization. *BMC Bioinformatics 13*, S19:S8, 2012.
2. Chunfang Zheng, David Sankoff. Fractionation, rearrangement and subgenome dominance. *Bioinformatics 28*, i402-i408, 2012.

Discovery of cell-type specific regulatory elements in the human genome using differential chromatin modification analysis

Chen Chen [1], Shihua Zhang [1,*], Xiang-Sun Zhang [1]

[1]: National Center for Mathematics and Interdisciplinary Sciences, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China.

[*]: To whom correspondence should be addressed.

Emails: CC (cchen@amss.ac.cn); SZ (zsh@amss.ac.cn); XSZ (zxs@amss.ac.cn)

Chromatin modifications have been comprehensively illustrated to play important roles in gene regulation and cell diversity in recent years. Given the rapid accumulation of genome-wide chromatin modification maps across multiple cell types, there is an urgent need for computational methods to analyze multiple maps to reveal combinatorial modification patterns and define functional DNA elements, especially those are specific to cell types or tissues. In this current study, we developed a computational method using differential chromatin modification analysis (dCMA) to identify cell-type specific genomic regions with distinctive chromatin modifications. We then apply this method to a public dataset with modification profiles of nine marks for nine cell types to evaluate its effectiveness. We found cell-type specific elements unique to each cell type investigated. These unique features show significant cell-type specific biological relevance and tend to be located within functional regulatory elements. These results demonstrate the power of a differential comparative epigenomic strategy in deciphering the human genome and characterizing cell specificity.

Eliminating nucleosome background from histone modification data

Jiao Chen[1], Yihua Zhu[1], Yumin Nie[1], Huan Huang[1] and Xiao Sun[1,*]

[1]: State Key Laboratory of Bioelectronics, School of Biological Science & Medical Engineering, Southeast University, Nanjing, China.

*: Corresponding author.

Emails: Jiao Chen (chjiao3456@gmail.com); Yihua Zhu (zhyh@njau.edu.cn); Yumin Nie (nieyum@126.com); Huan Huang (ahuanghuan@163.com); Xiao Sun* (xsun@seu.edu.cn)

Nucleosome positioning and histone modifications are very important epigenetic factors regulating gene expression. Recent high-throughput technologies, such as ChIP-seq and MNase-seq, which combine ChIP with next-generation sequencing technique, have generated genome-wide nucleosome position maps and histone modification distributions. Typical patterns of nucleosome positioning and histone modifications in the vicinity of transcription start sites (TSS), transcription termination sites (TTS) and transcription factor binding sites (TFBS) have also been studied. Existing data analyzing methods usually generate the average histone modification enrichment profiles by counting the number of shifted or extended tags that fall at each position along the region surrounding TSS, TTS or TFBS, without the control of underlying nucleosome occupancy. However, a genome region with low nucleosome occupancy but high histone modification level may be covered as many tags as the genome region with high nucleosome occupancy but low histone modification level. Here we develop a new method eliminating nucleosome background from histone modification data to generate the relative enrichment of histone modification tags surrounding TSS and TFBS.

Our results reveal that histone modification profiles after nucleosome background elimination exhibit distinct signal profiles comparing with original ones in the vicinity of TFBS and TSS. The depletion of most histone modification profiles at TFBS or TSS disappear in the relative tags enrichment profiles, indicating that the troughs of histone modification signal at these sites are mainly caused by nucleosome depletion. Besides, as the three lysine methylation states compete for the single lysine, H3K27 methylation signals after nucleosome background elimination exhibit higher and wider peaks at the TSSs as the modification moves from mono- to di- to trimethylation, suggesting that H3K27me2 and H3K27me3 may repress gene expression by their high occupancies at the TSSs.

In conclusion, our method, which tries to remove nucleosome background from histone modification profiles, provides a fresh perspective on histone modification data analysis.

Large Local Analysis of the Unaligned Genome and Its Application

Lianping Yang [1], Xiangde Zhang [1*], Tianming Wang [2], Hegui Zhu [1]

[1]: College of Sciences, Northeastern University, Shenyang, 110004, China.

[2]: School of Mathematical Sciences, Dalian University of Technology, Dalian 116024, China.

[*]: Corresponding author.

Emails: Yang L. (yangmath@yahoo.cn); Zhang X. (zhangxdmath@yahoo.cn); Wang T. (wangtm@dlut.edu.cn); Zhu H. (zhuhegui@126.com)

Relative feature methods, the most widely used method alignment method is one of them, are efficient and powerful tools in the field of sequence analysis. We describe a novel relative feature method for the local analysis of complete genomes. The method is based on the relationship between the longest common words and the shortest absent words of two genomes we compared. We find that the sum of all the lengths of the longest common words is an index reflecting the degree of the local segments belonging to the reference sequence even though the segments cover some gene arrangements. Based on that, a local distance measure called LODIST is proposed and it performs better than local alignment when the local region is large enough to cover some recombination genes. A distance measure called SILD.k.t with resolution k and step t is derived by the integral LODISTs of whole genomes. It is shown that the algorithm for computing the LODISTs and SILD.k.t is linear, which is fast enough to consider the problem of the genome comparison. We verify this method by recognizing the subtypes of the HIV-1 complete genomes and genome segments.

References

1. Lianping Yang, Xiangde Zhang, Tianming Wang, Hegui Zhu. Large Local Analysis of the Unaligned Genome and Its Application. *Journal of Computational Biology*, 20(1): 19-29, 2013.

MOABS: Model based analysis of bisulfite treated DNA methylation data

Deqiang Sun, Wei Li,

Division of Biostatistics, Dan L. Duncan Cancer Center
Department of Molecular and Cellular Biology
Baylor College of Medicine
One Baylor Plaza
Houston, TX 77030
deqiangs@bcm.edu

5-methylcytosine and 5-hydroxymethylcytosine can now be quantitatively measured at base level by whole genome bisulfite sequencing. However, lack of complete and accurate methods describing and utilizing digital methylation information from single base to region level, and lack of accurate and fast analysis pipeline are still two major challenges. They are now solved by MOABS, a complete, accurate and efficient solution for methylation data analysis. It seamlessly integrates alignment, methylation calling, identification of hypomethylation for one sample and differential methylation for multiple samples, and other downstream analysis. We show that it is aware of replicate reproducibility and accurate even at low coverage. It uses advanced algorithms and efficiently utilizes threads and clusters so that 2 billion aligned reads from two conditions can be processed lightening fast in 1 hour (vs more than 1 day by other pipelines) analyzing methylation on around 30 million CpGs.

Regulation of differentiation in atrial and ventricular myocytes

Zheng-Yu Liang[1], Monica C. Sleumer[1], Pu Li[2], Juntao Gao[1], Yue Ma[2,*], Michael Q. Zhang[1,3,*]

[1]: MOE Key Laboratory of Bioinformatics and Medicine School, Center for Synthetic and Systems Biology, TNLIST, Tsinghua University, Beijing 100084, China;

[2]: National Laboratory of Biomacromolecules, Institute of Biophysics, Chinese Academy of Sciences, 15 Datun Rd., Chaoyang District, Beijing 100101, China;

[3]: Department of Molecular and Cell Biology, Center for Systems Biology, The University of Texas at Dallas, Richardson, TX 75080, USA

[*]: Correspondence author.

Emails: Zheng-Yu L (zyliang.cn@gmail.com); Monica S (msleumer@mail.tsinghua.edu.cn); Pu L (puli8605@gmail.com); Juntao G (jtgao@biomed.tsinghua.edu.cn); Yue M (yuema@ibp.ac.cn); Michael Z (mzhang@cshl.edu)

Transcription factors play an important role during differentiation from human embryonic stem cells (hESCs) to cardiomyocyte. We earlier determined that retinoid signaling regulates the direction of cardiac differentiation to atrial and ventricular myocytes.

We have identified eighteen transcription factors important for cardiomyocyte development. We performed a literature search and generated a regulatory network showing how these signaling molecules and transcription factors regulate each other. We further found evidence for six protein-protein interactions between them in the BioGrid database.

Thereafter we measured gene expression levels of differentiated cardiomyocyte using microarrays and found 177 highly expressed genes in ventricular myocytes and 195 highly expressed genes in atrial myocytes. We discovered ten genes that are highly expressed in ventricular myocytes but completely suppressed by retinoid signaling and mildly suppressed by bone morphogenic protein (BMP) signaling.

Then we constructed a model of regulation to illustrate the regulation process. These findings demonstrate that retinoid signaling and BMP signaling function complementarily to specify atrial versus ventricular myocytes differentiation from hESCs.

Keywords: cardiomyocyte differentiation, gene regulation, expression pattern

References

1. Zhang, Q, Jiang J, Ma Y.et al. (2011). "Direct differentiation of atrial and ventricular myocytes from human embryonic stem cells by alternating retinoid signals." Cell Res 21(4): 579-587.

Using weakly conserved motifs hidden in secretion signals to identify type-III effectors from bacterial pathogen genomes

Xiaobao Dong [1*], Ziding Zhang [1*]

[1]: State Key Laboratory of Agrobiotechnology, College of Biological Sciences, China Agricultural University

[*]: To whom correspondence should be addressed.

Emails: X.D. (yndxb@126.com); Z.Z.(zidingzhang@cau.edu.cn);

Bacterial type-III effectors (TTEs) play a crucial role in pathogen-host interactions by directly influencing immune signaling pathways within host cells. Based on the hypothesis that type-III secretion signals may be comprised of some weakly conserved sequence motifs, here we used profile-based amino acid pair information to develop an accurate TTE predictor.

For a TTE or non-TTE, we first used a hidden Markov model-based sequence searching method to detect its weakly homologous sequences and extracted the profile-based k-spaced amino acid pair composition (HH-CKSAAP) from the N-terminal sequences. In the next step, the feature vector HH-CKSAAP was used to train a linear support vector machine model, which we designate as BEAN (Bacterial Effector ANalyzer). We compared our method with four state-of-the-art TTE predictors through an independent test set, and our method revealed improved performance. Furthermore, we listed the most predictive amino acid pairs according to their weights in the established classification model. Evolutionary analysis shows that predictive amino acid pairs tend to be more conserved. Some predictive amino acid pairs also show significantly different position distributions between TTEs and non-TTEs. These analyses confirmed that some weakly conserved sequence motifs may play important roles in type-III secretion signals. The webserver and stand-alone version of BEAN are available at http://protein.cau.edu.cn:8080/bean/.

References

1. Xiaobao Dong, Yong-Jun Zhang, Ziding Zhang. Using Weakly Conserved Motifs Hidden in Secretion Signals to Identify Type-III Effectors from Bacterial Pathogen Genomes. *PLoS ONE*, 8(2): e56632, 2013

A Mosaic of Transcriptional Fingerprints

Patrick Kemmeren[1,*], Katrin Sameith[1], Loes van de Pasch[1], Joris Benschop[1], Tineke Lenstra[1], Thanasis Margaritis[1], Tony Miles[1], Mariel Brok[1], Nathalie Brabers[1], Eoghan O'Duibhir[1], Sake van Wageningen[1], Dik van Leenen[1], Cheu Ko[1], Eva Apweiler[1], Sander van Hooff[1], Philip Lijnzaad[1], Marian Groot Koerkamp[1], Frank Holstege[1,*]

[1]: Holstege Lab, UMC Utrecht, Department of Molecular Cancer Research, Universiteitsweg 100, 3584 CG Utrecht, The Netherlands.

[*]: To whom correspondence should be addressed.

Emails: PK (p.kemmeren@umcutrecht.nl); FH (f.c.p.holstege@umcutrecht.nl)

Regulation of gene expression is pivotal for most biological processes. Through an intricate interplay between signaling pathways and the transcription machinery, cells are able to cope with a plethora of environmental cues and internal processes. Here, we have collected 1,484 gene deletion expression profiles in Saccharomyces cerevisiae, including protein kinases, protein phosphatases, components of ubiquitin and ubiquitin-like pathways, gene-specific transcription factors, chromatin modifiers, general transcription components and glucose sensing pathways. Of the 1,484 deletion mutants, 700 have an expression profile different from wildtype. Hierarchical clustering of the gene deletion expression profiles largely groups mutants together according to their protein complex or pathway membership. Clustering of the transcript profiles, groups the transcripts in biologically coherent clusters that in many cases correspond well with the binding sites of the known transcription factors. The degree to which genes affect each other transcriptionally is further investigated by constructing the genetic perturbation network. From this genetic perturbation network, significant recurrent network motifs are extracted and analyzed, revealing a metabolic regulatory circuit amongst others. By combining the genetic perturbation network with other functional genomics data, regulatory entry points for protein complexes can be pinpointed. Additional analyses aimed at further elucidating the transcriptional regulatory network are currently being pursued.

COUGER - a new framework for identifying co-factors associated with uniquely-bound genomic regions

Alina Munteanu [1,*], Raluca Gordân [2,*]

[1]: Faculty of Computer Science, Alexandru I. Cuza University, Iasi, Romania

[2]: Institute for Genome Sciences and Policy, Departments of Biostatistics & Bioinformatics, Computer Science, and Molecular Genetics and Microbiology, Duke University, Durham, NC 27708, USA

[*]: To whom correspondence should be addressed.

Emails: AM (alina.munteanu@info.uaic.ro); RG (raluca.gordan@duke.edu)

Transcription factors (TFs) regulate gene expression by binding to specific DNA sites in cis regulatory regions of genes. Most eukaryotic TFs are members of protein families that share a common DNA binding domain and have highly similar DNA binding preferences. However, individual TF family members (paralogous TFs) often have different functions and bind to different genomic regions *in vivo*. A potential mechanism for achieving regulatory specificity is through interactions with proteins co-factors.

We present **COUGER**, a general framework for identifying putative co-factors that provide specificity to paralogous TFs. Our framework uses state-of-the-art classification algorithms (support vector machines and random forests) with features that reflect the DNA binding specificities of putative co-factors. The features are generated either from high-throughput TF-DNA binding data (from protein binding microarray experiments), or from largely available DNA motif data. Our features take into account the fact that TF binding sites may occur in clusters [1].

COUGER can be applied to any two sets of genomic regions bound by paralogous TFs (e.g., regions derived from ChIP-seq experiments). The framework determines the genomic targets uniquely-bound by each paralogous TF, and identifies a small set of co-factors that best explain the genomic binding differences.

References

1. Alina Munteanu, Raluca Gordân. Distinguishing between Genomic Regions Bound by Paralogous Transcription Factors. *Lecture Notes in Bioinformatics 7821*, p.145, 2013.

An MRF-based Method for lncRNA Function Prediction

Xingli Guo[1], Lin Gao[1,*], Yongxuan Liu[1], Bingbo Wang[1]

[1]: School of computer science and technology, XiDian University, No.2 Taibai South Road, Xi'an, 710071, Shaanxi, China

[*]: Corresponding author, Lin Gao (lgao@mail.xidian.edu.cn)

Emails: XL Guo (xlguo@mail.xidian.edu.cn); YX Liu (lyxuk318@126.com);

BB Wang (w_bingbo@163.com)

Many genome-wide sequencing projects have identified vast amount of long non-coding RNAs (lncRNAs) in eukaryotic genomes. Further studies confirmed the key roles of lncRNAs in many biological processes. Only few of them have been functionally characterized, and to perform large-scale function annotation for lncRNAs is necessary and urgent. challenge. The network-based methods for lncRNA function prediction have been shown promising to tackle this challenge. Here we try to exploit a Markov random field (MRF) based method, which outperformed other methods in the task of protein function prediction, for large-scale function prediction of lncRNAs on a coding-non-coding bi-colored network. First, for each individual function annotation, a model representing the distribution of this function annotation in the bi-colored network is constructed based on the Markov random field theory. Then, a Gibbs sampler derived from the model is applied to predict probable functions for lncRNAs. To get a better performance, we restrict the application of MRF-method for those lncRNAs with at least one functionally annotated neighbor in the network. Altogether, there are 571 lncRNAs functionally characterized by our method. The presented MRF-based method can achieve a good performance with proper parameter tuning. Furthermore, the function annotation for lncRNAs are consistent with that by a newly published method—*lnc-GFP*. Especially, other novel and more specific functions are assigned to the lncRNAs, which are confirmed by other individual studies. As a conclusion, our MRF-method provided more experiences for lncRNA function annotation task, and can give more accurate results for well connected lncRNAs in the network.

Kernel-based method for measuring distance between RNA structures

Hanjoo Kim [1] and Sungroh Yoon [1,2,*]

[1]: Electrical and Computer Engineering, Seoul National University, Seoul 151-744, Korea.

[2]: Bioinformatics Institute, Seoul National University, Seoul 151-747, Korea.

[*]: To whom correspondence should be addressed.

Emails: H. K. (kprotoss@acl.snu.ac.kr); S. Y. (sryoon@snu.ac.kr)

The function of an RNA molecule is closely related to its structure, and comparative analysis of RNA structures can be useful to infer functions of novel RNAs. In such analysis, it is critical to have a means to measure distances between different RNA molecules, since pairwise distance is often the most fundamental information many data-analysis methods relies on. In this work, we propose a kernel-based approach to measuring distances between RNA structures. In this methodology, we first represent each RNA structure under study in an intermediate representation. The distance between two RNA structures then becomes the distance between their intermediate representations. To measure their distance, we use a special type of the Mercer kernel, effectively bypassing difficulties in handling RNAs of different lengths and structures. Using real RNA test data, we carry out data-mining tasks such as clustering to test the effectiveness of our approach. We also compare the proposed measure with the conventional edit distance metric for RNA structure comparison in terms of the fidelity to retrieving the labels of the test data.

Differential Methylation in t(8;21) AML and its association with AML1-ETO fusion protein binding profile

Zhirui HU [1,#], Xiaoning GAO [2,#], Yonghui LI [2], Yang CHEN [1], Li YU [2,*], Michael Q ZHANG [3,*]

[1]: MOE Key Laboratory of Bioinformatics and Bioinformatics Div, TNLIST / Department of Automation, Tsinghua University, Beijing 100084, China.

[2]: Department of Hematology and BMT Center, Chinese PLA General Hospital, Postgraduate Medical School, 28 Fuxing Road, Beijing 100853, China.

[3]: Department of Molecular and Cell Biology, Center of Systems Biology, The University of Texas at Dallas, TX 75080-3021, USA.

[*]: To whom correspondence should be addressed.

[#]: These authors contributed equally to this work.

Emails: ZRH (hzr07@mails.tsinghua.edu.cn); XNG (gaoxn@263.net); YHL (yonghuililab@yahoo.com.cn); YC (ebox.yc@gmail.com); LY (chunhuiliyu@yahoo.com); MQZ (mzhang@cshl.edu)

t(8;21) translocation, resulting in a fusion between transcription factor RUNX1 and an transcriptional co-repressor ETO, defines a distinct subtype of AML. Some studies demonstrated that the fusion protein recruits histone deacetylase complexes, while others said it would also activate genes. On the other hand, some studies have shown that DNA methylation can identify different subtypes of AML. To further investigate how RUNX1-ETO fusion protein (AE) alters epigenetic landscape and initiates leukemia, we measured DNA methylation of AML M2 patients with AE fusion and with normal cytogenetic and found distinct DNA methylation profile of AE-positive patients. Then, we integrated AE binding, RUNX1 binding, P300 binding and H3K9ac profile with DNA methylation, distinguished different binding patterns and built a Bayesian network to show how these components associate with each other leading to differential gene expression and malignancy in t(8;21) AML cells. The Bayesian network indicates that AE co-occupies with many other TFs, like HEB and FLT1, and AE binding affects H3K9ac through other proteins. We showed that AE plays a complex role in the leukemia cell, as we identified distinct patterns of AE binding promoters with respect to different methylation profiles. Globally, AE binding only slightly represses nearby gene expression, but AE binding influences more at less active chromatin site.

Discovering microRNA genes from insect transcriptome data

Ying Liu, Fei Li *

Department of entomology, Nanjing Agricultural University, Nanjing, 210095, China

*: To whom correspondence should be addressed.

Emails: Ying Liu (lying1125@gmail.com); Fei Li (lifei03@tsinghua.org.cn)

MicroRNAs (miRNAs) are small noncoding regulatory RNAs that play key roles in many diverse biological processes. With rapid development of next generation sequencing technique, lots of transcriptome data have been reported. Most analyses of transcriptome focus on the protein coding genes. However, only half of the assembled contigs in transcriptome can be annotated as protein coding transcripts. For those non-coding transcripts, they are either un-translated region (UTR) of mRNA or the transcripts of noncoding genes. Here we developed a pipeline to detect miRNA gene from insect transcriptome. The raw reads of transcriptome from brown planthopper *Nilaparvata lugens* were obtained from the SRA database. We re-assembled the transcriptome with Trinity, yielding 36,748 contigs. 20,674 contigs were annotated as protein coding genes and 16,074 contigs were treated as the noncoding contigs. We downloaded 16,495 mature miRNAs from the miRbase and mapped them with noncoding contigs. Two mismatches were allowed for aligning known insect nature miRNAs with contigs, whereas three mismatches were allowed for alignment of non-insect mature miRNAs with contigs. The putative miRNA precursors were obtained by extracting two ~200 bp sequences flanking the mapped regions in contigs (+10 upstream, ~20 bp mapped region and -160 downstream). The second structures were predicted with RNAfold and the maximum free energies were calculated. We kept the sequences with stable stem-loop secondary structures (MFE $\leq$-25 kcal/mol) and the stem part in the hairpin structure should be more than 22 nt. We then used triplet-SVM to find true miRNA precursors, producing 19 miRNA genes. Finding miRNAs from the transcriptome provides useful information to study mRNA-miRNA interactions since they the identified miRNAs and mRNAs are from a same sample.

Discovering Frequent Transcription Factor Interactions in Cis-Regulatory Module

Li Teng [1], Bing He [3], Kai Tan [1,2,3*]

[1]: Department of Internal Medicine, University of Iowa, Iowa City, IA 52242, USA.

[2]: Department of Biomedical Engineering, University of Iowa, Iowa City, IA, USA.

[3]: Interdisciplinary Graduate Program in Genetics, University of Iowa, Iowa City, IA, USA.

[*]: To whom correspondence should be addressed.

Emails: LT (li-teng@uiowa.edu); BH (bing-he@uiowa.edu); KT (kai-tan@uiowa.edu)

Enhancers are usually characterized by multiple transcription factor (TF) bindings in concert to target genomics DNA. In recent years, chromatin immunoprecipitation with massively parallel sequencing (ChIP-Seq) has become method of choice for genome-wide detection of the in vivo binding locations for individual TFs at high resolution. The wealth of data generated by these high-throughput experiments provides us with an opportunity to answer questions left open by previous analyses.

In this analysis we used a total of 108 different TFs that are currently available in ENCODE [1]. Genome-wide enhancer predictions were made for four human cell lines (GM12787, K562, HepG2 and H1) using CSI-ANN, a tool developed in our lab [2-3] that predicts enhancers based on histone modification marks. Confidence levels of TF binding and enhancer prediction were used in this analysis to construct a TF-enhancer binding matrix with uncertainty on both dimensions. Using frequent itemsets mining on uncertain data, we investigated combinatorial TF interactions in enhancer regions across the four human cells.

Our analyses show that the proposed method outperforms its predecessor by finding more TF interactions that are supported by known protein protein interaction data. We found that in different cell types the TF patterns are supported by enhancers close to genes enriched with very different GO terms. In addition, TFs that are involved in frequent combinatorial patterns have higher correlated expression profiles. We identified characteristic binding site spacing between TFs and binding orders of TFs in enhancers.

References

1. The ENCODE Project Consortium. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, 447(7146):799-816, 2007.
2. Hiram A. Firpi, Duygu Ucar and Kai Tan. Discover regulatory DNA elements using chromatin signatures and artificial neural network. *Bioinformatics*, 26(13):1579-86, 2010.
3. Li Teng, Hiram A. Firpi and Kai Tan. Enhancers in embryonic stem cells are enriched for transposable elements and genetic variations associated with cancers. *Nucleic Acids Res* 39(17):7371-9, 2011.

A Computational Prediction System for Identifying Human microRNA Target Sites

Ki-Bong Kim[1], Kiejung Park [2,*]

[1]: Dept. of Biomedical Technology, Sangmyung Univerisity.

[2]: Korean Bioinformation Center, Korea Research Institute of Bioscience & Biotechnology.

[*]: To whom correspondence should be addressed.

Emails: KB (kbkim@smu.ac.kr); KP (kjpark@kribb.re.kr)

MicroRNAs (miRNAs) are important regulators of gene expression and play crucial roles in many biological processes including apoptosis, differentiation, development, and tumorigenesis. Recent estimates suggest that more than 50% of human protein coding genes may be regulated by miRNAs and that each miRNA may bind to 300-400 target genes. Approximately 1,000 human miRNAs have been identified so far with each having up to hundreds of unique target mRNAs. However, the targets for a majority of these miRNAs have not been identified due to the lack of large-scale experimental detection techniques. Experimental detection of miRNA target sites is a costly and time-consuming process, even though identification of miRNA targets is critical to unraveling their functions in various biological processes. To identify miRNA targets, we developed *miRTar Hunter*, a novel computational approach for predicting target sites regardless of the presence or absence of a seed match or evolutionary sequence conservation. Our approach is based on a dynamic programming algorithm that incorporates more sequence-specific features and reflects the properties of various types of target sites that determine diverse aspects of complementarities between miRNAs and their targets. We evaluated the performance of our algorithm on 532 known human miRNA:target pairs and 59 experimentally-verified negative miRNA:target pairs, and also compared our method with three popular programs for 481 miRNA:target pairs. *miRTar Hunter* outperformed three popular existing algorithms in terms of recall and precision, indicating that our unique scheme to quantify the determinants of complementary sites is effective at detecting miRNA targets.

References

1. D.P. Bartel. MicroRNAs: target recognition and regulatory functions. *Cell*, 136: 215-233, 2009.
2. D. Betel, M. Wilson, A. Gabow, D.S. Marks, C. Sander. The microRNA.org resource: targets and expression. *Nucleic Acids Res*., 36: D149-153, 2008.
3. I. Hofacker. Vienna RNA secondary structure server. *Nucleic Acids Res.*, 31: 3429-3431, 2003.
4. P. Sethupathy, M. Megraw, A. G. Hatzigeorgiou. A guide through present computational approaches for the identification of mammalizan microRNA targets. *Nat. Methods*, 3: 881-886, 2006
5. M.S. Waterman, M. Eggert. A new algorithm for best subsequence alignments with application to tRNA-rRNA comparisons. *Journal of Mol. Biol.*, 197: 723-728, 1987.

Reference-gene-based normalization of microRNA expression data provides higher consistency in differential expression analysis

Xi Wang [1,2], Murray J. Cairns [1,2,3*]

[1]: School of Biomedical Sciences and Pharmacy, The University of Newcastle, Callaghan, New South Wales, Australia.

[2]: Hunter Medical Research Institute, New Lambton, New South Wales, Australia.

[3]: Schizophrenia Research Institute, Sydney, New South Wales, Australia.

[*]: To whom correspondence should be addressed.

Emails: XW (Xi.Wang@newcastle.edu.au); MC (Murray.Cairns@newcastle.edu.au)

Background: Normalization of microRNA (miRNA) expression profiles secures differential expression analysis between samples of different phenotypes or biological conditions, and facilitates comparison between experimental batches. There is mounting evidence that global shifts in miRNA expression patterns occur in specific circumstances, which pose a challenge for normalizing miRNA expression data. As an alternative to global normalization, which has the propensity to flatten large trends, normalization against constitutively expressed reference genes presents an advantage through their relative independence. This approach has been widely used in normalizing RT-qPCR expression data, and may be more consistent when large expression shifts exist in the biology of the samples.

Methods/Results: We investigated the performance of reference-gene-based normalization for differential miRNA expression analysis of microarray expression data, and compared the results with other normalization methods, including: quantile, variance stabilization, robust spline, simple scaling, rank invariant, and Loess regression. The comparative analyses were executed on miRNA expression data from peripheral blood mononuclear cells derived from a cohort of schizophrenia patients and non-psychiatric controls. We proposed a consistency criterion for evaluating methods by examining the overlapping of differentially expressed miRNAs detected using different partitions of the whole data. Based on this criterion, we found that reference-gene normalization generally outperformed other normalization methods regardless of the methods used for differential expression analysis.

Conclusions: We recommend the application of reference-gene-based normalization for miRNA expression data sets, and believe that this will yield a more consistent and useful readout of differentially expressed miRNAs, particularly in biological conditions characterized by large shifts in miRNA expression. We also believe this approach will aid the discovery and application of miRNA biomarkers and clinical diagnostics.

Impact of DNA Structure on Functional Regulatory Motifs

Qian Xiang

School of Information Science and Technology

Sun Yat-Sen University, Guangzhou, PR China

xiangq@mail.sysu.edu.cn

***Abstract*—**The three-dimensional structure of DNA has been proposed to be a major determinant for functional transcription factors (TFs) and DNA interaction, as it is a critical feature recognized by the regulatory machinery within a cell. Here we use hydroxyl radical cleavage pattern as a measure of local DNA structure, and the regulatory protein may recognize a variety of divergent nucleotide sequences that adopt the same local structures. We compared the conservation between DNA sequence and structure in terms of information content and attempted to assess the functional implications of DNA structures on regulatory motifs. We used statistical methods to evaluate the structural divergence of substituting a single position within a binding site and applied them to a collection of putative regulatory motifs. The following are our major observations: (i) We observed more information in structural alignment than the corresponding nucleotide sequence alignment for most of the transcriptional factors; (ii) For each TF, majority of positions have more information in the structural alignment as compared to the nucleotide sequence alignment; (iii) We further defined a DNA structural divergence score (SD-score) for each wild-type and mutant pair that are distinguished by single base mutation. The SD-score for benign mutations are significantly lower than that of switch mutations. This indicates structural conservation is important for TFBS to be functional. Based on these findings, we speculate that some of the functional information in the TFBS is conferred by DNA structure as well as by nucleotide sequence. DNA structures will provide previously unappreciated information for TF to realize the binding specificity. Our results should facilitate the prediction of the divergent functional TF regulatory interaction with binding site variations by altering the DNA structure.

***Keywords*—**DNA structure; transcriptional factor; regulatory motif; binding site mutation; structure divergence

High order intra-strand symmetry analysis between coding RNA and LncRNA

Shengqin Wang[1], Zuhong Lu[1*]
[1]:State Key Lab of Bioelectronics, School of Biological Science and Medical Engineering, Southeast University, Nanjing, 210096, China
[*]: To whom correspondence should be addressed.

Chargaff observed the equimolar frequencies of reverse complement nucleotide in one same DNA strand (Chargaff, 1979). Interesting, these intra-strand compositional symmetries are also found in dinucleotide and in higher order oligonucleotides(Qi & Cuticchia, 2001). Although still not much recognized, the phenomenon of these symmetries may provide clues to genome evolution. Here, we pooled data from two recent published papers (Chang et al., 2011; Volders et al., 2013), then quantify 8 mer symmetry as the increase of contiguous sequences for each RNA data. In order to ignore the order of pooled sequences, each data will be shuffled by 10 times. At last, we found the LncRNA has higher symmetry score than coding RNA at 8-mer. However, when we calculate the ratio of loop structure, LncRNA is less than coding RNA, which also proved that the contribution of stem-loop potential to compositional symmetry is limited(Zhang & Huang, 2010). Therefore, other possible causes for symmetry need to be taken. We also found that house keeping gene has a little higher than tissue specific gene, which may proved that evolution conserved gene has higher intra-strand symmetry.

Keywords: intra-strand symmetry, LncRNA, coding RNA

Chang, C.-W., Cheng, W.-C., Chen, C.-R., Shu, W.-Y., Tsai, M.-L., Huang, C.-L., & Hsu, I. C. (2011). Identification of human housekeeping genes and tissue-selective genes by microarray meta-analysis. *PLoS ONE*, *6*(7), e22859. doi:10.1371/journal.pone.0022859

Chargaff, E. (1979). How genetics got a chemical education. *Annals of the New York Academy of Sciences*, *325*, 344–360.

Qi, D., & Cuticchia, A. J. (2001). Compositional symmetries in complete genomes. *Bioinformatics (Oxford, England)*, *17*(6), 557–559.

Volders, P.-J., Helsens, K., Wang, X., Menten, B., Martens, L., Gevaert, K., Vandesompele, J., et al. (2013). LNCipedia: a database for annotated human lncRNA transcript sequences and structures. *Nucleic acids research*, *41*(Database issue), D246–51. doi:10.1093/nar/gks915

Zhang, S. H., & Huang, Y. Z. (2010). Limited contribution of stem-loop potential to symmetry of single-stranded genomic DNA. *Bioinformatics (Oxford, England)*, *26*(4), 478–485. doi:10.1093/bioinformatics/btp703

Shorter loop length regions show lower conservation score of stem region in Drosophila

Shengqin Wang[1], Zuhong Lu[1*]
[1]:State Key Lab of Bioelectronics, School of Biological Science and Medical Engineering, Southeast University, Nanjing, 210096, China
[*]: To whom correspondence should be addressed.

Functional residues usually have high conservation score, which measured from a multiple sequence alignment. Stem loop structure is a pattern that can commonly occur in function RNA, and the evolution information in stem region can be kept by paired nucleotide. Novel insert fragment will have lower conservation score by other genome sequence. The analysis of the development of these stem loop structure may provide clues to the evolution of DNA.
In this paper, we calculate the conservation score of stem region combine with loop length information in Drosophila. Our result shown shorter loop length regions have lower conservation score of stem region in all of four normal chromosomes. Previous stem-loop "kissing" model can explain the initiation of recombination and increase the length of genome size, however, this model will also destroy the stem structure. This phenomenon may give the insight that the increase of genome may also be originated from the increase of loop region by other mechanism. Further more, the stem loop structure often related to function RNA, which give us the hypothesis that the present genomes might be originated from the original RNA genes.

Keywords: stem loop, conservation score

Cloudbreak: A MapReduce Algorithm for Detecting Genomic Structural Variation

Christopher W. Whelan[1] and Kemal Sönmez[1,2]
[1]Institute on Development and Disability, Center for Spoken Language Understanding
[2]Department of Medical Informatics & Clinical Epidemiology
Oregon Health & Science University, Portland, OR, USA
cwhelan@gmail.com, sonmezk@ohsu.edu

The detection of genomic structural variations remains one of the the most difficult challenges in analyzing high-throughput sequencing data. Recent approaches have demonstrated that considering multiple mappings of all reads, rather than only uniquely mapped discordant fragments, can improve the performance of read-pair based detection methods. However, the computational requirements for storing and processing data sets with multiple mappings can be formidable. Meanwhile, the growing size and number of sequencing data sets have led to intense interest in distributing computation to cloud or commodity servers.

MapReduce, via its Hadoop implementation, is becoming a standard architecture for distributing processing across such compute clusters. In this work we describe a novel conceptual framework for structural variation detection in MapReduce/Hadoop based on computing local features along the genome. Our framework uses Hadoop to take advantage of distributed computing to find all possible read alignments using modern short-read aligners run with sensitive settings. We then provide an architecture to first compute features for each genomic location from the relevant alignments, and then to call structural variants from the set of all features across the genome.

In this framework, we have developed and evaluated a distributed deletion-finding algorithm based on fitting a Gaussian mixture model (GMM) to the distribution of mapped insert sizes spanning each location in the genome. A similar method was used in MoDIL[1]; however, our algorithm and the Hadoop framework drastically reduce the runtime requirements and overall difficulty of using this approach.

On simulated and real data sets of paired-end reads, our algorithm achieves performance similar to or better than a variety of popular structural variation detection algorithms, including read-pair, split-read, and hybrid approaches. Cloudbreak performs well on both small and medium size deletions, and in our simulations has greater sensitivity at most fixed levels of specificity. We also show increased performance in repetitive areas of the genome, identifying more deletions that overlap repeats than other approaches in both simulated and real data.

In addition, our algorithm can accurately genotype heterozygous and homozygous deletions from diploid samples. Using the parameters computed in fitting the GMM and a simple thresholding procedure, we were able to achieve 88.0% and 94.9% accuracy in predicting the genotype of the true positive deletions we detected in simulated and real data sets, respectively.

Finally, we have recently added the ability to detect insertions to Cloudbreak. Our implementation and source code are available at `https://github.com/cwhelan/cloudbreak`.

[1] Lee, S. et al., 2009. MoDIL: detecting small indels from clone-end sequencing with mixtures of distributions. *Nat. Methods*, 6(7), pp.473474.

Condition specific sub-network identification using a continuous optimization model

Bayarbaatar Amgalan [1], Hyunju Lee [1,*]

[1]Department of Information and Communications, Gwangju Institute of Science and Technology, Gwangju, Republic of Korea.

[*] To whom correspondence should be addressed.

Emails: [1]bayaraa@gist.ac.kr; [1,*]hyunjulee@gist.ac.kr;

Sub-networks can reveal the complex patterns of the whole bio-molecular network by extracting the interactions that depend on temporal or condition specific context. When genes interact with each other during cellular process, they may form differential co-expression patterns with other genes in different cell states. The identification of condition specific sub-networks is of great importance for investigating how a living cell adapts to changing environments.

In this work, we propose an optimization model, which uses scoring parameters that jointly measure the condition-specific changes of both individual genes and gene-gene co-expression, to identify the condition specific sub-network that has maximal score. Finding maximal scoring sub-network is generally formulated as a combinatorial optimization problem. Bio-molecular networks are often large in scale. It is impossible to solve such a large combinatorial optimization problem exactly in reasonable time. To address this issue, we formulate the sub-network identification problem as a continuous optimization problem which is an approximation of the general combinatorial problem based on the theorem due to Motzkin and Straus. It relates maximum cliques of a weighed graph to the optimization of a quadratic function under sparsity constraints. The optimization problem can be efficiently solved by the continuous genetic algorithm to find a single optimal sub-network which maximizes the quadratic objective function under sparsity constraints.

We applied this model to analyze a real prostate data set. Compared with previous methods, the optimization model is more robust in identifying truly significant sub-networks of appropriate size and meaningful biological relevance.

Adjusted z-score approach to pathway analysis incorporating dependencies among biomolecules

En-Yu Lai[1,2], Yi-Hau Chen[3], Kun-Pin Wu[1,*]

[1]: Institute of Biomedical Informatics, National Yang Ming University, Taipei 11221, Taiwan.

[2]: Bioinformatics Program, Taiwan International Graduate Program, Institute of Information Science, Academia Sinica, Taipei 11529, Taiwan.

[3]: Institute of Statistical Science, Academia Sinica, Taipei 11529, Taiwan.

[*]: To whom correspondence should be addressed.

Emails: E. Y. Lai (junelai@iis.sinica.edu.tw); Y. H. Chen (yhchen@stat.sinica.edu.tw); K. P. Wu (kpwu@ym.edu.tw)

With the popularization of high-throughput technology, a huge amount of quantitative data has been produced by microarray or mass spectrometry. Consequently, approaches to detect pathways or other functional categories, which are relevant to the underlying molecular mechanisms behind the expression profiles, have been developed in recent years.

Most of the available methods only focus on the expression of genes or proteins and assume these biomolecules as independent units. This assumption ignores the correlation among biomolecules with similar functions or cellular localization, as well as the interactions among them manifested as changes in expression ratios. As a result, the independent assumption leads to inaccurate small *p*-values and often causes serious false positives.

In this study, we present a method based on the statistical meta-analysis incorporating correlations of expression profiles among genes or proteins. The concept of meta-analysis integrates individual evidences for each of biomolecules into a total evidence for the functional category being evaluated. The proposed method uses the probabilities provided by the STRING database to estimate the correlation/interaction structure between biomolecules. After taking the dependencies into consideration, the proposed method will adjust the test statistics and provide a list of functional categories with more precise *p*-values.

Network construction and analysis for the human gut metagenome

Peng He [1], Rui Jiang [1,*]

[1]: MOE Key Laboratory of Bioinformatics and Bioinformatics Division, TNLIST/Department of Automation, Tsinghua University, Beijing 100084, China.

[*]: To whom correspondence should be addressed.

Emails: Peng He (hep10@mails.tsinghua.edu.cn); Rui Jiang (ruijiang@tsinghua.edu.cn)

Metagenomic studies of the human gut microbiome have revealed that intestinal living organisms play an important role in the human health. Most researches in this area focused on the taxonomical diversity and functional variety of the metagenome sampled from different individuals. From a system-level point of view, we try to construct and analyze a series of metagenomic networks, which reveal the underlying interactions between different units that compose the human gut microbiome. We collect two deep shotgun sequencing datasets of the gut microbial DNA, containing about 500 samples from Chinese and European individuals with different health status. All sequences are aligned to the assembled gene catalogue, as well as some existing databases to obtain the abundance of different genes, genera, KEGG orthologous groups (KO) and eggNOG orthologous groups (OG). With these profiles, we apply several state-of-the-art methods to construct metagenomic networks of different levels from different classes of samples. We calculate and compare the topological properties of the networks constructed from healthy samples and samples with different diseases. We also look into the node-level properties of different groups of units within the metagenomic networks. These networks and the analysis would be helpful for us to have a better understanding of the human gut microbiome, the interactions within it and its correlation with host properties.

SYSTEMS BIOLOGY ANALYSIS OF COMPLEX DISORDERS

Keywords: translational medicine, biological networks, gene prioritization

Progress in understanding of molecular mechanisms underlying complex heritable disorders (*e.g.*, autism, schizophrenia, diabetes) depends on new bioinformatics approaches for systems-level analysis and identification of disease-specific patterns of inheritance.
We present an approach and a supporting computational platform LYNX (http://lynx.ci.uchicago.edu/) for the analysis of complex heritable disorders from the systems biology perspective. Our approach is based on a large-scale integration of genomic and clinical data and various classes of biological information from over 35 public and private databases. This data is used for the identification of genes and molecular networks contributing to phenotypes of interest, as well as for the prediction of additional high-confidence disease genes to be tested experimentally. Our analytical strategy includes: (a) the enrichment analysis of high-throughput genomic data; (b) feature-based gene prioritization and (c) the development of network-based disease models for the identification of molecular mechanisms involved in disease pathogenesis. Networks-based gene prioritization leverages previous work of Dr. Börnigen-Nitsch on PINTA system. The algorithms were modified to accommodate a variety of weighted data types for gene prioritization. Our analysis allowed uncovering some of the molecular mechanisms that underlie the brain connectivity disorders. This knowledge will eventually lead to the development of efficient diagnostic and therapeutic strategies.

A Log-Linear Graphical Model for Inferring Genetic Networks from High-Throughput Sequencing Data

Genevera I. Allen [1,2], Zhandong Liu[1*]

[1]: Neurological Research Institute, Baylor College of Medicine, Houston, TX, U.S.A.

[2]: Department of Statistics, Rice University, Houston, TX, U.S.A.

[*]: To whom correspondence should be addressed.

Emails: G.A. (gallen@bcm.edu); Z.L.(Zhandong.liu@bcm.edu);

Gaussian graphical models are often used to infer gene networks based on microarray expression data. With the development of the next generation sequencing technology, many scientists have begun using high-throughput sequencing technologies to measure gene expression. As the resulting high-dimensional count data consists of counts of sequencing reads for each gene, Gaussian graphical models are not optimal for modeling gene networks based on this discrete data. We develop a novel method for estimating high-dimensional Poisson graphical models, the Log-Linear Graphical Model, allowing us to infer networks based on high-throughput sequencing data. Our model assumes a pair-wise Markov property: conditional on all other variables, each variable is Poisson. We estimate our model locally via neighborhood selection by fitting 1-norm penalized log-linear models. Additionally, we develop a fast parallel algorithm, an approach we call the Poisson Graphical Lasso, permitting us to fit our graphical model to high-dimensional genomic data sets. In simulations, we illustrate the effectiveness of our methods for recovering network structure from count data. A case study on breast cancer microRNAs, a novel application of graphical models, finds known regulators of breast cancer genes and discovers novel microRNA clusters and hubs that are targets for future research.

References

1. G. I. Allen and Z. Liu. A Log-Linear Graphical Model for Inferring Genetic Networks from High-Throughput Sequencing Data. In IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2012

A Prognostic CNA Signature Sub-Stratifies Intermediate-Risk Prostate Cancer Patients

Emilie Lalonde[1,2*], Adrian Ishkanian[3,4], Jenna Sykes[4], Nathalie Moon[1], Gaetano Zafarana[3,4], John Thoms[3,4], Cherry Have[1], Chad Malloff[5], Varune Ramnarine[3,4], Alice Meng[3,4], Denise Mak[1], Lauren Chong[1], Dorota Sendorek[1], Omer Ahmed[3,4], Jeremy A. Squire[6], Igor Jurisica[2,3,4], Alan Dal Pra[3,4], Melania Pintilie[3,4], Theo van der Kwast[3,4], Wan L. Lam[5], Michael Milosevic[3,4], Paul C. Boutros[1,2*], Robert G. Bristow[2,3,4*]

[1] Informatics and Bio-Computing, Ontario Institute for Cancer Research, Toronto, Ontario

[2] Department of Medical Biophysics, University of Toronto

[3] Departments of Radiation Oncology, Medical Biophysics, Laboratory Medicine and Pathology and Dalla Lana School of Public Health, University of Toronto

[4] Ontario Cancer Institute/Princess Margaret Hospital-University Health Network

[5] Department of Integrative Oncology, British Columbia Cancer Research Centre

[6] Department of Pathology and Oncology, Queen's University, Kingston, Ontario

[*]: To whom correspondence should be addressed.

Emails: EL (emilie.lalonde@oicr.on.ca); PCB (paul.boutros@oicr.on.ca); RGB (rob.bristow@rmp.uhn.on.ca)

Men with prostate cancer (CaP) are stratified into low, intermediate and high risk groups based on clinical factors such as pre-treatment prostate-specific antigen (PSA) levels, tumour grade and tumour stage. Intermediate-risk patients vary widely in clinical outcome, with a 20-40% recurrence rate, as measured by a rise in post-treatment PSA concentration (biochemical recurrence). Unfortunately, there is no way to accurately identify the intermediate-risk patients that derive benefit from therapy. To address this issue, we developed prognostic signatures to further stratify these patients into sub-groups with distinct risk-profiles by applying machine learning to gene copy number profiles from intermediate-risk patients.

Array comparative genomic hybridization (aCGH) was applied to frozen biopsies from 126 intermediate-risk CaP patients prior to image-guided radiotherapy. Copy number aberrations (CNAs) were extracted and used to develop signatures which were then evaluated in an independent cohort of 129 low to intermediate risk patients treated by radical prostatectomy. With unsupervised hierarchical clustering, we identified four distinct patient groups within the radiotherapy cohort. Patients from the surgery cohort were matched to these clusters and the resulting clusters have statistically different biochemical recurrence rates. We also used a supervised learning approach to develop a CNA-signature. This signature is effective at identifying patients at risk of biochemical recurrence, while accounting for clinical covariates (HR= 6.12, $p = 1.29 \times 10^{-9}$).

We have recapitulated known genomic heterogeneity and have developed a clinically-relevant CNA-signature which stratifies intermediate-risk patients into two refined risk groups. This genomic biomarker is promising in improving clinical management of intermediate-risk CaP patients.

A Probabilistic Framework of Constructing Gene Network from Multifold Gene Expression Data

Xueying Xie[1], Yunfei Bai[2]

1. The Research Center of Learning Science, Southeast University, China

2. School of Biological Science and Medical Engineering, Southeast University, China

Email: xxying.rcls@seu.edu.cn

I*n silico* reconstruction of gene regulatory networks becomes one of the most challenging issues in functional genomics, especially with the advance of high-throughput gene expression data. Though a lot of work had been done to model the gene regulatory networks, they didn't propose a generalized framework for the construction and analysis of gene regulatory networks from the multifold gene expression data. In our work, we propose a probabilistic framework to model the gene regulatory network with the consideration of multi-level regulatory mechanisms. The mathematic basis of our model is the factor graph which is used widely in many fields owning to its particular features, including flexibility, established theory foundation and linear computational complexity. In order to handle the manifold diversity of the data (including regulatory factors, miRNA, and regulated genes), we also put forward two methods to set up the model parameters adaptively in case to avoid the biases caused by subjective choose of parameters. We apply our model to infer the gene regulatory networks from two datasets. One is the simulated data which is generated randomly according to the predetermined network structure and the characteristics of the real gene expression data. The other is the RNA-seq data from TNF-alpha treated HepG2 cell. We successfully find all relationships that have been pre-set in the simulated data. The result validates the feasibility and the efficiencyof our model. From the real data, our method finds many significant candidate interactions between miRNA and mRNA, and a high percentage of which are highly consistent with ones validated experimentally in public database.

GIANT：Genome-wide Identification of Somatic Aberrations from Paired Normal and Tumor Samples

Ao Li*, Yuanning Liu, Minghui Wang

Centers for Biomedical Engineering

School of Information Science and Technology

University of Science and Technology of China

P.O.Box 4, West Campus

Hefei, AH230037, China

*: To whom correspondence should be addressed.

Emails: A L (aoli@ustc.edu.cn); Y L (lynn100@mail.ustc.edu.cn); M W (mhwang@ustc.edu.cn)

Cancer genome often encompasses numerous of aberrations originated during tumorigenesis, which play an important role in tumor evolution and progression [1]. Single-nucleotide polymorphism (SNP) genotyping array technique provides a great opportunity to profile the genomic aberrations with high resolution, but sophisticated computational methods are sorely needed for accurately recognizing genomic aberrations from genotyping signals dramatically affected by normal cell contamination, tumor aneuploidy, and GC content. In this study, we introduced a novel bioinformatics toolkit — GIANT, for dissecting paired normal-tumor SNP-array data. GIANT mainly consists of two components: 1) a statistical model called PSHMM(paired samples hidden Markov model), which is designed to identify somatic aberrations in tumor sample by borrowing the genotype information of paired normal sample; 2) genome-wide permutation test for discovering statistically significant aberrations across the cancer genome. Results of GIANT on simulated and real paired normal-tumor samples showed that GIANT outperformed current "state of the art" approaches for paired SNP-array data analysis, such as OncoSNP [2] and ASCAT [3], with higher sensitivity and accuracy. Moreover, by using GIANT we successfully discovered novel somatic/germline aberrations in HER2+ breast cancer and urothelial cancer samples. Finally, Applying GIANT to 112 paired normal-tumor samples depicted the gnome-wide breast cancer landscape including significant amplification, deletion and loss of heterozygosity (LOH) harboring many oncogenes and tumor suppressor genes, which provide an exquisite genome profile for further cancer driver gene studies. GIANT is freely available at http://bioinformatics.ustc.edu.cn/giant.

References

1. Micheal RS, Peter JC, P Andrew W. The cancer genome. *Nature*, 458: 719-724, 2009.
2. Yau C, Mouradov D, Jorissen RN, et al. A statistical approach for detecting genomic aberrations in heterogeneous tumor samples from single nucleotide polymorphism genotyping data, *Genome Biol*, 11: R92, 2010.
3. Van Loo P, Nordgard SH, Lingjarde OC, et al. Allele-specific copy number analysis of tumors. *Proc Natl Acad Sci*, 107: 16910-16915, 2010.

A miR-21-PDCD4 sub-network effects TGF-beta induced apoptosis in liver cancer cells

Lingyun YIN [1], Qi WANG [1], Yang CHEN [1], Michael Q ZHANG[1,2,*]

[1]: MOE Key Laboratory of Bioinformatics; Bioinformatics Division/Center for Synthetic & System Biology, TNLIST; Department of Automation; Tsinghua University, Beijing 100084, China.

[2]: Department of Molecular and Cell Biology, Center for Systems Biology, The University of Texas at Dallas, Richardson, Texas 75080, USA

[*]: To whom correspondence should be addressed.

Emails: L.Y.YIN (yly.thu@gmail.com); Q.WANG (wqwq945@163.com); Y.CHEN (yc@tsinghua.edu.cn); M.Q.ZHANG (michael.zhang@utdallas.edu).

Transforming growth factor beta (TGF-beta), a major inflammatory cytokine, plays a key role in several disease, such as Parkinson's disease, heart disease and cancer[1-3]. In normal cells, TGF-beta usually triggers apoptosis though DAXX pathway and canonical signaling pathway SMAD pathway[4].

However, in transformed cells parts of the TGF-beta pathway are mutated. To illustrate the molecular mechanism underlying this process, we constructed TGF-beta induced tumor cell Huh7, a hepatocellular carcinoma cell line. mRNA, miRNA and genome-wide PPI datasets were used to construct responsive molecular networks. Then, we analyzed the topological properties of the networks and calculated the node degree and node betweenness centralization of each gene in the molecular networks. We found that PDCD4 which is one of direct target of oncogene,miR-21[5], is closely related to the center of TGF-beta induced apoptosis pathway. At the same time, it was predicted that miR-21 could increase the degradation of PDCD4 by PI3K/mTOR pathway indirectly.

After that, according to a series biological experiments, we proved that aberrant expressed miR-21 can inhibit the translation and increase the degradation of PDCD4 which makes the liver cancer cells hyposensitive to TGF-beta induced apoptosis. And this mutated part of pathway may serve as a target for effective molecular cancer therapies.

References

1. Rubio-Perez JM, Morillas-Ruiz JM: **A review: inflammatory process in Alzheimer's disease, role of cytokines**. *ScientificWorldJournal* 2012, **2012**:756357.
2. Garside VC, Chang AC, Karsan A, Hoodless PA: **Co-ordinating Notch, BMP, and TGF-beta signaling during heart valve development**. *Cell Mol Life Sci* 2012.
3. Derynck R, Akhurst RJ, Balmain A: **TGF-beta signaling in tumor suppression and cancer progression**. *Nat Genet* 2001, **29**(2):117-129.
4. Butz H, Racz K, Hunyady L, Patocs A: **Crosstalk between TGF-beta signaling and the microRNA machinery**. *Trends Pharmacol Sci* 2012, **33**(7):382-393.
5. Chen Y, Liu W, Chao T, Zhang Y, Yan X, Gong Y, Qiang B, Yuan J, Sun M, Peng X: **MicroRNA-21 down-regulates the expression of tumor suppressor PDCD4 in human glioblastoma cell T98G**. *Cancer Lett* 2008, **272**(2):197-205.

Transcriptional regulation analysis in the peripheral blood from cervical cancer patients undergoing concurrent chemoradiation

Wei-Hsiang Kung [1,2*], Jui-Hung Hung [1], Hsien-Da Huang [1]

[1]: Institute of Bioinformatics and System Biology, National Chiao Tung University.

[2]: Department of Medical Imaging Technology, Su-Zen Junior College of Medicine and Management.

[*]: To whom correspondence should be addressed.

Emails: Wei-Hsiang[*] (brach@ms.szmc.edu.tw); Jui-Hung (juihunghung@gmail.com); Hsien-Da (bryan@mail.nctu.edu.tw)

In the United States about 12,200 new cases of cervical cancer as well as 4,210 deaths from cervical cancer are estimated for 2010. Cervical cancer continues to be an important worldwide health problem for women, especially in developing countries without established screening programs. It can be cured by radical surgery or radiotherapy with equal effectiveness. In this study, we investigated the contribution of gene variants to the treatment of cervical cancer from public datasets. Firstly, the standard deviation was used to filter the low-variation genes (SD > 0.5) and student t test (p value < 0.05) was performed between any two times to select differentially expressed genes among four times during and after concurrent chemoradiation. Secondly, 397 genes were subject to DAVID bioinformatics database for functional annoatation. According to molecular function annotation, a cluster of structural constituent ribosome genes (RPL12, RPL23, RPL23A, RPL24, RPS2, RPS15, RPS27A, UBA52, UBC) (p value < 0.01) were found in relation to cervix neoplasia (UP_TISSUE annotation, p value < 0.00001).

Pearson correlation (r > 0.7) was applied to search structural constituent ribosome related co-expression genes (ATP1A1, CAPN1, CHUK, DLX2, IRF2, KRT7, RPL12, RPL23, RPL23A, RPL24, RPS2, RPS15, RPS27A, UBA52, UBC). The regulation for these genes used CORE_TF database to search conserved and over-represented transcription factors. The factors (AR, ATF, DEAF1, NFKB1, NRF2, TAXCREB) (p value < 0.01) showed significant roles on ribosomal genes regulation in human, mouse, and rat. Our findings at the gene expression and transcription regulation suggest that DEAF1 had a opposite tendency compared to ribosomal genes expression and may play a critical role in cervical carcinogenesis.

Disease Module Identification from an Integrated Transcriptomic and Interactomic Network Using Evolutionary Community Extraction

Yunpeng Liu [1], Daniel A. Tennant [2], John K. Heath [3], Shan He [1, 3*]

[1]: the School of Computer Science, the University of Birmingham.

[2]: the School of Cancer Sciences, the University of Birmingham.

[3]: the School of Biological Sciences, the University of Birmingham.

[*]: To whom correspondence should be addressed.

Emails: Y. Liu (YXL221@bham.ac.uk); D.A. Tennant (d.tennant@bham.ac.uk); J.K. Heath (j.k.heath@bham.ac.uk), S. He (s.he@cs.bham.ac.uk)

Molecular mechanisms underlying the transition from low grade glioma to a high grade Glioblastoma Multiforme may be revealed by the identification of putative disease modules in glioma progression, i.e., group of interacting biological network components that collectively contribute the development from low- to high-grade gliomas. In this paper, we proposed a novel genetic algorithm based community extraction algorithm based on [1]; and applied it to an integrated network of glioma transcriptomic and interactomic data to extract putative disease modules that differentiate Grade II gliomas from Grade IV Glioblastoma Multiforme. Extraction revealed two putative disease modules of 33 and 15 genes, respectively. In the module with the highest score, we found three guanine nucleotide exchange factors (GEFs) TRIO, ECT2 and VAV3, which have been shown to mediate the invasive behaviour of glioblastoma [2]. In addition, the cell cycle regulatory protein CDC42 was also included in the module with dense interactions with other GEFs, consistent with recent studies [3]. The second putative disease module primarily consisted of several members of the mini-chromosome maintenance (MCM) gene family, e.g., MCM2-5, which have been recently identified as high-grade glioma markers [4]. In this module, we also found CDK2, another key kinase in cell cycle progression, and POLE2, a cancer gene recently identified to be associated with bowel cancer [5], suggesting their potential roles in glioma transformation and invasion.

References

1. Zhao Y, Levina E, Zhu J. (2011). Community extraction for social networks. *Proc Natl Acad Sci USA*. 108(18): 7321-26.
2. Salhia B, et al. (2008). The guanine nucleotide exchange factors trio, Ect2, and Vav3 mediate the invasive behavior of glioblastoma. *Am J Pathol*. 173(6): 1828-38.
3. Fortin SP, et al. (2012). Cdc42 and the guanine nucleotide exchange factors Ect2 and trio mediate Fn14-induced migration and invasion of glioblastoma cells. *Mol Cancer Res*. 10(7): 958-68.
4. Bie L, Zhao G, McClland M, Ju Y. (2012). Minichromosome maintenance (MCM) family as potential prognostic tumor markers for human gliomas, *Neuro-Oncology*. 14(6): 50-52.
5. Palles C, et al. (2012) Germline mutations in the proof-reading domains of POLE and POLD1 predispose to colorectal adenomas and carcinomas. *Nature Genetics*. 45(2): 136-144.

Retrofitting Functional Prediction Methods to Fill Gaps in Metabolic Networks

Nam Ninh Nguyen[1], Wanwipa Vongsangnak[2], Hon Wai Leong[1,*]

[1]: Dept of Computer Science, National University of Singapore, Singapore 117417.

[2]: Center for Systems Biology Soochow University, Suzhou, China.

[*]: To whom correspondence should be addressed.

Emails: NNN (nguyennn@comp.nus.edu.sg); WV (wanwipa@suda.edu.cn); LHW (leonghw@comp.nus.edu.sg)

The recently reconstructed metabolic networks of *Saccharomyces cerevisiae* (SC) [1] and *Aspergillus oryzae* (AO) [2] contain a significant number of metabolic gaps (i.e. reactions without gene identified), which is a bottleneck for understanding the cellular metabolism and physiology of these species. This work presents an attempt to fill the gaps and thus enhance these metabolic models. We develop a missing enzyme predictor, called EnzPro, based on profile hidden Markov models. We also retrofit two general function predictors (PFP and Blast2GO) and two metabolic function predictors (PRIAM and EFICAz) for missing gene prediction. First, we run these methods for whole genome annotation. Second, we use the EC2GO mappings to map between predicted GO-terms and the EC numbers, if necessary. Finally, we consolidate the result and output candidates for each gap. Performance of these methods is evaluated by self-testing on already-identified enzyme datasets.

We found that, the general function predictors are not sufficient enough due to their low precision (24-77%) and recall (51-83%). On the other hand, the metabolic function predictors get sufficient accuracy (precision: 72-95%, recall: 78-94%) in testing datasets, but they make very few predictions for the gaps. Our method, EnzPro, gets a tolerable accuracy (precision: 53-80%, recall: 54-92%), and is able to predict candidates for 61% of the gaps. Thus, to achieve higher coverage for gap prediction and utilize the high confidence of common prediction, integration of all methods should be considered.

Applying the above methodology, we are able propose candidates for 38/61 gaps in AO network (38/52 gaps in SC). For example, the pantetheine-phosphate adenylyltransferase (PPAT) reaction (EC:2.7.7.3) in AO is predicted to be catalyzed by AO090023000706. This protein matches with the PPAT domain in Conserved Domains Database, and cytidylyltransferase domain in Pfam, with similar structures to coaD gene (EC:2.7.7.3) in *Escherichia coli* and *Bacillus subtilis*. As another example in SC, YNL168C is a candidate for fumarylacetoacetase (EC:3.7.1.2). This protein matches to InterPro entry IPR002529 (fumarylacetoacetase, C-terminal, EC:3.7.1.2), but currently annotated as unknown function in SGD database.

References

1. I. Nookaew et al. The genome-scale metabolic model iIN800 of Saccharomyces cerevisiae and its validation: a scaffold to query lipid metabolism. *BMC Sys. Biol*. 2:71, 2008.
2. W. Vongsangnak et al. Improved annotation through genome-scale metabolic modeling of Aspergillus oryzae. *BMC Genomics*, 9:245, 2008.

Computational analysis of synthetic lethality in DNA repair pathways with application to cancer treatment

Inna Kuperstein, Emmanuel Barillot and Andrei Zinovyev[1,2,3]

[1]Institut Curie, Paris, France,
[2] INSERM U900, Paris, France
[3] Ecole des Mines ParisTech, Paris, France

We present a study on characterizing known and predicting new synthetic lethal combinations of genes in DNA repair pathways. We first classify known types of synthetic lethal interactions and introduce a new type of within non-essential reversible pathway synthetic lethality, which is observed experimentally in Homologous Recombination Repair pathway. Second, we present a detailed reconstruction of DNA repair machinery in mammals in the form of comprehensive map of molecular interactions. Using this map, we are able to predict a number of synthetic lethal combinations of genes. In particular, we are able to predict in which genetic context some well-known synthetic pairs (such as BRCA+PARP) should be lethal. Our approach is particularly relevant for developing new treatment strategies in cancer therapy, like better stratifying patients, complementing genotoxic chemotherapy, or targeting specifically cancer cells harbouring certain mutations.

Keywords: DNA repair, synthetic lethality, cancer treatment

Identifying conserved protein complexes between species by constructing interolog interaction networks

Phi Vu Nguyen [1], Sriganesh Srihari [2,*], Hon Wai Leong [1,*]

[1] Department of Computer Science, National University of Singapore, Singapore 117417.

[2] Institute for Molecular Bioscience, The University of Queensland, St. Lucia, Australia.

[*] To whom correspondence should be addressed.

Emails: PVN (nphivu@comp.nus.edu.sg); SS (s.srihari@uq.edu.au); HWL (leonghw@comp.nus.edu.sg)

Protein complexes *conserved* across species indicate processes that are *core* to cellular machinery (*e.g.* cell-cycle or DNA damage-repair complexes conserved across human and yeast). While numerous computational methods have been devised to identify complexes from the protein interaction networks (PINs) of individual species, these are severely limited by noise and errors (false positives) in currently available datasets [1]. Consequently, our analysis using human and yeast PINs revealed that these methods overlook several important complexes including those conserved between the two species (24/42 yeast and 68/118 human conserved complexes missed – *e.g.* the MLH1-MSH2-PMS2-PCNA mismatch-repair complex).

Here, we propose to identify conserved complexes by constructing *interolog interaction networks* (IINs) by developing a novel method (inspired from [2]) to align PINs using protein-homolog information from species. IINs leverage the conservation of interactions between species, thereby reducing the number of false positives. We employ state-of-the-art methods (MCL, CMC and HACO) to cluster the IINs, and map these clusters back to the original PINs to identify complexes conserved between the species.

Evaluation of our IIN-based approach using human and yeast interaction data identified several additional complexes (16 in human and 9 in yeast) compared to direct complex detection from the original PINs. Our analysis revealed that the IIN-construction removed non-conserved interactions many of which were false positives, thereby improving complex prediction. In fact removing non-conserved interactions from the original PINs also resulted in higher number of conserved complexes (additional 12 in human and 2 in yeast), thereby substantiating our IIN-based approach. These complexes included the mismatch repair complex, MLH1-MSH2-PMS2-PCNA, and other important ones namely, RNA polymerase-II, EIF3 and MCM complexes, all of which constitute core cellular processes known to be conserved across the two species.

References

1. Srihari S., Leong HW. A survey of computational methods for protein complex prediction from protein interaction networks. *Journal of Bioinformatics and Computational Biology*, 11(2):1230002, 2013.
2. Sharan R., Ideker T., Kelley B., Shamir R., Karp RM**.** Identification of protein complexes by comparative analysis of yeast and bacterial protein interaction data. *Journal of Computational Biology*, 12(6): 835-846, 2005.

Network Analysis of Mutations Across Cancer Types

Mark D.M. Leiserson[1,*], Hsin-Ta Wu[1], Fabio Vandin[1], Benjamin Raphael[1]

[1]: Department of Computer Science and Center for Computational Molecular Biology, Brown University, Providence, RI, USA

[*]: To whom correspondence should be addressed.

Emails: ML (mdml@cs.brown.edu); HW (hsin-ta_wu@brown.edu); FV (vandinfa@cs.brown.edu) ; BR (braphael@cs.brown.edu)

Recent cancer sequencing studies from The Cancer Genome Atlas (TCGA), the International Cancer Genome Consortium (ICGC) and others have shown that relatively few genes are recurrently mutated in many samples from the same cancer type. Rather a large number of genes are mutated in a small number of samples, and thus most mutations are indistinguishable from random mutations at a reasonable level of statistical significance. One reason for this mutational heterogeneity is cancer mutations target different cellular signaling and regulatory pathways, and different genes in these pathways may be mutated in different individuals.

We performed a pan-cancer analysis of mutated pathways/networks using whole-exome sequencing and copy number aberration data in 2359 TCGA samples from twelve different cancer types: AML, BLCA, BRCA, COADREAD, GBM, HNSC, KIRC, LUAD, LUSC, OV, and UCEC. Rather than restrict our analysis to known cancer pathways, we use the HotNet algorithm [1] to identify subnetworks of a protein-protein interaction (PPI) network that are mutated in a statistically significant number of samples. HotNet identifies these subnetworks by modeling the mutations in a gene as a source of heat on the corresponding protein in the interaction network. HotNet employs a two-stage multiple hypothesis test to rigorously bound the false discovery rate of the list of "significantly hot" subnetworks.

We apply HotNet to the TCGA mutation data using the iRefIndex and HPRD PPI networks, each consisting of tens of thousands of interactions among thousands of proteins. We identify 13 (respectively 30) significantly mutated subnetworks ($P < 0.01$). These subnetworks overlap well-known cancer signaling pathways (e.g. p53, RTK, and RB signaling), but also include subnetworks with less characterized roles in cancer; e.g. the cohesin complex and the SLIT-ROBO pathway, the latter involved in cell migration. We also identify subnetworks that are significantly enriched for mutations in a specific cancer type (e.g. ARID1A, PBRM1 and other interacting proteins in kidney cancer samples). We show that some of the resulting subnetworks correspond to known pathways, other sets result from (sub)type-specific mutations, and the remaining subnetworks suggest novel groups of mutated genes. Thus, this pan-cancer analysis reveals mutational patterns in sets of genes that are not readily apparent from analysis of individual genes.

References

1. Vandin F, Upfal E, Raphael B: Algorithms for Detecting Significantly Mutated Pathways in Cancer. *J Comput Biol* 2011, 3**(18)**:507–522.

Network Prioritization and Functional Characterization of Candidate Disease Genes

Nadezhda T. Doncheva[1], Tim Kacprowski[2], Mario Albrecht[1,2,*]

[1]: Max Planck Institute for Informatics, Saarbrücken, Germany.

[2]: Department of Bioinformatics, Institute of Biometrics and Medical Informatics, University Medicine Greifswald, Greifswald, Germany.

[*]: To whom correspondence should be addressed.

E-mail: M.A. (info@mario-albrecht.de)

Many efforts are still devoted to the discovery of genes involved with specific phenotypes, in particular, diseases. High-throughput techniques are thus applied frequently to detect dozens or even hundreds of candidate genes. However, the experimental validation of many candidates is often an expensive and time-consuming task. Therefore, a great variety of computational approaches has been developed to support the identification of the most promising candidates for follow-up studies (Doncheva, N.T et al., WIREs Syst. Biol. Med., 4(5):429-442, 2012). The biomedical knowledge already available about the disease of interest and related genes is commonly exploited to find new gene–disease associations and to prioritize candidates. Our approach particularly integrates heterogeneous data sources and uses disease-specific network information. This also provides more insights into the functional characteristics of the underlying phenotypes, for example, of complex autoinflammatory diseases.

ContigScape: a Cytoscape plugin facilitating microbial genome gap closing

Qi Wang[3§], Biao Tang[1,2§], Minjun Yang[2], Feng Xie[3,6], Yongqiang Zhu[2], Ying Zhuo[3], Shengyue Wang[2], Hong Gao[3], Xiaoming Ding[1], Huajun Zheng[2*], Lixin Zhang[3*], Guoping Zhao[1,2,4,5*]

[1]: State Key Laboratory of Genetic Engineering, Department of Microbiology, School of Life Sciences, Fudan University, Shanghai 200433, China;

[2]: Shanghai-MOST Key Laboratory of Health and Disease Genomics, Chinese National Human Genome Center at Shanghai, Shanghai 201203, China;

[3]: CAS Key Laboratory of Pathogenic Microbiology & Immunology, Institute of Microbiology, Chinese Academy of Sciences, Beijing 100190, China;

[4]: CAS Key Laboratory of Synthetic Biology, Institute of Plant Physiology and Ecology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200032, China;

[5]: Department of Microbiology and Li KaShing Institute of Health Sciences, The Chinese University of Hong Kong, Prince of Wales Hospital, Shatin, New Territories, Hong Kong SAR, China;

[6]: Graduate School of Chinese Academy of Sciences, Beijing, 100049, China.

[*]: To whom correspondence should be addressed.

[§]: These authors contributed equally to this work.

Emails: LXZ (zhanglixin@im.ac.cn); HJZ (zhenghj@chgc.sh.cn); GPZ (gpzhao@sibs.ac.cn)

Due to complex genomic structures or repeat sequences, gap-closing is often a rate-limiting step in next generation sequencing. Therefore, we developed a Cytoscape plugin for analyzing the relationships between genomic contigs derived from sequence assemblers. By displaying the relationships using networks in Cytoscape, user has a more straightforward feeling of the orders of contigs. Moreover, the pattern of plasmids or special repeats (IS elements, ribosomal RNAs, terminal repeats, etc) can be easily identified. ContigScape is a nice tool in guidance of PCR verification for gap closing.

References

1. Boetzer,M. et al. (2011) Scaffolding pre-assembled contigs using SSPACE. Bioinformatics, 27, 578–579.
2. Nielsen,C.B. et al. (2009) ABySS-Explorer: visualizing genome sequence assemblies. IEEE Trans. Vis. Comput. Graph., 15, 881–888.
3. Oksana R.G.et al. (2010) Visualization and quality assessment of de novo genome assemblies. Bioinformatics, 27: Bioinformatics, 24(2) : 3425–3426.
4. Shannon,P. et al. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res., 13, 2498-2504.

A Quantitative Approach to Study microRNAs Regulation in Breast Cancer

Yu Liu[1,†], Peng Xie[1,†], Michael Q. Zhang[1,2], Xiaowo Wang[1,*]

[1]: MOE Key Laboratory of Bioinformatics and Bioinformatics Div, TNLIST/Department of Automation, Tsinghua University, Beijing 100084, China

[2]: Affiliation 2. Department of Molecular and Cell Biology, Center for Systems Biology, University of Texas at Dallas, Dallas, TX 75080, USA

[†]: These authors contributed equally to this work

[*]: To whom correspondence should be addressed.

Emails: YL (yuliu0819@gmail.com); PX (x-p-06@mails.thu.edu.cn); MZ (mzhang@cshl.edu); XW (xwwang@tsinghua.edu.cn)

MicroRNAs (miRNAs) are small noncoding RNA molecules, which known to play an important role in suppressing protein synthesis or promoting the degradation of their target mRNAs; therefore, deciphering the interaction of miRNA targets is crucial for understanding the regulatory process and benefits to diseases diagnostics and therapeutics. Recently, numbers of computational methods have been developed based on sequence signatures and cross-species conservation, but all suffered from high false positive rates and lack of overlap between each other; in the last few years, several improved approaches which took into account of the expression values of miRNA and mRNA with the sequence based predictions in order to achieve more accurate relationships. However, such approaches mainly use the model from correlation, linear regression and Bayesian inference, are highly depended on the given database and cannot draw a specific map of how miRNA and its targets work under the specific cellular condition. In our study, we proposed a quantities model based on equilibrium statistical mechanics to estimate the miRNA occupation rate at each potential target site under a specific cellular condition by considering the expression of miRNA and mRNA, and their interaction energy. Our prediction showed high correlaiton with the known high-throughput sequencing of RNAs isolated by crosslinking immunoprecipitation (HITS-CLIP) data. And we applied our method on the Cancer Genome Atlas (TCGA) breast cancer dataset to predict the prominent miRNA-target pairs among the tumor and normal samples. Comparing with the sate-of-the-art miRNA-target prediction methods which combine both expression data with sequence information (Migia, GenMiR++ and TALASSO), our methods showed better consistency with the literature reported breast cancer signature miRNA-targets pairs, and gave a detailed view of how regulation changes between cancer and normal tissue.

References

1. Jim C Huang, Tomas Babak, Timothy W Corson, et al. Using expression profiling data to identify human microRNA targets. *Nat Methods*, 4(12): p. 1045-9, 2007
2. Ander Muniategui, Ruben Nogales-Cadenas, Miguel Vazquez, et al. Quantification of miRNA-mRNA interactions. *PLoS One*, 7(2): p. e30766, 2012
3. TCGA Data Portal (http://tcga-data.nci.nih.gov/tcga/findarchives.htm)

Mathematical model of cancer treatment response in the presence of cooperative intercellular interactions

Chanchala D. Kaddi[1] and May D. Wang[1,*]

[1]: Wallace H. Coulter Department of Biomedical Engineering, Georgia Institute of Technology and Emory University, Atlanta, GA, USA 30332

[*]: To whom correspondence should be addressed.

Emails: C.D.K. (gtg538v@mail.gatech.edu); M.D.W. (maywang@bme.gatech.edu)

Cancers are complex diseases involving the interactions of heterogeneous cell types with each other and with their microenvironment, mediated through a large variety of bio-molecules. The complexity of the cancer 'ecosystem' has led to recent research into how ecological principles, such as cooperation, may be applied to study cancer development and progression.

In cooperative interactions, one or more of the entities (e.g. cancer cells) involved receives a fitness benefit, and none experience a reduction in fitness. Cooperative behavior has been linked with major processes in cancer, such as growth factor production, angiogenesis, invasion, and metastasis. In this study, we use an agent-based model of a cancer cell population to investigate how cooperative intercellular interactions can affect the response to drug treatment.

We extend upon an earlier model of cooperation in cancer to examine how cooperative interactions, drug exposure, and drug resistance combine to affect cancer cell population behavior. We identify examples of qualitative agreement between the model predictions and experimental observations in the literature. We present this study as a step towards the development of more detailed models for investigating how specific mechanisms of intercellular cooperation in cancer may affect treatment response.

A Bayesian Approach to Reasoning on a Causal Biological Network

Robert Ness [1], Halima Bensmail [2], Olga Vitek [1,3,*]

[1]: Department of Statistics, Purdue University, West Lafayette, Indiana, USA.
[2]: Qatar Computational Research Institute, Qatar Foundation, Doha, Qatar.
[3]: Department of Computer Science, Purdue University, West Lafayette, Indiana, USA.
[*]: To whom correspondence should be addressed.

A biological causal network is defined as a graph where nodes are molecular events, such as a change in abundance or concentration of a compound or gene product. Directed edges between two events imply a causal relationship between them. Causal networks are composed from biomedical literature: each edge is annotated with a source literary reference which contains evidence of the relationship. The motivation of applying causal networks to systems biology experimental data (transcriptomics, proteomics, etc) is to identify the causal mechanism of the observed changes in the data. This approach is particularly useful in drug discovery and repositioning, for example, because of the potential to identify causal molecular mechanisms susceptible to drug intervention.

A general approach called "reverse causal inference" maps state changes in the data to corresponding events in the network, and searches in reverse along directed paths to find possible causes of the observed changes. Current methods for reverse causal inference typically reduce expression data to a dichotomous variable representing significant change in abundance, which results in loss of information. Further, they only consider the shortest paths between hypotheses and the data-mapped nodes, ignoring the information hidden in the graph's topological complexity. We propose a reverse causal inference method that makes use of the information in graph topology and all the explanatory power in the data.

We look at cases of microarray data in case/control experiments, where the data-mapped nodes are state changes involving genes. We use a Markov random walk based algorithm similar to Pagerank, to acquire a vector that quantifies the proximity of a given candidate hypothesis node to each of the data-mapped nodes. To evaluate a the candidate hypothesis, a LASSO logistic regression model is fit with case/control status as the response, each gene as a predictor, and where cross validation is used to select a base penalty. The proximity value for each gene, with respect to a given hypothesis, is used in forming gene-specific penalties used in the LASSO model. The result is that the further a gene's node is from the given hypothesis node in the network, the stronger the penalty, and therefore the more likely its regression coefficient will be shrunk to 0 in the LASSO fitting, depending on its explanatory power. Genes for which no path from the hypothesis node exist in the network are excluded from the linear model. Thus, for each candidate hypothesis a different LASSO model is fit. The penalties determine a unique network neighborhood for each hypothesis, and each hypothesis is evaluated by how much explanatory power is its neighborhood. The candidate hypothesis is scored by a goodness-of-fit measure for the linear model called the Bayesian Information Criterion (BIC). Since the ratio of two BIC scores for two different models is the Bayes factor, we are effectively using Bayesian model selection to compare the explanatory power of candidate hypotheses.

Using a causal network of 22898 nodes and 114027 edges, we simulated random data for all the RNA expression events in the network. Then, for a given candidate causal event, we resimulated significant fold change for all the RNA expression events with which it has a causal relationship, and tested our model's ability to recover that event. We compare our performance to other reverse causal inference methods. We also apply it to expression data from a cigarette smoking study using a network with 730 nodes and 778 causal edges specific to the context of cellular stress. The resulting putative regulators constitute directly testable hypotheses for follow-up in the laboratory. We present biologists with a method for capturing the information hidden in complex prior knowledge, and computing a simple comparison statistic (the goodness of fit measure BIC) for evaluating hypotheses. Using simulated data we quantify the recoverability of embedded signals from regulators for our causal graph under various kinds of noise; and we give a concrete example where our methodology helps elucidate biological phenomena when presented with real data.

Funding: This work was supported by the Qatar Computational Research Institute, Qatar Foundation, Doha, Qatar

The comparison of epigenomes of 22 mouse tissues recapitulates the cellular differentiation pathway

Song Yang [1*], Inna Dubchak [1], Dario Boffelli [2]

[1]: Genomics Division, Lawrence Berkeley National Lab

[2]: Children's Hospital Oakland Research Institute

[*]: To whom correspondence should be addressed.

Emails: SY (songyang@lbl.gov); ID (ildubchak@lbl.gov); DB (dboffelli@chori.org)

With the advances in next-generation sequencing technology, epigenomes mapping projects are revealing epigenomics information at different level of biological complexity, from cells to species. At the organismal level, comparison of epigenomes of diverse cell types may provide insights into cellular differentiation and organism development. The ENCODE and modENCODE project provide functional genome annotations of human and other model organism at various levels, including DNA methylation patterns and Histone modification marks. In this study, we investigated the epigenomes of three histone marks (H3k4me1, H3k4me3 and H3k27ac) from 22 different mouse tissues, and showed that the comparison of genome-wide histone modification can recapitulate the cellular differentiation pathways of different tissues, which are represented in a tree-like structure. Tissues developed from the three primary layers, endoderm, mesoderm and ectoderm are separated into three clades in the differentiation tree. The clustering and maximum parsimony analysis of the differentiation tree revealed the gain and loss of histone modification marks at each transition branches. Functional analysis (GO terms) of these histone marks and their associated genes were compatible to the cellular innovations during organism development.

GEOGLE: context mining tool for the correlation between gene expression and the phenotypic distinction

Yao Yu [#,1,2], Kang Tu [#,1], Siyuan Zheng [1], Yun Li [1], Guohui Ding [1], Jie Ping [4], Xuan Li [*,1,3], Pei Hao [*,1,3] and Yixue Li [*, 1,2,3,4,5]

[1]: Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200031, PR China;

[2]: Graduate School of the Chinese Academy of Sciences, Shanghai 200031, PR China;

[3]: Shanghai Center for Bioinformation Technology, 100 Qinzhou Road, Shanghai 200235, PR China;

[4]: College of life science and biotechnology, Shanghai Jiaotong University, Shanghai 200240, PR China;

[5]: College of life science and biotechnology, Shanghai Tongji University, Shanghai 200331, PR China

#: Equal contributors.

*: To whom correspondence should be addressed.

Emails: Yao Yu - yyu01@sibs.ac.cn; Kang Tu - ktu@sibs.ac.cn; Siyuan Zheng - syzhenger@gmail.com; Yun Li - yli01@sibs.ac.cn; Guohui Ding - gwding@sibs.ac.cn; Jie Ping - pjtalent@sjtu.edu.cn; Xuan Li* - lixuan@sibs.ac.cn; Pei Hao* - phao@sibs.ac.cn; Yixue Li* - yxli@sibs.ac.cn

In the post-genomic era, the development of high-throughput gene expression detection technology provides huge amounts of experimental data, which challenges the traditional pipelines for data processing and analyzing in scientific researches. In our work, we integrated gene expression information from Gene Expression Omnibus (GEO), biomedical ontology from Medical Subject Headings (MeSH) and signaling pathway knowledge from sigPathway entries to develop a context mining tool for gene expression analysis – GEOGLE. GEOGLE offers a rapid and convenient way for searching relevant experimental datasets, pathways and biological terms according to multiple types of queries: including biomedical vocabularies, GDS IDs, gene IDs, pathway names and signature list. Moreover, GEOGLE summarizes the signature genes from a subset of GDSes and estimates the correlation between gene expression and the phenotypic distinction with an integrated p value. This approach performing global searching of expression data may expand the traditional way of collecting heterogeneous gene expression experiment data. GEOGLE is a novel tool that provides researchers a quantitative way to understand the correlation between gene expression and phenotypic distinction through meta-analysis of gene expression datasets from different experiments, as well as the biological meaning behind. The web site and user guide of GEOGLE are available at: http://omics.biosino.org:14000/kweb/

GsVIN: An analytical platform for genome-scale virus-human interaction network

Chunyan Li[1], Jia Sheng[2,3], Lulu Zheng[4,5], Yixue Li[3,5], Xuan Li[2,3], Pei Hao[1,3,*]

[1]: Pathogen Diagnostic Center, Institut Pasteur of Shanghai, Chinese Academy of Sciences.

[2]: Institute of Plant Physiology and Ecology, Chinese Academy of Sciences.

[3]: Shanghai Center for Bioinformation Technology.

[4]: Hubei Bioinformatics and Molecular Imaging Key Laboratory, Huazhong University of Science and Technology, Wuhan, Hubei, China.

[5]: Key Laboratory of System Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai, China.

[*]: To whom correspondence should be addressed.

Abstract: A virus-host interaction network offers a broad perspective on viral infection mechanism and disease etiology. While the concept of studying virus-host interaction network is not new, there are no applications or tools to facilitate such study, and it has not been performed on the genome scale. Therefore, we developed an analytical platform, GsVIN, to implement the analysis and comparison of the virus-host interaction networks for all human viruses. First, we collected 13,058 virus-host protein-protein interactions (PPIs) involving 674 viral proteins and 2,388 human proteins. Secondly, because the virus-host PPI data were concentrated on several most studied human viruses, i.e. HIV, HCV and EBV, we sought to expand the virus-host PPI networks by two methods: i) adding more proteins to the network nodes based on sequence similarity between proteins from different viruses; ii) using protein domain-domain interaction (DDI) data to construct PPIs between viral proteins and human ones, thus adding new nodes to existing networks. Thirdly, human protein atlas data were incorporated into GsVIN to make it possible to construct context-specific virus-human interaction networks. Lastly, we developed new functions/tools to view the virus-human interaction networks, and to compare multi-virus-human interaction networks or same-virus-human interaction networks of different tissues. Importantly, the analytical tool can apply GO enrichment technique to sort out the similar and the most different GO modules between virus-human interaction networks. It is particularly useful to shed light on the functions of newly emerging viral genomes. GsVIN analytical platform is freely accessible at http://www.viralportal.org/GsVIN/.

Prognostic Prediction for Locally Advanced Nasopharyngeal Carcinoma by Integration of Molecular and Pathological Markers via Machine Learning Techniques

Hongmin Cai[1,*], Xiangbo Wan[2], Ming-Huang Hong[2], Quentin Liu[2]

[1]: School of Computer Science and Technology, South China University of Technology, Guangzhou, China

[2]: Key Laboratory of Oncology in Southern China, The Sun Yat-sen University, Guangzhou, China

[*]: To whom correspondence should be addressed.

Emails: HM Cai (hmcai@scut.edu.cn);

ABSTRACT

Background: Accurate prognostication of locally advanced nasopharyngeal carcinoma (NPC) , besides standard TNM staging system, will benefit patients for tailored therapy. Since data from epithelial-mesenchymal-transition (EMT)-related biomarkers hold an enormous amount of biological information. We sought to develop a prognostic tool for NPC patients by integration of clinicopathologic variables and biomarker expression.

Methods: One hundred and thirty-six locally advanced NPC patients in a randomized controlled trial (RCT) with 5-year follow-up were studied. Various experiments by using classical machine learning techniques were conducted on this dataset. Feature selection method was firstly employed to find the variables subset with high prognostic potentials. Seven variables from thirty-eight tissue molecular biomarkers and one variable from eighteen clinicopathologic markers were selected. We designed three classification models based on Adaboost techniques and Support Vector Machine to refine prognosis of NPC with 5-year follow-up.

Results: The model for prediction of overall survival (OS) displayed a high predictive capacity (sensitivity, specificity were 88.0% and 90.3%, respectively) by integrating the expression of seven molecular biomarkers and one clinicpathologic variable. The model for failure free survival (FFS) possessed highly predictive specificity (sensitivity, specificity were 86.0% and 60%, respectively) by using the expression level of 8 molecular biomarkers. The third model for distant metastasis-free survival (DMFS) obtained a high predictive sensitivity and specificity (sensitivity, specificity were 90.9.0% and 73%, respectively).

We found that two biomarkers, Aurora-A and MMP9 had strong discrimination power in classification of prognosis. Survival analysis via Cox multivariate regression analysis confirmed these models were all the significant independent prognostic model for various survival states.

Conclusions: Integration of molecular biomarkers and clinicpathologic significantly improves the prognostic accuracy of survival stats in patients with NPC and facilitates the identification of patients with counseling and individualize management of patient.

Statistical Validation of Protein Quantification in Label-Free Quantitative Proteomics

Mingon Kang [1], Dong-Chul Kim [1], Jean Gao [1,*]

[1]: The University of Texas at Arlington

[*]: To whom correspondence should be addressed.

Emails: M. Kang (mingon.kang@mavs.uta.edu); D. Kim (dongchul.kim@mavs.uta.edu); J. Gao (gao@uta.edu)

Label-free proteomics is a promising technology to provide qualitative and quantitative high-throughput analysis for determining the differential expression level of proteins in proteomics. Label-free proteomics has been gaining interest due to its capacity for identifying and quantitating large complex biological samples. Protein quantification using mass spectrometry data plays a key role in analyzing proteins quantitatively and lays a foundation for further research such as biomarker discovery and signaling pathway construction in proteomics. In order to obtain significant protein biomarker candidates among thousands of proteins in the samples, a well-designed method to validate quantitative protein measurements is required as well as a high quality of protein identification and protein quantification. In this paper, we propose a statistical framework for validating protein quantitation using a fusion methodology (Dezert-Smarandache theory). This framework validates quantitative measurements of proteins with statistical models and computes protein ratios for comparative analysis. The experimental result with NCI-funded data shows this framework consequently reduces the quantification errors.

The Relationship between Experimentally Validated Intracellular Human Protein Stability and the Features of its Solvent Accessible Surface

Yan Jing[1], Ping Han[2, *], Xiaofeng Song[1, *]

1: Department of Biomedical Engineering, Nanjing University of Aeronautics and Astronautics, Nanjing 210016 China;

[2]: Department of Gynecology and Obstetrics, The First Affiliated Hospital with Nanjing Medical University, Nanjing 210029 China

[*]: To whom correspondence should be addressed.

Emails: PH(hanping200701@163.com); XFS(xfsong@nuaa.edu.cn);

Protein degradation is critical for most cellular processes, including cell cycle progression, signal transduction, and differentiation. In depth investigating the degradation signals in the protein sequence and structure is beneficial for analyzing the protein stability based on high-throughput dataset. Being similar to other protein function, the key residues exposed to the solvent surface and local specific structure such as cleft and pockets in the protein surface may play a pivotal role in protein degradation. In this paper, we investigated in depth the intrinsic factors affecting the protein degradation based on the sequence and structure features in the protein solvent accessible surface. The results indicated that there are more hydrophobic residues on the surface of short-lived protein than the long-lived protein; the secondary structure such as coil tends to be on the surface of short-lived protein; There are more serine phosphorylation sites on the short-lived protein surface; And there is higher possibility for the short-lived proteins to start the degradation by signal of PEST motif than long-lived proteins. We also found that almost all of N terminal residues in protein dataset are exposed to be on the surface; therefore the specific features of the solvent accessible surface residues are the key factors affecting intracellular protein stability.

Acknowledgements

The study was supported by grants from National Natural Science Foundation of China (No. 61171191, No. 81270700) and Natural Science Foundation of Jiangsu Province in China (BK2010500).

References

1. Yen Hsueh-Chi Sherry, Xu Qikai, Chou Danny M. et al. (2008) Global Protein Stability Profiling in Mammalian Cells. *Science*, 322(5903): 918-923
2. Tompa P, Prilusky J, Silman I. (2008) Structural disorder serves as a weak signal for intracellular protein degradation. *Proteins*, 71: 903-909.
3. Alexander Varshavsky. (1997) The N-end rule pathway of protein degradation. *Genes to Cell*, 2: 13-28.
4. Alessandro Pintar, Oliviero Carugo and S´andor Pongor. (2003) DPX: for the analysis of the protein core . *Bioinformatics* , 19(2): 313-314.
5. Stewart H Lecker, Alfred L Goldberg, William E Mitch. (2006) Protein Degradation by the Ubiquitin-Proteasome Pathway in Normal and Disease States. *JANS*, 17: 1807-1819

Core cancer proteome profiling of the NCI-60 cell line panel

Amin Moghaddas Gholami[1,*], Hannes Hahne[1], Zhixiang Wu[1], Florian Auer[1], Chen Meng[1], Mathias Wilhelm[1] and Bernhard Kuster[1,2*]

[1]: Chair of Proteomics and Bioanalytics, Technische Universität München, Freising, Germany

[2]: Center for Integrated Protein Science Munich

[*]: To whom correspondence should be addressed.

Emails: AMG (amin@tum.de); HH (hannes.hahne@tum.de); ZW (wu@wzw.tum.de); FA (florian.j.auer@gmail.com); CM (chen.meng@tum.de); MW (mathias.wilhelm@tum.de); BK (kuster@tum.de)

Advances in the high-throughput omic technologies have made it possible to profile cells in a large number of ways at different biological levels (DNA, RNA, Protein, Chromosomal, Functional, and Pharmacological). Cancer cell biology and response to drug treatment have benefited from new molecular technologies for integrating information from multiple sources. The NCI-60, a panel of 60 diverse human cancer cell lines, has been used by the National Cancer Institute to screen >100,000 chemical compounds for anticancer activity and has been extensively molecularly characterized.

To complement the existing NCI-60 datasets, we have measured global proteome profiling as well as kinase centric proteomics screen of the NCI-60 panel. Integration with transcriptome data and modeling drug response profiles for 108 FDA approved drugs identified known and potential novel protein markers for drug sensitivity and resistance. To enable community access to this unique resource, we incorporated it into a public database for comparative and integrative analysis.

Vibrational spectral signature of peptides with different secondary structures: new insight from molecular dynamics simulations with approximate density-functional theory

Xijun Wang, Soran Jahangiri, Gilles H. Peslherbe*

Center for Research in Molecular Modeling, Department of Chemistry and Biochemistry, Concordia University, Montreal, Canada, H4B 1R6 (xijun@cermm.concordia.ca)

*: To whom correspondence should be addressed.

Emails: X. W. (Xijun@cermm.concordia.ca); S. J. (soran@cermm.concordia.ca); G. H. P. (gilles@cermm.concordia.ca)

Vibrational spectroscopy plays a key role in probing the interactions between biological macromolecules and other species as well as their structural changes. The assignment of experimental spectral bands is usually based on empirical rules and can certainly benefit from the insight of molecular modeling. However, normal modes analysis based on quantum-chemical calculations do not include anharmonic effects and are restricted to small model systems. In this presentation, we demonstrate the performance and strength of molecular dynamics simulations with approximate density-functional theory in reproducing the infrared (vibrational) spectra of peptides in different secondary structures, i.e. α-helix, parallel and anti-parallel β-sheet, extended coil and turns, at finite temperature. The calculated band shapes and positions are qualitatively consistent with experimental data. With the help of group-specific vibrational-spectral decomposition, we are able to address the relationship between the vibrational frequencies of the amide modes and their chemical environment, i.e. with factors such as hydrogen-bonding strength and geometry. This work brings insight into the vibrational spectral signature of amide groups in diverse biological models, and the approach can be easily extended to larger systems at reasonable computational cost.

A count-based approach for discovery of translatome differences

Olga Nikolayeva [1,*], Knud Nairz [2], Alexander Kanitz [3], André P. Gerber [4] , Mark D. Robinson [1]

[1]: University of Zurich

[2]: ETH Zurich

[3]: University Basel

[4]: University of Surrey

[*]: To whom correspondence should be addressed.

Email: ON (olga.nikolayeva@uzh.ch)

Ribosome affinity profiling (RAP) is a rapid and accurate method allowing isolation of ribosomes and the associated mRNAs in S. cerevisiae. RAP has previously been used to demonstrate that in mild stress conditions, the translatome (mRNAs attached to ribosomes and in the process of translation) and the transcriptome become uncorrelated, implying that alteration of protein synthesis might be a prominent mechanism of modulating gene expression in response to mild stress.

In this study, we use RAP-RNA-seq (capture of full mRNAs followed by high throughput sequencing) to further examine yeast translatome behaviour by comparing expression of stressed and non-stressed translatomes. Specifically, wild type untreated translatome, wild type untreated transcriptome and DTT-treated translatome are compared by using a standard counting approach and an existing statistical framework (edgeR; Bioconductor) to assess "differential translation". Importantly, we cast the net of features as wide as possible, including cryptic unstable transcripts (CUTs) and stable uncharacterized transcripts (SUTs) and explore the role of untranslated regions (UTRs) as well.

A library of TALE-based transcription repressors in mammalian cells

Yinqing Li[1], He Chen[2*], Yun Jiang[2*], Zhihua Li[2,3], Zhen Xie[2]† and Ron Weiss[1]†

[1]:Department of Biological Engineering, Massachusetts Institute of Technology, 500 technology square, Cambridge MA 02139, USA

[2]:Center for Synthetic and Systems Biology, Bioinformatics Division, TNLIST, Tsinghua University, Beijing, China.

[3]:Institute of Medical Biology, Chinese Academy of Medical Sciences and Peking Union Medical College, 935 Jiaoling Road, Kunming, Yunnan, 650118..

[*]:These authors contribute equally.

†:Correspondence addresses to zhenxie@tsinghua.edu.cn, rweiss@mit.edu..

TALE (Transcription activator of-like effectors) is a DNA-binding protein found in the plant pathogen Xanthomonas spp. TALE DNA binding domain is composed by a string of highly conserved motifs in which the 12th and 13th amino acids residues determine a single nucleotide DNA recognition. Unlike zinc finger or other specific DNA binding domains, a concatenation of motifs can be designed to specifically bind to a given DNA sequence without the need of directed evolution. Artificial TALE transcription repressors have been demonstrated by fusing TALE DNA binding domains with chromatin remodeling domains, resulting in irreversible repression. However, many important biological investigations and bioengineering applications require a dynamic control of gene expression. In this study, we show that TALE proteins without a functional repression domain reversibly inhibit gene expression by using a transient reporter system in mammalian cells when two TALE DNA binding sites flank a DNA cis-element that is essential for transcription initiation. We also demonstrate that efficient repression requires optimal distance and orientation of TALE DNA binding sites. Based on these results, we have constructed a library of TALE-based repressors (TALERs) and their cognate promoters, and evaluated the pair-wise orthogonality of top ten strong repressors. This library of orthogonal transcription repressor will be a valuable tool kit for both basic biological research and biomedical applications.

References

1. Boch, J., Scholze, H., Schornack, S., Landgraf, A., Hahn, S., Kay, S., Lahaye, T., et al.(2009). Breaking the code of DNA binding specificity of TAL-type III effectors. *Science* 326(5959), 1509-12. doi:10.1126/science.1178811
2. Li, T., Huang, S., Jiang, W. Z., Wright, D., Spalding, M. H., Weeks, D. P., & Yang, B. (2011).TAL nucleases (TALNs): hybrid proteins composed of TAL effectors and FokI DNA cleavage domain. *Nucleic acids research*, 39(1), 359-72. doi:10.1093/nar/gkq704

Functional distinctive CTCF bindings revealed by a novel motif discovery pipeline

Rongxin Fang [1], Zhihua Zhang [1,*]

[1]: CAS Key Laboratory of Genome Sciences and Information, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 100029, China.

[*]: To whom correspondence should be addressed.

Emails:RF (fangrongxinwilliam@gmail.com); ZZ (zhangzhihua@big.ac.cn).

CTCF, the 11-zinc finger protein, is a key factor in regulating gene expression, DNA loop formation and maintain chromatin high order 3D architecture. It has been hypothesized that it exhibits distinct properties through binding various DNA sequence motifs by different combinations of the zinc fingers. To better understand the relationship between the binding patterns and its versatile properties, we reanalyzed the ENCODE data of CTCF chromatin immunoprecipitation followed by high-throughput sequencing (ChIP-seq) experiments in 14 cell types, obtaining 12,761 “Ubiquitous” CTCF binding events. We developed a novel motif discovery pipeline, by which three distinct core CTCF motifs were revealed. As strikingly differ in GC content, we termed the three motifs as HighGC-, MidGC- and LowGC-CTCF, respectively. We showed strong evidences that only the HighGC-CTCF behaved like enhancer and promoter. We also found the HighGC-CTCFs, in aid with P63 and several other cofactors, are more positively correlated with loop formation. The genes which interacted with the HighGC-CTCFs have significant higher expression levels. Long-range loops mediated by the HighGC-CTCFs are more conserved between tissues. We found the LowGC-CTCFs behaved like insulator, as they are frequently interacted with cohesin, and we demonstrated the LowGC-CTCFs are better elements to define function unit than the HighGC-CTCFs. As the MidGC-CTCFs are enriched in boundaries regions of topological domains, we showed evidences that the potential of the motifs to be barriers of the domains. Our results provide a comprehensive scope on the relationship between the CTCF sequence motifs and its versatile properties.

Consistent Phenotype Discrimination and Biomarker Discovery in Translational Bioinformatics

Xiaoxu Han[1*], Xiao-Li Li[2,], See-Kiong Ng[2]

[1]: Department of Computer and Information Science, Fordham University, New York NY 10458 USA

[2]: Institute for Infocomm Research, Agency for Science, Technology and Research (A*STAR), Singapore 138632.

[*]: To whom correspondence should be addressed.

Emails: X.Han (xhan9@fordham.edu); X.Li (xlli@i2r.a-star.edu.sg); S.Ng (skng@i2r.a-star.edu.sg)

While high-throughput technologies such as gene and protein expression microarray are expected to play a critical role in clinical translational research for complex disease diagnosis, the ability to accurately and consistently discriminate disease phenotypes by determining the gene and protein expression patterns as signatures of different clinical conditions have remained a challenge for translational bioinformatics.

In this study, we propose a novel multi-resolution feature selection algorithm: MultiResoluTion-test (*MRT-test*) that can produce significantly accurate and consistent phenotype discrimination across a series of gene and protein expression data. Our algorithm can capture those features contributing to subtle data behaviors instead of selecting the features contributing to global data behaviors. The advantage of our algorithm is essential to achieve high-performance diagnosis in translational bioinformatics, because the samples sharing the global characteristics but with similar local characteristics are usually hard to classify and lead to false positives and false negatives in diagnosis. We apply our *MRT-test* to *omics* data classification by combining it with the state-of-the-art classifiers and achieve exceptional results compared with the tradition methods. Moreover, our results demonstrate that high-dimensional *omics* data classification is actually a linear-separable problem under our technologies.

In addition, based on the seed biomarkers detected by the *MRT-test*, we design a novel subnetwork marker synthesis algorithm to decipher the underlying molecular mechanisms of tumorigenesis from a systems point of views. Unlike the existing top-down gene network building approaches, our subnetwork marker synthesis methods can not only avoid large computing overhead but also capture the essential bridge genes that connecting different subnetwork markers. Experimental evaluation showed that *MRT-test* based classification is able to generate consistent and robust clinical-level phenotype separation for various diseases. The biomarkers detected by the *MRT-test* and subnetwork marker synthesis algorithm seem to be able to provide biologically meaningful insights for understanding the genetic basis of complex diseases.

Real Time Classification of Viruses in 12 Dimensions

Troy Hernandez[1*], Chenglong Yu[1], Hui Zheng[1], Shek-Chung Yau[2], Hsin-Hsiung Huang[1], Rong Lucy He[3], Jie Yang[1], Stephen S.-T. Yau[4]

[1]: University of Illinois at Chicago.

[2]: The Hong Kong University of Science and Technology

[3]: Chicago State University

[3]: Chicago State University

[4]: Tsinghua University

[*]: To whom correspondence should be addressed.

Emails: TH (Troy.A.Hernandez@gmail.com); CY (clyu7@uic.edu); HZ (hzheng9@uic.edu)

The International Committee on Taxonomy of Viruses authorizes and organizes the taxonomic classification of viruses. The detailed classifications for all viruses are neither complete nor free from dispute. For example, the current missing label rates in GenBank are 12.1% for family label and 30.0% for genus label. Using the proposed Natural Vector representation, all 2,044 single-segment referenced viral genomes in GenBank can be embedded in R^{12}. Unlike other approaches, this allows us to determine phylogenetic relations for all viruses at any level (e.g., Baltimore class, family, subfamily, genus, and species) in real time. Based on cross-validation results, the accuracy rates of our predictions are as high as 98.2% for Baltimore class labels, 96.6% for family labels, 99.7% for subfamily labels and 97.2% for genus labels.

Recursive Longest Common Subsequence: A Novel Similarity Measure for Sequences

Ribel Fares [1], Byron J. Gao [1,*]

[1]: Texas State University–San Marcos.

[*]: To whom correspondence should be addressed.

Longest common subsequence (LCS) is a standard similarity measure for sequences. We argue that LCS has a weakness as it ignores the structural commonality in LCS residues, failing to represent the full-spectrum similarity between sequences. This is especially the case for sequences with large alphabets, where the LCS is usually very short and a large portion of sequences becomes irrelevant.

LCS has a known weakness in capturing structural commonality. Let us consider the sequences S1 = {A, A, A, B, B}, S2 = {B, B, A, A, A}, and S3 = {C, C, A, A, A}. Clearly, S1 and S2 are more similar to each other than S1 and S3 are. However, LCS_{length}(S1, S2) = LCS_{length}(S1, S3) = 3. To address this weakness, we introduce recursive longest common subsequence (rLCS) that generalizes LCS by aggregating the structural commonality of LCS residues recursively. Applying rLCS to the same sequence pairs above, rLCS(S1, S2) would return {A, A, A} in the first recursion, and {B, B} on the second. On the other hand, rLCS(S2, S3) would only make one recursion, terminating after the subsequence {A, A, A} is removed from both sequences. An rLCS score that captures a full-spectrum similarity could then be calculated as a weighted sum of the LCS lengths at each recursion. One way to compute the rLCS score could be via equations 1 and 2, given the LCS function randomly returns a single LCS if there is more than one.

$$rLCS_{length}(S_1, S_2) = \begin{cases} 0, & \text{if } LCS_{length}(S_1, S_2) = 0 \\ rLCS_{length}\left(S_1 \setminus LCS(S_1, S_2), S_2 \setminus LCS(S_1, S_2)\right) + \dfrac{LCS(S_1, S_2)}{recursion\_number}, & \text{otherwise} \end{cases} \quad (1)$$

$$rLCS(S_1, S_2) = \frac{\left(rLCS_{length}(S_1, S_2)\right)^2}{\min(|S_1|, |S_2|)\ \max(|S_1|, |S_2|)} \quad (2)$$

At each recursion, the sequence lengths are reduced, thus the computation is guaranteed to terminate. The rLCS score calculated by equations 1 and 2 gives a similarity score between 0 (least similar) and 1 (most similar).

We ran kNN classification experiments using 6 UCI benchmark datasets. Specifically, we used the Balance Scale, Reuters Transcribed Subset, 20 Newsgroups (40 samples per class), Farm Ads (100 random instances), Mushroom (200 random instances), and Australian Sign Language (2 classes) datasets. We ran 1NN, 5NN, 9NN, 13NN, and 17NN for all of the datasets using both LCS and rLCS as the similarity measure. On average, the kNN accuracy for rLCS was 5.3% higher than for LCS.

Performance Assessment of BLAST and H-Tuple Methods in Comparison of Biological Sequences Using the ROC Curve

Afshin Fayyaz movaghar *[1] and Musa Ghahremanzadeh Barugh[1]

[1]Department of Statistics University of Mazandaran, IRAN

An important step in learning the function of a new biological sequence (DNA or protein) is to compare the new sequence with existing sequences belonging to a database whose biological functions are known. To compare these sequences, there are different methods such as BLAST [1] and H-tuple [2]. These methods derive a statistical significance based on a computed gapped local score.

H-tuple method stands on combining an adapted scoring scheme that includes the gaps and an approximate distribution of the ungapped local score of two independent sequences of i.i.d. random variables. The new scoring scheme is defined on h-tuples of the sequences, using the gapped global score. Then, p-value of gapped alignment be derived from the one of ungapped case, proposed by Mercier and Daudin [3].

Comparing performance of the BLAST and h-tuple method is the aim of our work. For this, the receiver operating characteristics (ROC) curve is employd, that states the relationship between sensitivity and specificity of a binary classifier. A ROC curve is a technique for visualizing, organizing and selecting classifiers based on their performance. We apply the methods on the SCOP1.75 detabase and the ROC corve of classification results. Shows the performance of h-tuple method against the BLAST one.

Keywords: ROC curve, Compare biological sequences, BLAST, H-Tuple method.

References

[1] Altschul, S., Gish, W., Miller, W., Myers, E. & Lipman, D. (1990). Basic Local Alignment Search Tool. *Journal of Molecular Biology, 215,* 403-410.

[2] Fayyaz movaghar, A., Mercier, S. & Ferre, L. (2007). H-Tuple Approach to Evaluate statistical Significance of Biological Sequence Comparison with Gaps. *Statistical Applications in Genetics and Molecular Biology, 6*, Iss. 1, Art. 22.

[3] Mercier, S. & Daudin, J. (2001). Exact distribution for the local score of one i.i.d. random sequence. *Journal of Computational Biology, 8*, 373-380.

*A_fayyaz@umz.ac.ir; Corresponding author

Representation of Protein Complexes as Multilayer Graphs

Nadav Rappoport [1], Nathan Linial [1], Michal Linial [2,*]

[1]: School of Computer Science and Engineering, The Hebrew University of Jerusalem, Israel.

[2]: Department of Biological Chemistry, Institute of Life Science, The Hebrew University of Jerusalem, Israel.

[*]: To whom correspondence should be addressed.

Emails: NR (nadavrap@cs.huji.ac.il); NL (nati@cs.huji.ac.il); ML (michall@cc.huji.ac.il)

The basic machineries in all forms of life are the proteins complexes (PCs). Some PCs are found in all organisms (*e.g.*, replisome, ribosome) while others PCs are specific to a phylogenetic branch (*e.g.*, photosystem 2) or specific cells (*e.g.*, exosome). Protein-protein interaction technologies allowed cataloguing stable (coined obligatory) PCs. The number of PCs increases with the complexity with 180, 500 and 600 PCs found in *E. coli*, yeast and human, respectively. PC machines function to execute a function such as activation of gene expression, electron transfer, degradation and more. Herein, we transform each obligatory CP to a graph where the nodes are the proteins and the edges represent the physical interactions between the nodes. However, CPs as 'cellular machines' may carry an intrinsic dynamic component. A transformation of a static graph to a series of dynamic representations captures the ability of the CP proteins to vary (*e.g.*, coding SNP, phosphorylation). Eventually, a variability on every node in the graph leads to an exponential number of graph variants. In our work, we model such multilayered graphs of CPs and classify them according to the dynamic robustness as an internal measure among CP graphs. Using the known outcome of the CPs to variation, we aim to rank the key nodes in any obligatory CP. Moreover, a gain in CP complexity along evolution will be tested by comparing the properties of the CP multilayered graphs. We illustrate our model on the proteasome and the transcription complex of RNA polymerase II.

Statistical Significance of Comparison Between a Protein Sequence and a 3D Structure

Afshin Fayyaz movaghar *[1] and Milad Asadi[1]

[1]Department of Statistics University of Mazandaran, IRAN

The three-dimensional structure of a protein is the mediator between its sequence and its function[1]. So comparing the sequence of a protein with a known three-dimensional structure,for realizing the function of a protein is effective.Threading is a tool for recognize a remote relationship between a query protein and a protein of known three-dimensional structure. since the threading score alone can not evaluate the remote relationship, So we need to check the significance of these score that in this thesis we investigate it. Here we show numerically that threading score distribution is generalized pareto distribution and this distribution fitted to threading score obtained from simulations. we also offer a method to calculate the threshold for generalized Pareto distribution fitted that Before fitting the threshold can be calculated.

Key words: Threading, Generalized Pareto Distribution, Fold, threshold, Three-Dimensional.

References

[1] Fayyaz Movaghar, A., Launay, G., Schbath, S., Gibrat, J.-F. & Rodolphe, F. (2012). Statistical Significance of Threading Scores. *Journal of Computational Biology 19*, 1-18.

*A_fayyaz@umz.ac.ir; Corresponding author

Coalescent-based Estimation of Population History in the Presence of Admixture from Genome-Scale Variation Data

Ming-Chi Tsai [1], Guy Blelloch [2], R. Ravi [3], Russell Schwartz [4,*]

[1]: Joint Carnegie Mellon/University of Pittsburgh Ph.D. Program in Computational Biology.

[2]: Computer Science Department, Carnegie Mellon University.

[3]: Tepper School of Business, Carnegie Mellon University.

[4]: Department of Biological Sciences and Lane Center for Computational Biology, Carnegie Mellon University.

[*]: To whom correspondence should be addressed.

Emails: M-CT (mingchit@andrew.cmu.edu); GB (guyb@cs.cmu.edu); RR (ravi@andrew.cmu.edu); RS (russells@andrew.cmu.edu);

Learning precisely how the human population is structured and how different subgroups are related to one another is not only a fundamental issue in understanding human origins, but also has practical relevance to improving models of genome evolution. Despite considerable attention to the general problem of identifying population substructure in large-scale variation data, the field lacks automated methods for reconstructing the relationships among population subgroups and inferring correct orders and timing of events in the presence of admixture. We describe here a novel two-step approach for inference of quantitative population history in the presence of admixture from large variation datasets. The method first identifies a set of phylogenetic splits that are likely to have occurred during the evolutionary history and assign each observed variation site to one of the splits. The resulting split set and the number of variation sites assigned to each split are then used to infer a model of population-level evolution using a coalescent-based Markov chain. This population model describes inferred times of divergence and admixture events as well as admixture proportions for the population history. Evaluation on simulated three- and four-population data sets suggest that fully automated reconstruction of population histories in the presence of admixture is feasible, although further algorithmic improvements may be needed to infer more complicated scenarios.

Echo: Evolutionary CHaracterization of fixed-length biological sequence mOtifs

Miaomiao Zhao [†,1], Zhao Zhang [†,1,2], Guoqin Mai [†,1], Youxi Luo [1,3], Fengfeng Zhou [*,1]

[1]: Shenzhen Institutes of Advanced Technology, and Key Laboratory of Health Informatics, Chinese Academy of Sciences, Shenzhen, Guangdong, China, 518055.

[2]: Tianjin Polytechnic University, Tianjin, China, 300387.

[3]: School of Science, Hubei University of Technology, Wuhan, Hubei, China, 430068.

[*]: Corresponding author: Fengfeng Zhou, Emails: FengfengZhou@gmail.com or ff.zhou@siat.ac.cn. Web site: http://www.HealthInformaticsLab.org/ffzhou/ .

[†]: These authors contribute equally to this work.

Detection of fixed-length biological sequence motifs is a classic bioinformatics problem, with applications in both nucleotide and peptide sequence motifs. For DNA sequences, transcription factor binding sites (TFBSs) are assumed to be fixed-length for a given transcription factor, and a position weight matrix (PWM) is usually built based on the known TFBSs. For peptide sequences, the fixed-length flanking sequences of the post-translational modification (PTM) residues are assumed to harbor the selective signals for the catalytic enzymes, and a neural network or support vector machine model may be built from the known PTM residues and their flanking sequences. One of the major obstacles for the fixed-length motif finding problem is the unsatisfying screening accuracies. This study proposed an evolutionary algorithm (Echo) to iteratively optimize the fixed-length motif finding model. We chose the predictions of SREBP binding motifs and phosphorylation modification residues as examples to illustrate Echo's prediction performance. Sterol regulatory element binding proteins (SREBPs) are transcription factors (TFs) involved in the lipid balance regulation, by controlling the expression of synthesis enzymes for endogenous cholesterol and fatty acid. Echo outperforms the widely used Position Weight Matrix (PWM) algorithm by 2-30% in sensitivity at the similar level of specificity. Echo also achieves at least 97.8% for both sensitivity and specificity for all the four SREBPs. For the phosphorylation site prediction of the kinases MAPK14, Abl and S6K, Echo outperforms one of the best algorithms, GPS 2.1, by over 20% in the sum of sensitivity and specificity. Large-scale further optimization by parallel computing and cloud computing are underway to obtain better prediction performance.

Key words:

Fixed-length motif finding, evolutionary algorithm, transcription factor binding site, SREBP, post-translational modification.

Evaluating the statistical significance of rare protein modifications detected by high-throughput

Yan Fu

National Center for Mathematics and Interdisciplinary Sciences, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China

Email: yfu@amss.ac.cn

High-throughput mass spectrometry enables systematic identification of proteins and their post-translational modifications. Current proteomic experiments and followed data analysis produce thousands to millions of hypothetical peptide identifications. The common way to control the false discovery rate (FDR) of peptide identifications is the target-decoy database search strategy, which is accurate for large data sets only. On the other hand, the legality of the target-decoy strategy for modification-centric studies has not been rigorously discussed. Because the FDRs of modified and unmodified peptides may be dramatically different at the same score threshold, the FDR of modified peptides should be separately estimated, instead of in combination with unmodified ones. However, low-abundance modifications may result in a very small number of modification identifications from a large set of mass spectra, making the direct separate FDR estimation very inaccurate. This work presents a method, still based on the target-decoy strategy, for accurate FDR estimation for protein modifications in arbitrary abundances. The proposed method is validated on both simulated and real mass spectral data.

References

1. Yan Fu. Bayesian false discovery rates for post-translational modification proteomics. *Statistics and Its Interface*, 5:47–59, 2012.
2. Yan Fu. False discovery rate control for post-translational modifications detected by mass spectrometry. To appear, 2013.

Protein Side-Chain Prediction and Inference Using Continuous Variables

Laleh Soltan Ghoraie [1,*], Forbes Burkowski [1], Mu Zhu [2]

[1]: University of Waterloo, Cheriton School of Computer Science

[2]: University of Waterloo, Department of Statistics and Actuarial Science

[*]: To whom correspondence should be addressed.

Emails: L.S.G. (lsoltang@uwaterloo.ca); F.B. (fjburkow@uwaterloo.ca); M.Z. (m3zhu@uwaterloo.ca)

The protein folding problem is of crucial importance to bioinformatics. Side chain prediction is an important subtask, and many attempts have been made to tackle it. Given a protein sequence, the objective is to find the minimum-energy conformations for the side chains at the various residue sites. A common approach is to discretize the search space by considering a list of distinct rotamers at each residue. However, the space of all possible rotamer combinations is exponentially large, and the search problem is NP-hard. In this work, we avoid discretizing the search space altogether and propose a new framework, in which the side chain conformations are represented by continuous variables. To find the minimum-energy conformation, we model the protein by a Markov random field, approximate the node- and edge-potential functions by mixtures of von-Mises distributions, and apply an approximate continuous inference method called Particle Belief Propagation. For some proteins whose native structures are known to contain conformations that are not covered by the discrete rotamer library, our method has already achieved some successes – we found closer-to-native conformations than some state-of-the-art methods such as SCWRL.

Plot3: A Online Data Management, Exploration and Visualization Platform

Robert Edwards[1*], David Lougheed[2], Benoit Valin[1], Guillaume de Lazzer[1]

1. Research and Development Solutions Private Limited, #33-03, Hong Leong Building, 16 Raffles Quay Singapore 048581
2. Loyalist Collegiate and Vocational Institute, 153 Van Order Drive, Kingston, Ontario, K7M 1B9, Canada

* Email: r@rnd.sg

Plot3 (www.plot3.com) is a web based scalable-vector-graphics data-visualization platform tailored to address the complexity associated with Big Data. Natively capable of managing, processing, exploring and analyzing both static and live data, Plot3 aims to seamlessly integrate federated data sources (from databases, instrumentation and analytical platforms) into a single unified data structure.

Through the creation of custom data exploration dashboards, users can explore their data through rich, interactive and self-updating analysis and visualization widgets, such as graphs, plots and tables. These dashboards can be saved and shared with others through a collaboration work-space or through graphic-rich and interactive presentations.

Plot3 can also be made to integrate with a myriad of analysis, modeling and simulation platforms, such as Accelrys Pipeline Pilot, Discovery Studio and Material Studio; with knowledge exploration and management platforms such as Linguamatics I2E and Electronic Lab Notebooks, as well as with Open Street Map to explore geo-spatial data.

Plot3 is a free to use Python and Javascript based web application, its functionalities can be enhanced and extended by users, to integrate new data types and to create new visualization widgets, without any software programing.

Structural analysis of B-cell epitopes and protein binding pockets

Jens Vindahl Kringelum[1*], Olivier Taboureau[1] and Ole Lund[1]

[1]: Center for Biological Sequence Analysis, Institute of Systems Biology, Technical University of Denmark, Copenhagen Denmark

[*]: To whom correspondence should be addressed.

Emails: jk (jkgm@cbs.dtu.dk); ot (otab@cbs.dtu.dk); ol (lund@cbs.dtu.dk)

The interaction between antibodies and antigens is one of the most important events of the immune system and is involved in multiple mechanisms including allergic reactions and clearing of infectious organisms. Antibodies bind to antigens at sites known as antigenic determinant regions, which are also referred to as B-cell epitopes. The precise location of B-cell epitopes on the antigen surface is essential in the development of several biomedical application such as; rational vaccine design, disease diagnostic, immune-therapeutics and potentially in assessment of protein immunogenicity and alleginicity. However, experimental mapping of the epitope area is costly and time consuming, thereby making in silico methods an appealing alternative. To date, the performance of methods for in silico mapping of B-cell epitopes has been moderate, which to some extent can be explained by our incomplete understanding of what constitute a B-cell epitope. Here, we present a method that enables structural superimposing of unrelated B-cell epitopes and describes the size, shape and spatial amino acid distribution of epitopes and they related paratopes [1]. Furthermore, by applying the method on protein pockets known to bind small molecules we illustrate the molecular basis for off-target activity of small molecule drugs.

References

1. J. V. Kringelum, M. Nielsen, S. B. Padkjær, and O. Lund, 'Structural analysis of B-cell epitopes in antibody:protein complexes', *Molecular Immunology*, vol. 53, no. 1–2, pp. 24–34, Jan. 2013.

A nonparametric model of haplotypes in admixed populations

Lloyd T. Elliott [1,*], Yee Whye Teh [2]

[1]: Gatsby Unit, University College London, 17 Queen Square, London WC1N 3AR, U.K.

[2]: Department of Statistics, University of Oxford, 1 South Parks Road, Oxford OX1 3TG, U.K.

[*]: To whom correspondence should be addressed.

Emails: LTE (elliott@gatsby.ucl.ac.uk); YWT (y.w.teh@stats.ox.ac.uk)

Clustering methods based on hidden Markov models (HMMs) are useful for approximating the haplotype structure of admixed populations. In such models, latent population indicators can specify the ancestral populations (clusters) from which a given set of loci originate. In this work, we present ADAMIX: a Bayesian nonparametric model of haplotype structure based on a nonhomogeneous version of the hierarchical Dirichlet process hidden Markov model (HDP-HMM). ADAMIX can be seen as a nonparametric generalization of the popular fastPHASE model, and as such it well describes genetic sequences in which the population proportions of an unspecified number of admixed ancestral populations vary along the chromosome.

Bayesian nonparametric statistics provide a class of flexible distributions and inference methods in which the latent parameter space grows with the size of the data. The Dirichlet process is a Bayesian nonparametric prior which is natural for allele sampling processes[1]. By using the Dirichlet process, ADAMIX can learn the number of admixed populations.

In the ADAMIX model, the genetic material at each locus of each genetic sequence in the observed data is assigned to a latent cluster representing the ancestral population from which that material originates. Between each pair of consecutive loci the latent cluster of each sequence transitions according to an HMM such that the latent cluster either remains the same, or, with a probability proportional to a nonhomogeneous jump rate, is chosen randomly according to the population proportions of a Dirichlet process specific to that locus. The jump rate represents the relative rates of recombination events occurring in the ancestry of the population. To ensure that the resulting model is well defined, the Dirichlet processes at each location are linked through a hierarchical Dirichlet process, forming a nonhomogeneous version of the HDP-HMM.

Previous work has recognized the possibility of using an HDP-HMM as a genetic model[2]. But [2] does not take into account the nonhomogeneity of population proportions that is often observed in admixed populations. Further, unlike fastPHASE, [2] learns different HMM transition probabilities for each cluster, whereas dependence in the proportions of admixed populations tends to be shared across all clusters.

In a series of experiments involving data simulated from the coalescent with recombination in which populations are admixed, we compare the performance of ADAMIX and fastPHASE.

References

1. W. Ewens. The sampling theory of selectively neutral alleles. *Theoretical population biology*, 3(1), 1972.
2. K.-A. Sohn, E. P. Xing. Hidden Markov Dirichlet process: modeling genetic inference in open ancestral space. *Bayesian Analysis*, 2(3), 2007.

Identity by Descent in Admixed Populations

Itamar Eskin [1,*], Eran Halperin [2,3,4]

[1]: Applied Mathematics Department, School of Mathematical Sciences, Tel-Aviv University.

[2]: Computer Science Department, Tel-Aviv University.

[3]: International Computer Science Institue, Berkeley, California.

[4]: Molecular Microbiology and Biotechnology Department, Tel-Aviv University.

[*]: To whom correspondence should be addressed.

Emails: IE (itamares@post.tau.ac.il); EH (heran@post.tau.ac.il)

The concept of Identity By Descent (IBD) has gained considerable focus over the past several years, with applications in population-based linkage analysis (IBD mapping), haplotype phase inference, genotype imputation, and inference of population structure. Standard techniques for IBD detection require a careful estimation of shared haplotype frequency in order to distinguish true IBD from merely common haplotypes. This remains particularly challenging when relatively short IBD segments (< 5cM) are considered, and when the samples are obtained from structured populations such as recently admixed populations (e.g. African Americans).

Presented here is a novel approach that leverages the structure of linkage disequilibrium (LD) in the ancestral populations, allowing higher power to detect short tracts of IBD, while controlling for the false-positive rate. This involves jointly modeling the local ancestry and IBD status of pairs of individuals, based on available reference panels of the ancestral populations. We have successfully applied this approach on simulated data, and claim that it is beneficial not only for IBD detection, but also for improving the accuracy of local ancestry inference.

Computational Biology in eTRIKS

On behalf of the eTRIKS consortium: Ioannis Pandis, Ibrahim Emam, Xiang Yang and Yi-ke Guo*

Discovery Sciences Group, Department of Computing, Imperial College London, UK

Corresponding author: y.guo.at.imperial.ac.uk

eTRIKS (European Translational Information & Knowledge Management Services) is an EU Innovative Medicines Initiative (IMI) project aiming to provide a cloud based knowledge management (KM) platform and service infrastructure capable of the efficient storage and effective analysis of experimental data from studies in man, in animals and in pre-clinical models. http://www.transmartproject.org/

The main eTRIKS objectives are:

1. **Service**: Deploy and host the eTRIKS platform based on the tranSMART[1] technology and provide training, support and consultation activities to all IMI project partners on using the platform.
2. **Platform**: Develop and maintain a sustainable, interoperable, collaborative, re-usable, open source and scalable translational research (TR) KM platform, as well as conduct research & development into effective analytics methods and tools to support TR and computational molecular biology research in general, thus evolving and extending the platform with tools for omics, imaging data and text analysis that can leverage cloud-based operations for system biology research.
3. **Content**: Establish eTRIKS as a unique European TR data resource supporting cross-organisation TR studies, including clinical studies and pre-clinical studies, omics data analysis for biomarker discovery and validation, genetics and NGS studies and populate eTRIKS with existing and active data from TR studies and supporting the integration of standardised legacy TR study data.
4. **Community**: Promote and lead an active international TR analytics & informatics community, centred around eTRIKS, through active stakeholder engagement and by disseminating tools and expertise worldwide and engage in, and influence, international standardisation activities in areas relating to TR informatics.

In this poster, we will summarise the computational molecular biology aspect of the project, especially in the area of cross omics data analysis and modelling support.

This work is funded by the Innovative Medicines Initiative (IMI) EU program.

CAGI: The Critical Assessment of Genome Interpretation, a community experiment to evaluate phenotype prediction

Steven E. Brenner[1*], Susanna Repo[1,3], John Moult[2], CAGI participants

[1]: Department of Plant and Microbial Biology, University of California, Berkeley, CA 94720
[2]: IBBR, University of Maryland, Rockville, MD 20850
[3]: Currently at: EMBL-EBI, Wellcome Trust Genome Campus, Hinxton, Cambridgeshire, CB10 1SD, UK
[*]: To whom correspondence should be addressed.

Email addresses: SEB (Brenner@compbio.berkeley.edu); SR (srepo@compbio.berkeley.edu); JM (jmoult@umd.edu)

The Critical Assessment of Genome Interpretation (CAGI, \'k**ā**-j**ē**\) is a community experiment to objectively assess computational methods for predicting the phenotypic impacts of genomic variation. In this assessment, participants are provided genetic variants and make predictions of resulting phenotype. These predictions are evaluated against experimental characterizations by independent assessors. The CAGI experiment culminates with a community workshop and publications to disseminate results, assess our collective ability to make accurate and meaningful phenotypic predictions, and better understand progress in the field. A long-term goal for CAGI is to improve the accuracy of phenotype and disease predictions in clinical settings.

The CAGI 2011 experiment consisted of 11 diverse challenges exploring the phenotypic consequences of genomic variation. In two challenges, CAGI predictors applied the state-of-the-art methods to identify the effects of variants in a metabolic enzyme and oncogenes. This revealed the relative strengths of each prediction approach and the necessity of customizing such methods to the individual genes in question; these challenges also offered insight into the appropriate use of such methods in basic and clinical research. CAGI also explored genome-scale data, showing unexpected successes in predicting Crohn's disease from exomes, as well as disappointing failures in using genome and transcriptome data to distinguish discordant monozygotic twins with asthma. Complementary approaches from two groups showed promising results in predicting distinct response of breast cancer cell lines to a panel of drugs. Predictors also made measurable progress in predicting a diversity of phenotypes present in the Personal Genome Project participants, as compared to the CAGI predictions from 2010.

Current CAGI experiments are presently underway and, further information is available at the CAGI website at http://genomeinterpretation.org.

Modelling Translation

Dominique Chu [1] and Tobias von der Haar [1]

[1]: University of Kent, CT2 7NF Canterbury, UK.

Proteomes evolve under many different constraints including the minimization of the energy required to produce them, and the establishment of biochemical networks that optimize metabolic fluxes or increase fitness by other means. Another constraint, which is not widely studied, is that viable proteomes must be producible with a limited gene expression machinery. Gene expression is in essence a two-step process, whereby protein templates are produced during transcription, and the proteins proper during translation. Although specific limitations apply at every level, translation is overall the more resource intense step. The main components of the translation machinery are tRNAs, mRNAs and ribosomes. Particularly the latter are very costly to produce for the cell and have been proposed to limit gene expression and cell growth as a whole.

The optimality of a particular proteome is not only a function of its environment, but will also depend on its metabolic maintenance costs. It appears to be generally accepted knowledge that cell resources somehow limit the achievable proteomes, yet at present we do not have a detailed understanding of this limitation. For example, it seems to be a widely-held belief that initiation is a major limiting factor of translation. Similarly, it is also sometimes stated that tRNA availability is a limiting factor of translation. While these statements are normally inferences of from particular experiments, it has never been systematically investigated to what extent these insights are compatible with our wider quantitative and qualitative understanding of translation. At least for translation it is now possible to do this. For some of the best studied organisms we have a wealth of numerical information. What is lacking so far is a unifying framework that would allow us to understand what these detailed, but ultimately disparate, pieces of information entail and how they relate to one another.

In this contribution we do precisely this. We present a dynamical and stochastic model of translation in baker's yeast (*Saccharomyces cerevisiae*) that encodes the best known data about this organisms and enables us to combine the pieces of the jigsaw-puzzle into a coherent and dynamical picture of translation.

Our key findings include: (I) Ribosome-ribosome interactions (traffic jams) are likely to be at most a sub-ordinate effect in translation; (II) According to best estimates of the availability of tRNA availability is not limiting. Instead, tRNA seems to be available in large excess; (III) Limitation through initiation is more complicated than appears at first. The statement that "translation is limited by initiation" is imprecise in that it could mean that it is limited by a lack of ribosomes, a low affinity of ribosomes to bind to the 5'-cap or crowding at the 5' UTR. Detailed simulations suggest that the latter is likely to be a minor effect only. Similarly, available data suggests that ribosome affinity is not substantially limiting translation, at least not globally. However, the number of ribosomes is a limiting factor for the translation rate in yeast.

A Model Based Approach for Analysis of Spatial Structure in Genetic Data

Wen-Yun Yang[1,2], John Novembre[1,3], Eleazar Eskin[1,2,4,7,8], and Eran Halperin[5,6,7]

[1]Interdepartmental Program in Bioinformatics,
[2]Department of Computer Science,
[3]Department of Ecology and Evolutionary Biology and
[4]Department of Human Genetics, University of California, Los Angeles, California, USA
[5]International Computer Science Institute, Berkeley, California, USA
[6]School of Computer Science and the Department of Molecular Microbiology and Biotechnology, Tel-Aviv University, Tel-Aviv, Israel
[7]These authors contribute equally to this work
[8]Correspondence should be addressed to E.E. (eeskin@cs.ucla.edu)

- Published Journal: Nature Genetics 44, 725 - 731 (2012)
- URL: `http://www.nature.com/ng/journal/v44/n6/abs/ng.2285.html`

Abstract

Characterizing genetic diversity within and between populations has broad applications in studies of human disease and evolution. Here, we describe a new approach, spatial ancestry analysis (SPA), for the modeling of genetic variant in two- or three-dimensional space. We explicitly model the spatial distribution of the allele frequency of each genetic variant as a continuous function over a geographic region. We show how this modeling allows for the accurate prediction of an individuals location of origin within several hundred kilometers. SPA can also predict the location of origin of the parents of individuals with mixed ancestry. SPA can also identify genes which are under selection in human populations by identifying the locations of the genome which have dramatic changes in frequency over a geographic region. The SPA software is available in our website http://genetics.cs.ucla.edu/spa.

List of Authors

List of Authors

Brun, Christine	TAGC, Inserm	France
Bryant, A William	Imperial College, London	UK
Bryant, William	Imperial College, London	UK
Burkowski, Forbes	University of Waterloo	Canada
Cai, Hongmin	South China University of Technology	China
Cai, Jun	Beijing Institute of Genomics, CAS	China
Cairns, J. Murray	University of Newcastle	Australia
Caiyan, Jia	Beijing Jiaotong University	China
Califano, Andrea	Columbia University	USA
Cao, Changchang	Southeast University	China
Cao, Lihua	Beijing Institute of Genomics, CAS	China
Chang, Suhua	Institute of Psychology, CAS	China
Chen, Chen	Academy of Mathematics and Systems Science, CAS	China
Chen, Chun-Long	CENTRE DE GENETIQUE MOLECULAIRE	France
Chen, Gang	BGI-Shenzhen	China
Chen, He	Tsinghua University	China
Chen, Hsuan-Yu	Institute of Statistical Science, Academia Sinica	Taiwan
Chen, Jiao	Southeast University	China
Chen, Ke	Tianjin Polytechnic University	China
Chen, Liang	University of Southern California	USA
Chen, Shijian	Academy of Mathematics and Systems Science, CAS	China
Chen, Ting	University of Southern California	USA
Chen, Tong	Institute of Genetics and Developmental Biology, CAS	China
Chen, Yang	Tsinghua University	China
Chen, Yi	Fudan University	China
Chen, Yi-Hau	Institute of Statistical Science, Academia Sinica	Taiwan
Cheng, Jeffrey	UCSF	USA
Cheng, Jiefeng	Shenzhen Institutes of Advanced Technology, CAS	China
Chi, Hao	Chinese Academy of Sciences	China
Cho, Dong-Yeon	NCBI	USA
Choi, Pui Kwok	National University of Singapore	Singapore
Chong, Lauren	Ontario Cancer Institute	Canada
Choudhury, Roy Kingshuk	University College Cork	Ireland
Chu, Dominique	University of Kent	UK
Chu, Kwang-Hua	Inner Mongolia University of Science and Technology	China
Chu, Xiaowen	Hong Kong Baptist University	Hong Kong
Clote, Peter	Boston College	USA
Conde, Lucia	University of California, Berkeley	USA
Coronnello, Claudia	University of Pittsburgh	USA
Cosentino, Salvatore	Technical University of Denmark	Denmark
Costello, Joseph	UCSF	USA
Dai, Huanqin	Institute of Microbiology, CAS	China
D'Aubenton-Carafa, Yves	CNRS	France
Dave, Paul	Computation Institute	USA
Davila, Jaime	Mayo Clinic	USA
Dawson, John	University of Wisconsin Madison	USA
Deelman, Ewa	Information Science Institute	USA
Deng, Chao	Peking University	China
Deng, Lei	Tongji University	China
Deng, Minghua	Peking University	China
Di, Chao	Tsinghua University	China
Ding, Xiaoming	Fudan University	China
Djekidel, Nadhir Mohamed	Tsinghua University	China
Donald, Bruce	Duke University	USA
Doncheva, T. Nadezhda	Max Planck Institute for Informatics	Germany
Dong, Lili	Beijing Institute of Genomics, CAS	China
Dong, Xiaobao	China Agriculturial University	China
Dotu, Ivan	Boston College	USA
Dubchak, Inna	Lawrence Berkeley National Lab	USA
Durbin, Richard	Wellcome Trust Sanger Institute	UK
Eckloff, Bruce	Mayo Clinic	USA
Edwards, Robert	Research and Development Solutions Private Limited	Singapore
Elledge, Stephen	Harvard Medical School	USA
Elliott, Lloyd	University College London	UK
El-Mashtoly, Samir	Ruhr University Bochum	Germany

Emam, Ibrahim	Imperial College London	UK
Eskin, Eleazar	University of California, Los Angeles	USA
Eskin, Itamar	Tel-Aviv University	Israel
Fan, Zhonghua	Beijing Institute of Genomics, CAS	China
Fang, Rongxin	Beijing Institute of Genomics, CAS	China
Fares, Ribel	Texas State University - San Marcos	USA
Faruqi, A Ali	Imperial College, London	UK
Fei, Li	Nanjing Agricultural University	China
Feng, Bo	Illinois Institute of Technology	USA
Feng, Huijuan	Tsinghua University	China
Feng, Shengzhong	Shenzhen Institutes of Advanced Technology	China
Feng, Zhixing	Tsinghua University	China
Figueroa, Nathan	Miami Unviersity	USA
Firth, E. Andrew	University of Cambridge	UK
Fox, Gearoid	University College Dublin	Ireland
Freindman, Yitzhak	Hebrew University of Jerusalem	Israel
Fricke, Inka	Ruhr University Bochum	Germany
Fristensky, Brian	University of Manitoba	Canada
Fu, Limin	University of California San Diego	USA
Fu, Yan	Chinese Academy of Sciences	China
Fujiwara, Ricardo	Federal University of Minas Gerais	Brazil
Furmanek, Tomasz	Institute of Marine Research	Norway
Gao, Hong	Institute of Microbiology, CAS	China
Gao, J. Byron	Texas State University	USA
Gao, Jean	University of Texas at Arlington	USA
Gao, Juntao	Tsinghua University	China
Gao, Lin	XiDian University	China
Gao, Shaorong	National Institute of Biological Sciences	China
Gao, Xiaoning	Chinese PLA General Hospital	China
Ge, Qinyu	Southeast University	China
Gerber, P. Andre	University of Surrey	UK
Gerwert, Klaus	Ruhr Universität Bochum	Germany
Gholami, Moghaddas Amin	Technical University of Munich	Germany
Ghoraie, Soltan Laleh	University of Waterloo	Canada
Giemza, Joanna	University of Warsaw	Poland
Gifford, David	CSAIL - MIT	USA
Gilliam, Conrad	University of Chicago	USA
Gordan, Raluca	Duke University	USA
Gould, Michael	University of Wisconsin Madison	USA
Großerüschkamp, Frederik	Ruhr Universität Bochum	Germany
Gu, Jin	Tsinghua University	China
Gu, Wanjun	Southeast University	China
Guénoche, Alain	IML, CNRS	France
Guimerà, Roger	Universitat Rovira i Virgili	Spain
Guo, An-Yuan	Huazhong University of Science and Technology	China
Guo, Fei	City University of Hong Kong	Hong Kong
Guo, Xingli	XiDian University	China
Guo, Xuejiang	Nanjing medical university	China
Guo, Yi-Ke	Imperial College London	UK
Gurevich, Alexey	Saint Petersburg Academic University	Russia
Haag, Jill	University of Wisconsin Madison	USA
Haar, der von Tobias	University of Kent	UK
Hahne, Hannes	Technical University of Munich	Germany
Haining, Nicholas	Dana Farber Cancer Institute	USA
Halperin, Eran	Tel Aviv University	Israel
Han, Ping	Nanjing Medical University	China
Han, Xiaoxu	Fordham University New York NY USA	USA
Hao, Jingjing	Tsinghua University	China
Hao, Lingtong	Beijing Institute of Genomics, CAS	China
Hao, Pei	Shanghai Institutes for Biological Sciences, CAS	China
Harris, Curtis	National Cancer Institute	USA
Have, L. Cherry	Ontario Cancer Institute	Canada
Hayes, Matthew	Case Western Reserve University	USA
He, Bing	University of Iowa	USA
He, Chao	Tsinghua University	China
He, Dan	IBM T.J. Watson	USA
He, Peng	Tsinghua University	China
He, Shan	University of Birmingham	UK
He, Zengyou	Dalian University of Technology	China
Heath, K John	University of Birmingham	UK
Hengel, Shawna	PNNL	USA

Hernandez, Troy	University of Illinois at Chicago	USA
Hoeng, J.	Philip Morris Research and Development	Switzerland
Holstege, Frank	UMC Utrecht	Netherlands
Hong, Ming-Huang	Sun Yat-sen University	China
Honig, Barry	Columbia University	USA
Hooff, van Sander	UMC Utrecht	Netherlands
Hormozdiari, Farhad	University of California, Los Angeles	USA
Hou, Lixia	Beijing Institute of Genomics, CAS	China
Hou, Minmei	Northern Illinois University	USA
Hsu, Yi-Chiung	Institute of Statistical Science, Academia Sinica	Taiwan
Hu, Jialu	Freie Universität Berlin	Germany
Hu, Long	Tsinghua University	China
Hu, Wenqian	Whitehead Institute	USA
Hu, Xiaohui	South China Normal University	China
Hu, Zhibin	Nanjing medical university	China
Hu, Zhirui	Tsinghua University	China
Huang, Da Hsien	National Chiao Tung University	Taiwan
Huang, Fuzhuo	Boston University	USA
Huang, Haiyan	University of California, Berkeley	USA
Huang, Shi	Central South University	China
Huang, Ting	Dalian University of Technology	China
Huang, Yu	MCB, USC, Graduate student	USA
Huang, Yuanhua	Tsinghua University	China
Hung, Hung Jui	National Chiao Tung University	Taiwan
Hwang, Ming-Jing	Institute of Biomedical Scienecs, Academia Sinica	Taiwan
Hyrien, Olivier	Ecole Normale Supérieure de Paris	France
Ilicic, Tomislav	Wellcome Trust Sanger Institute	UK
Ingolia, T. Nicholas	Carnegie Institution for Science	USA
Ishkanian, Adrian	Ontario Cancer Institute	Canada
Jacquemin, Thibault	Tsinghua University	China
Jahangiri, Soran	Concordia Univerisuty	Canada
Jang, Ho	GIST	Korea
Jang, Sung Jin	Mayo Clinic	USA
Jansson, Jesper	Kyoto University	Japan
Jen, Jin	Mayo Clinic	USA
Jeong, Kyowon	University of California, San Diego	USA
Jiang, Haitao	Shandong University	China
Jiang, Qingshan	Shenzhen Institutes of Advanced Technology	China
Jiang, Rui	Tsinghua University	China
Jiang, Tao	University of California, Riverside	USA
Jiang, Yun	Tsinghua University	China
Jing, Yan	Nanjing University of Aeronautics and Astronautics	China
Jurisica, Igor	Ontario Cancer Institute	Canada
Kacprowski, Tim	University Medicine Greifswald	Germany
Kaddi, Chanchala	Georgia Institute of Technology	USA
Kaderali, Lars	University of Technology Dresden	Germany
Kai, Wang	Nanjing Agricultural University	China
Kakaradov, Boyko	University of California, San Diego	USA
Kallenbach, Angela	Ruhr Universität Bochum	Germany
Kambadur, Prabhanjan	IBM TJ Watson Research Center	USA
Kang, Mingon	University of Texas at Arlington	USA
Kanitz, Alexander	University Basel	Switzerland
Karro, John	Miami University, Computer Science, Microbiology	USA
Kartasalo, Kimmo	Tampere University of Technology	Finland
Kashef-Haghighi, Dorna	Stanford Univeristy	USA
Kehr, Birte	Freie Universität Berlin	Germany
Kellis, Manolis	Massachusetts Institute of Technology	USA
Kemmeren, Patrick	UMC Utrecht	Netherlands
Kendziorski, Christina	University of Wisconsin Madison	USA
Khavari, David	Stanford University	USA
Kim, Dongchul	University of Texas at Arlington	USA
Kim, Hanjoo	Seoul National University	Korea
Kim, Ki-Bong	Sangmyung University	Korea
Kim, Sangtae	UCSD	USA
Knowles, James	University of Southern California	USA

Ko, Cheu	UMC Utrecht	Netherlands
Koerkamp, Groot Marian	UMC Utrecht	Netherlands
Kondrashov, Alexey	Moscow State University	Russia
Korobeynikov, Anton	St. Petersburg State University	Russia
Kostem, Emrah	UCLA Computer Science	USA
Kozakov, Dima	Boston University	USA
Krasheninnikova, Ksenia	Saint Petersburg Academic University	Russia
Krauß, Sascha	Ruhr University Bochum	Germany
Kringelum, Vindahl Jens	Technical University of Denmark	Denmark
Kuang, Shuzhen	Huazhong University of Science and Technology	China
Kuang, Yuming	Peking University	China
Kumar, Senthil	University of Maryland	USA
Kung, Hsiang Wei	National Chiao Tung University	Taiwan
Kuo, Jay C.-C.	University of Southern California	USA
Kuperstein, Inna	Institut Curie	France
Kuster, Bernhard	Technical University of Munich	Germany
Kwast, der van Theo	Ontario Cancer Institute	Canada
Lai, En-Yu	Institute of Information Science, Academia Sinica	Taiwan
Lalonde, Emilie	Ontario Cancer Institute	Canada
Lam, L. Wan	British Columbia Cancer Research Centre	Canada
Lapidus, Alla	Saint Petersburg Academic University	Russia
Lasken, Roger	J. Craig Venter Institute	USA
Latonen, Leena	University of Tampere	Finland
Lazzer, de Guillaume	Research and Development Solutions Private Limited	Singapore
Le, Rongrong	National Institute of Biological Sciences	China
Lederman, Roy	Yale University	USA
Lee, Byunghan	Seoul National University	Korea
Lee, Guipeng	Tsinghua University	China
Lee, Hyunju	Gwangju Institute of Science and Technology	Korea
Lee, Seunghak	Carnegie Mellon University	USA
Leenen, van Dik	UMC Utrecht	Netherlands
Lei, Chengwei	University of Texas at San Antonio	USA
Lei, Hongxing	Beijing Institute of Genomics, CAS	China
Leiserson, Mark	Brown University	USA
Leng, Ning	University of Wisconsin Madison	USA
Lenstra, Tineke	UMC Utrecht	Netherlands
Leong, Wai Hon	National University of Singapore	Singapore
Leppänen, Simo-Pekka	Tampere University of Technology	Finland
Levens, David	National Cancer Institute	USA
Lewitter, Fran	Whithead Institute	USA
Li, Ao	University of Science and Technology of China	China
Li, Cheng Shuai	City University of Hong Kong	Hong Kong
Li, Chunyan	Institut Pasteur of Shanghai, CAS	China
Li, Guoliang	Genome Institute of Singapore	Singapore
Li, Hua	Institute of Computing Technology, CAS	China
Li, Jessica Jingyi	University of California, Berkeley	USA
Li, Jing	Case Western Reserve University	USA
Li, Jingrui	China Agricultural University	China
Li, Jun Mulin	University of Hong Kong	Hong Kong
Li, Junji	Southeast University	China
Li, Kenli	Hunan University	China
Li, Ker-Chau	Institute of Statistical Science, Academia Sinica	Taiwan
Li, Lei	Academy of Mathematics and Systems Science, CAS	China
Li, Lianshuo	Tsinghua University	China
Li, Ming	Institute of Computing Technology, CAS	China
Li, Musheng	Southeast University	China
Li, Pengping	nanjing university	China
Li, Pu	Institute of Biophysics, CAS	China
Li, Rong	Stowers Institute for Medical Research	USA
Li, Tianyang	Tsinghua University	China
Li, Tiecheng	South China Normal University	China
Li, Tingting	Peking University	China
Li, Wei	Baylor College of Medicine	USA
Li, Weizhong	University of California San Diego	USA
Li, Xiaoli	Institute for Infocomm Research	Singapore
Li, Xinna	Beijing Institute of Genomics, CAS	China

Li, Xuan	Shanghai Institutes for Biological Sciences	China
Li, Yanda	Tsinghua University	China
Li, Yanjian	Tsinghua University	China
Li, Yi	University of California, Irvine	USA
Li, Yinqing	Massachusetts Institute of Technology	USA
Li, Yixue	Shanghai Center for Bioinformatics Information Technology	China
Li, Yonghui	Chinese PLA General Hospital	China
Li, You	Hong Kong Baptist University	Hong Kong
Li, Yu-Cheng	Institute of Statistical Science, Academia Sinica	Taiwan
Li, Zhe	Tsinghua University	China
Li, Zhihua	Chinese Academy of Medical Sciences	China
Liang, Zheng-Yu	Tsinghua University	China
Lihua, Cao	Beijing Institute of Genomics, CAS	China
Lijnzaad, Philip	UMC Utrecht	Netherlands
Lili, Dong	Beijing Institute of Genomics, CAS	China
Ling, Shaoping	Beijing Institute of Genomics, CAS	China
Linghu, Bolan	Novartis Institutes for BioMedical Research	USA
Linial, Michal	Hebrew University of Jerusalem	Israel
Linial, Nathan	Hebrew University of Jerusalem	Israel
Liu, Chenglin	Methodist Hospital Research Institute	USA
Liu, Fei	Beijing Institute of Genomics, CAS	China
Liu, Guiming	Beijing Institute of Genomics, CAS	China
Liu, Guojing	Beijing Institute of Genomics, CAS	China
Liu, Honglei	Tsinghua University	China
Liu, Jiahan	Institute of Software, Chinese Academy of Sciences	China
Liu, Li	Southeast University	China
Liu, Lin	Xiamen University	China
Liu, Quentin	Sun Yat-sen University	China
Liu, Xiaoshuang	nanjing university	China
Liu, Xiaowen	Indiana University-Purdue University Indianapolis	USA
Liu, Yifang	Tsinghua University	China
Liu, Yihua	MD Anderson Cancer Research Center	USA
Liu, Yongxuan	XiDian University	China
Liu, Yu	Tsinghua University	China
Liu, Yuanning	University of Science and Technology of China	China
Liu, Yunpeng	University of Birmingham	UK
Liu, Zhandong	Baylor College of Medicin	USA
Lo, Christine	University of California, San Diego	USA
Lobo, Francisco	Federal University of Minas Gerais	Brazil
Lodish, Harvey	Whitehead Institute	USA
Logacheva, Maria	Moscow State University	Russia
Loh, Po-Ru	Massachusetts Institute of Technology	USA
Lohmann, Volker	Universität Heidelberg	Germany
Lokshtanov, Daniel	University of California, San Diego	USA
Lougheed, David	Loyalist Collegiate and Vocational Institute	Canada
Lozano, Aurelie	IBM TJ Watson Research Center	USA
Lu, Xuemei	Beijing Institute of Genomics, CAS	China
Lu, Youtao	CAS-MPG Partner Institute, CAS	China
Lu, Yu	Peking University	China
Lu, Zhi	Tsinghua University	China
Lu, Zuhong	Southeast University	China
Luiz-Rodrigues, Gabriela	Federal University of Minas Gerais	Brazil
Lund, Ole	Technical University of Denmark	Denmark
Luo, Youxi	Shenzhen Institutes of Advanced Technology	China
Lv, Pengyun	Tsinghua University	China
M?hl, Mathias	Albert-Ludwigs-University Freiburg	Germany
Ma, Shining	Tsinghua University	China
Ma, Wenji	City University of Hong Kong	Hong Kong
Ma, Xinyun	Tsinghua University	China
Ma, Ying-Ke	Institute of Genetics and Developmental Biology, CAS	China
Ma, Yue	Institute of Biophysics, CAS	China
Ma, Zhaowu	Huazhong University of Science and Technology	China
Mahmoody, Ahmad	Brown University	USA

Mai, Guoqin	Shenzhen Institutes of Advanced Technology, CAS	China
Mak, Denise	Ontario Cancer Institute	Canada
Malde, Ketil	Institute of Marine Research	Norway
Malloff, A. Chad	British Columbia Cancer Research Centre	Canada
Maltsev, Natalia	University of Chicago	USA
Mancuso, Nicholas	Georgia State University	USA
Mangul, Serghei	UCLA	USA
Margaritis, Thanasis	UMC Utrecht	Netherlands
Markowetz, Florian	Cambridge Research Institute	UK
Martin, Florian	Philip Morris International	Switzerland
Mathis, C.	Philip Morris Research and Development	Switzerland
Mayani, Rajiv	Information Science Institute	USA
Maycock, Matthew	Transition Technologies S.A.	Poland
McLean, Jeffrey	J. Craig Venter Institute	USA
Mendes, Tiago	Federal University of Minas Gerais	Brazil
Meng, Alice	Ontario Cancer Institute	Canada
Meng, Cehn	Technical University of Munich	Germany
Meng, Jintao	Shenzhen Institutes of Advanced Technology, CAS	China
Mesirov, Jill	Broad Institute	USA
Messer, Karen	Moores UCSD Cancer Center	USA
Meyer, P.	IBM Thomas J. Watson Research Center	USA
Mi, Kaixia	Institute of Microbiology, CAS	China
Mi, Shuangli	Beijing Institute of Genomics, CAS	China
Miao, Gengxin	University of California, Santa Barbara	USA
Miao, Xuexia	Beijing Institute of Genomics, CAS	China
Michel, M. Audrey	University College Cork	Ireland
Miladi, Milad	Albert-Ludwigs-University Freiburg	Germany
Miles, Tony	UMC Utrecht	Netherlands
Milosevic, Michael	Ontario Cancer Institute	Canada
Moffa, Giusi	University of Regensburg	Germany
Moghadasi, Mohammad	Boston University	USA
Moon, Nathalie	Ontario Cancer Institute	Canada
Mosig, Axel	Ruhr Universität Bochum	Germany
Moult, John	University of Maryland	USA
Movaghar, Fayyaz Afshin	Mazandaran	Iran
Munteanu, Alina	University of Iasi	Romania
Nagarajan, Niranjan	Genome Institute of Singapore	Singapore
Nairz, Knud	ETH Zurich	Switzerland
Nakaya, Helder	Emory University Vaccine Center	USA
Nardini, Christine	CAS-MPG Partner Institute, CAS	China
Ness, Robert	Purdue University	USA
Nesteruk, Monika	Warsaw Medical Center for Postgraduate Education	Poland
Newburger,Daniel	Stanford University	USA
Ng, See-Kiong	Institute for Infocomm Research	Singapore
Nguyen, Ninh Nam	National University of Singapore	Singapore
Nguyen, Vu Phi	National University of Singapore	Singapore
Nikolayeva, Olga	University of Zurich	Switzerland
Ning, Kang	Qingdao Institute of Bioenergy and Bioprocess Technology, CAS	China
Norel, R.	IBM Thomas J. Watson Research Center	USA
Nurk, Sergey	Saint Petersburg Academic University	Russia
Nykter, Matti	University of Tampere	Finland
O'Duibhir, Eoghan	UMC Utrecht	Netherlands
Oesper, Layla	Brown University	USA
Oles, Karl	Mayo Clinic	USA
P.K., Suresh	VIT University	India
Panda, Roshni	VIT University	India
Pandis, Ioannis	Imperial College London	UK
Pandit, Manish	University of Windsor	Canada
Pang, Yu	Chinese Center for Disease Control and Prevention	China
Papoutsaki, Alexandra	Department of Computer Science	USA
Parida, Laxmi	IBM	USA
Park, Kiejung	Korean Bioinformation Center	Korea
Participants, Cagi	Multiple Institutions	USA
Pasa-Tolic, Ljiljana	PNNL	USA

Pasch, de van Loes	UMC Utrecht	Netherlands
Paschalidis, Ioannis	Boston University	USA
Pei, Jian	Simon Fraser University	Canada
Peitsch, C. M.	Philip Morris Research and Development	Switzerland
Penin, Aleksey	Moscow State University	Russia
Perun, Serhiy	Institute of Physics PAS	Poland
Peslherbe, Gilles	Concordia Univerisuty	Canada
Petersen, Dennis	Ruhr University Bochum	Germany
Petrey, Donald	Columbia University	USA
Pevzner, Pavel	UCSD	USA
Pilpel, Yitzhak	Weizmann Institute of Science	Israel
Pinney, John	Imperial College London	UK
Pinney, W John	Imperial College, London	UK
Pintilie, Melania	Ontario Cancer Institute	Canada
Podsiad?o, Agnieszka	University of Warsaw	Poland
Ponty, Yann	Ecole Polytechnique	France
Poussin, C.	Philip Morris Research and Development	Switzerland
Pra, Dal Alan	Ontario Cancer Institute	Canada
Prjibelsky, Andrey	Saint Petersburg Academic University	Russia
Przytycka, Teresa	NIH	USA
Pu, Minya	Moores Cancer Center	USA
Pulendra, Bali	Emory University Vaccine Center	USA
Puniyani, Kriti	Carnegie Mellon University	USA
Pyshkin, Alexey	Saint Petersburg Academic University	Russia
Qi, Yijun	Tsinghua University	China
Qian, Mingping	Peking University	China
Qian, Minping	Peking University	China
Qin, Zhiyi	Tsinghua University	China
Qiu, Peng	University of Texas MD Anderson Cancer Center	USA
R?tsch, Gunnar	Friedrich Miescher Laboratory	USA
Ramnarine, Rohan Varune	Ontario Cancer Institute	Canada
Raphael, Ben	Brown University	USA
Raphael, Benjamin	Brown University	USA
Rappoprt, Nadav	Hebrew University of Jerusalem	Israel
Ravi, R.	Carnegie Mellon University	USA
Reeder, Christopher	CSAIL - MIT	USA
Reinert, Knut	FU Berlin	Germany
Reinharz, Vladimir	McGill University	Canada
Ren, Jie	Peking University	China
Repo, Susanna	EMBL-EBI	UK
Riby, Jacques	University of California, Berkeley	USA
Rice, J.J.	IBM Thomas J. Watson Research Center	USA
Rissman, Anna	University of Wisconsin Madison	USA
Robinson, D. Mark	University of Zurich	Switzerland
Robisson, Benoit	TAGC, Inserm	France
Robles, Ana	National Cancer Insitute	USA
Rocha, Wanderson	Federal University of Parana	Brazil
Rodrigues, Thiago	Centro Federal de Educação Tecnológica de Minas Gerais	Brazil
Rodriguez, Jesse	Stanford University	USA
Rogé, Xavier	Tsinghua University	China
Ronen, Roy	University of California, San Diego	USA
Rueda, Luis	University of Windsor	Canada
Ruotti, Victor	Morgridge Institute for Research	USA
Rydzak, Thomas	University of Manitoba	Canada
Sadeh, Javad Mohammad	University of Regensburg	Germany
Safonova, Yana	St. Petersburg Academic University	Russia
Salah, Ahmad	Hunan University	China
Salari, Raheleh	Stanford Univeristy	USA
Saleh, Shayon Syed	Stanford Univeristy	USA
Sameith, Katrin	UMC Utrecht	Netherlands
Sankoff, David	University of Ottawa	Canada
Saveliev, Vladislav	Saint Petersburg Academic University	Russia
Scaravilli, Mauro	University of Tampere	Finland
Schmiedl, Christina	Albert-Ludwigs-University Freiburg	Germany
Schwartz, Russell	Carnegie Mellon University	USA
Sendorek, Dorota	Ontario Cancer Institute	Canada
Senter, Evan	Boston College	USA
Sergushichev, Alexey	NRU ITMO	Russia
Serocka, Peter	CAS-MPG Partner Institute	China
Sha, Jiahao	Nanjing medical university	China
Sham, Chung Pak	University of Hong Kong	Hong Kong

Shaoping, Ling	Beijing Institute of Genomics, CAS	China
Sheikh, Saad	University of Florida	USA
Shen, Chuanqi	Stanford University	USA
Shen, Qi	Fudan University	China
Sheng, Jia	Shanghai Center for Bioinformation Technology, CAS	China
Shi, Jiahai	Whitehead Institute	USA
Sidow, Arend	Stanford University	USA
Sirotkin, Alexander	Saint Petersburg Academic University	Russia
Sirotkin, Yakov	Saint Petersburg Academic University	Russia
Skibola, Chris	University of California, Berkeley	USA
Sleumer, Monica	Tsinghua University	China
Smits, Bart	University of Wisconsin Madison	USA
Song, Baoxing	Qingdao Institute of Bioenergy and Bioprocess Technology, CAS	China
Song, Gao	National University of Singapore	Singapore
Song, Xiaofeng	Nanjing University of Aeronautics and Astronautics	China
Song, Youqiang	University of Hong Kong	Hong Kong
Sonmez, Kemal	Oregon Health & Science University,	USA
Spang, Rainer	University of Regensburg	Germany
Sparling, Richard	University of Manitoba	Canada
Spiewak, Dan	Novartis Institutes for BioMedical Research	USA
Squire, A. Jeremy	Qu ’	Canada
Srihari, Sriganesh	University of Queensland	Australia
Srivastava, Aashish	Warsaw Institute of Oncology	Poland
Stępniak, Piotr	Transition Technologies S.A.	Poland
Stepanauskas, Ramunas	Bigelow Laboratory for Ocean Sciences	USA
Sternberg, Mike	Imperial College London	UK
Stevens, Michael	Washington University	USA
Stewart, Ron	Morgridge Institute for Research	USA
Stolovitzky, G.	IBM Thomas J. Watson Research Center	USA
Su, Xiaoquan	Qingdao Institute of Bioenergy and Bioprocess Technology, CAS	China
Su, Yao	Beijing Institute of Genomics, CAS	China
Sulaimanov, Nurgazy	Universität Heidelberg	Germany
Sulakhe, Dinanath	Computation Institute	USA
Sun, Beili	Southeast University	China
Sun, Deqiang	Baylor College of Medicine	USA
Sun, Fengzhu	University of Southern California	USA
Sun, Huan	University of California, Santa Barbara	USA
Sun, Ren	University of California, Los Angeles	USA
Sun, Xiao	Southeast University	China
Sung, Wing-Kin	National University of Singapore	Singapore
Św k , K ad	Warsaw University of Technology	Poland
Sykes, Jenna	Ontario Cancer Institute	Canada
Szustakowski, Joseph	Novartis Institutes for BioMedical Research	USA
Taboureau, Olivier	Technical University of Denmark	Denmark
Tamayo, Pablo	Broad Institute	USA
Tan, Kai	University of Iowa	USA
Tan, Xu	Harvard Medical School	USA
Tan, Yan	Boston University	USA
Tang, Biao	Fudan University	China
Taylor, Andrew	University of Chicago	USA
Teh, Whye Yee	University of Oxford	UK
Teixera, Santuza	University of Minas Gerais	Brazil
Teng, Li	University of Iowa	USA
Tennant, Daniel	University of Birmingham	UK
Tesler, Glenn	University of California, San Diego	USA
Thermes, Claude	CNRS	France
Thoms, John	Ontario Cancer Institute	Canada
Thomson, James	Morgridge Institute for Research	USA
Toh, Kim-Chuan	National University of Singapore	Singapore
Tolic, Nikola	PNNL	USA
Tran, Hieu Ngoc	National University of Singapore	Singapore
Tripathy, Chittaranjan	Duke University	USA
Tsai, Ming-Chi	Carnegie Mellon University	USA
Tsarev, Fedor	NRU ITMO	Russia
Tu, Jing	Southeast University	China
Udpa, Nitin	University of California, San Diego	USA

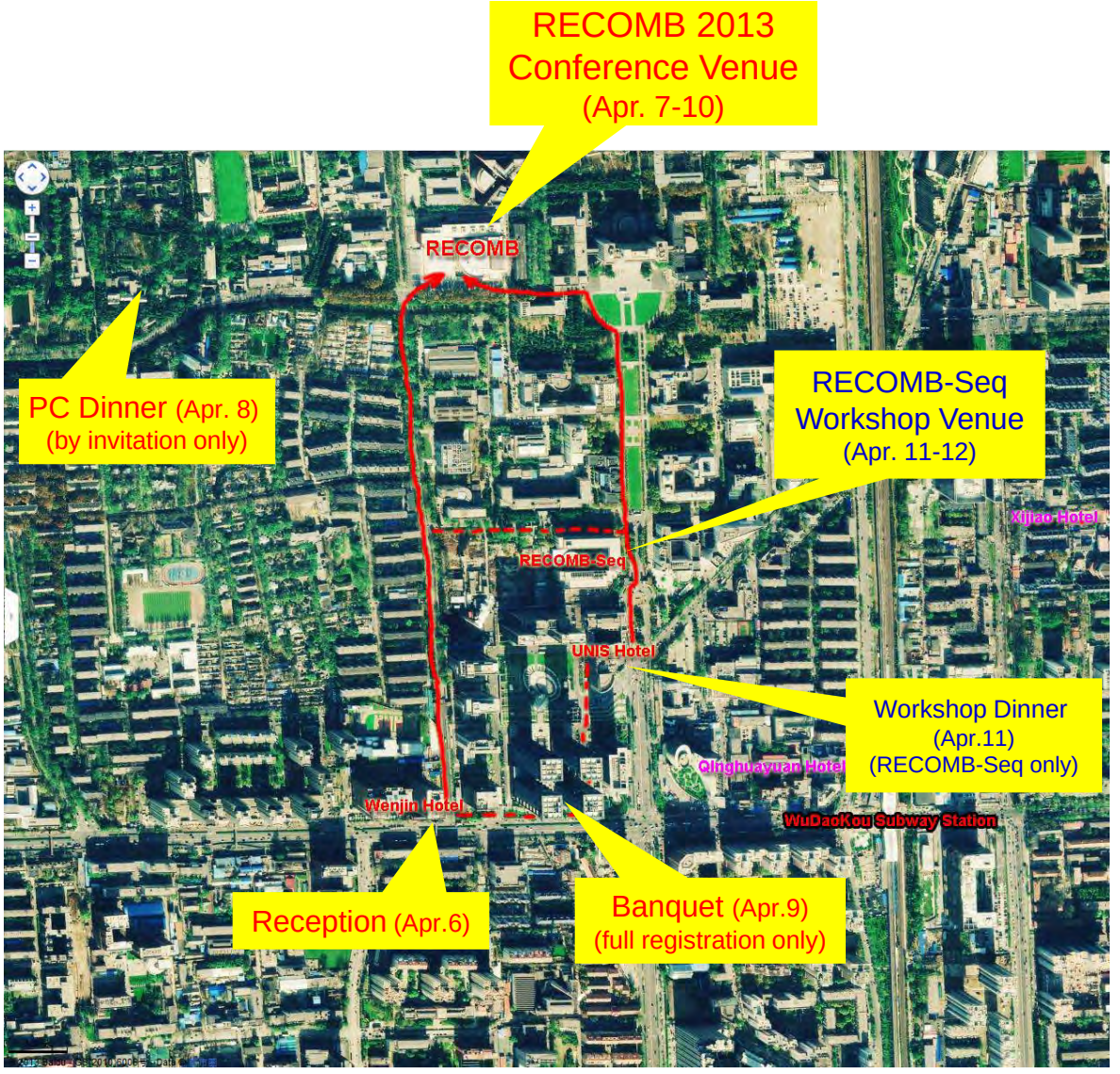
Upfal, Eli	Brown University	USA
Vajda, Sandor	Boston University	USA
Vakili, Pirooz	Boston University	USA
Valin, Benoit	Research and Development Solutions Private Limited	Singapore
Vandin, Fabio	Department of Computer Science	USA
Visakorpi, Tapio	University of Tampere	Finland
Vitek, Olga	Purdue University	USA
Vongsangnak, Wanwipa	Soochow University	China
Vyahhi, Nikolay	Saint Petersburg Academic University	Russia
Wageningen, van Sake	UMC Utrecht	Netherlands
Waldispühl, Jér?me	McGill University	Canada
Waltering, Kati	Tampere University of Technology	Finland
Wan, Xiangbo	Sun Yat-sen University	China
Wang, Anqi	Academy of Mathematics and Systems Science, CAS	China
Wang, Bingbo	XiDian University	China
Wang, Bingqiang	Beijing Institute of Genomics, CAS	China
Wang, Chenyang	Fudan University	China
Wang, D. May	Georgia Institute of Technology	USA
Wang, Dapeng	Beijing Institute of Genomics, CAS	China
Wang, Haitian Maggie	Chinese University of Hong Kong	Hong Kong
Wang, Hongyan	Methodist Hospital Research Institute	USA
Wang, Jin	nanjing university	China
Wang, Jing	Institute of Psychology, CAS	China
Wang, Junwen	University of Hong Kong	Hong Kong
Wang, Kai	Foundation Medicine	USA
Wang, Lei	Beijing Institute of Genomics, CAS	China
Wang, Lusheng	City University of Hong Kong	Hong Kong
Wang, Meng	Institute of Genetics and Developmental Biology, CAS	China
Wang, Minghui	University of Science and Technology of China	China
Wang, Mingxun	UCSD	USA
Wang, Panwen	University of Hong Kong	Hong Kong
Wang, Qi	Tsinghua University	China
Wang, Quan	Peking University	China
Wang, Shengqin	Southeast University	China
Wang, Shengrui	University of Sherbrooke	Canada
Wang, Shengyue	Chinese National Human Genome Center at Shanghai	China
Wang, Tianming	Dalian University of Technology	China
Wang, Ting	Washington University	USA
Wang, Weichen	Tsinghua University	China
Wang, Weixin	University of Hong Kong	Hong Kong
Wang, Xi	University of Newcastle, Australia	Australia
Wang, Xiaofei	Southeast University	China
Wang, Xiaowo	Tsinghua University	China
Wang, Xijun	Concordia Univerisuty	Canada
Wang, Xiu-Jie	Institute of Genetics and Developmental Biology, CAS	China
Wang, Xumin	Beijing Institute of Genomics, CAS	China
Wang, Ying	Xiamen University	China
Wang, Yunfeng	University of California, Irvine	USA
Wang, Zhanyong	UCLA	USA
Wei, Dan	Xiamen University	China
Wei, Yanjie	Shenzhen Institutes of Advanced Technology, CAS	China
Weiss, Ron	Massachusetts Institute of Technology	USA
West, B Robert	Stanford University	USA
Whelan Christopher W.	Oregon Health & Science University	USA
Whittaker, Joe	Lancaster University	UK
Wieben, Eric	Mayo Clinic	USA
Wilczynski, Bartek	University of Warsaw	Poland
Wilcz ń k , Bartek	Warsaw University	Poland
Wilhelm, Mathias	Technical University of Munich	Germany
Will, Sebastian	Albert-Ludwigs-University Freiburg	USA
Wojdan, Konrad	Warsaw University of Technology	Poland
Wojtowicz, Damian	University of Warsaw	Poland
Wu, Chung-I	Beijing Institute of Genomics, CAS	China
Wu, Chung-I	Beijing Institute of Genomics, CAS	China
Wu, Dingming	Tsinghua University	China
Wu, Hsin-Ta	Brown University	USA

Wu, Hua-Jun	Institute of Genetics and Developmental Biology, CAS	China
Wu, Jiaxin	Tsinghua University	China
Wu, Kun-Pin	National Yang Ming University	Taiwan
Wu, Ling-Yun	Academy of Mathematics and Systems Science, CAS	China
Wu, Nicholas	University of California, Los Angeles	USA
Wu, Si	PNNL	USA
Wu, Sitao	University of California San Diego	USA
Wu, Yang	Tsinghua University	China
Wu, Yufeng	University of Connecticut	USA
Wu, Yuliang	Huazhong University of Science and Technology	China
Wu, Zheng	Peking University	China
Wu, Zhixiang	Technical University of Munich	Germany
Wyrwicz, S. Lucjan	Warsaw Institute of Oncology	Poland
Xi, Ruibin	Peking University	China
Xia, Leihao	Hong Kong Baptist University	Hong Kong
Xia, Yan	Beijing Institute of Genomics, CAS	China
Xiang, Qian	Sun Yat-Sen University	China
Xiang, Y.	Philip Morris Research and Development	Switzerland
Xiang, Yang	Philip Morris International	Switzerland
Xiao, Fen	Xiangtan University	China
Xie, Bingqing	Illinois Institute of Technology	USA
Xie, Feng	Institute of Microbiology, CAS	China
Xie, Jianming	Southeast University	China
Xie, Mingchao	Washington University	USA
Xie, Peng	Tsinghua University	China
Xie, Xiaohui	University of California, Irvine	USA
Xie, Xueying	Southeast University	China
Xie, Zhen	Tsinghua University	China
Xie, Zhong-Ru	Institute of Biomedical Sciences, Academia Sinica	Taiwan
Xing, Eric	Carnegie Mellon University	USA
Xing, Jing	China University of Geosciences	China
Xing, Qingfeng	Peking University	China
Xiong, Jie	Tsinghua University	China
Xiong, Yun	Fudan University	China
Xu, Dong	Sanford-Burnham Medical Research Institute	USA
Xu, Feng	University of Hong Kong	Hong Kong
Xu, Jian	Qingdao Institute of Bioenergy and Bioprocess Technology, CAS	China
Xu, Jinfeng	National University of Singapore	Singapore
Xu, Kui	Tianjin Polytechnic University	China
Xue, Yun	South China Normal University	China
Xuemei, Lu	Beijing Institute of Genomics, CAS	China
Yan, Anthony	Duke University	USA
Yan, Jun	Huazhong University of Science and Technology	China
Yan, Weili	Fudan University	China
Yan, Xifeng	University of California, Santa Barbara	USA
Yang, Chen	Tsinghua University	China
Yang, Fan	Novartis Institutes for BioMedical Research	USA
Yang, Lianping	Northeastern University	China
Yang, Minjun	Chinese National Human Genome Center at Shanghai	China
Yang, Song	Lawrence Berkeley National Lab	USA
Yang, Xian	Imperial College London	UK
Yin, Junming	Carnegie Mellon University	USA
Yin, Lingyun	Tsinghua University	China
Yin, Longhui	Xiangtan University	China
Ying, Liu	Nanjing Agricultural University	China
Ylipää, Antti	Tampere University of Technology	Finland
Yoon, Sungroh	Seoul National University	Korea
Yu, Jun	Beijing Institute of Genomics, CAS	China
Yu, Li	Chinese PLA General Hospital	China
Yu, S. Philip	University of Illinois at Chicago, USA	USA
Yu, Yao	Shanghai Institutes for Biological Sciences	China
Yuan, Bingbing	Whitehead Institute	USA
Yuan, Shin-Sheng	Institute of Statistical Science, Academia Sinica	Taiwan
Yuan, Yinyin	Institute of Cancer Research	UK
Zafarana, Gaetano	Ontario Cancer Institute	Canada
Zaki, Nazar	UAE University	United Arab Emirates

Zelikovsky, Alex	Georgia State University	USA
Zeng, Tianying	Tsinghua University	China
Zhang, Chaoqun	Nanjing Agricultural University	China
Zhang, Cliff Qiangfeng	Stanford University	USA
Zhang, Hong	Fudan University	China
Zhang, Hongmei	Huazhong University of Science and Technology	China
Zhang, Jing	University of Southern California	USA
Zhang, Junyi	Southeast University	China
Zhang, Lixin	Institute of Microbiology, CAS	China
Zhang, Louxin	National University of Singapore	Singapore
Zhang, Michael	Tsinghua University	USA
Zhang, Peijin	Massachusetts Institute of Technology	USA
Zhang, Q Michael	Tsinghua University	China
Zhang, Shihua	Chinese Academy of Science	China
Zhang, Wei	The University of Texas	USA
Zhang, Xiangde	Northeastern University	China
Zhang, Xiangli	University of Manitoba	Canada
Zhang, Xiang-Sun	Academy of Mathematics and Systems Science, CAS	China
Zhang, Xuegong	Tsinghua University	China
Zhang, Yang	University of Michigan	USA
Zhang, Yubin	Beijing Institute of Genomics, CAS	China
Zhang, Zhao	Shenzhen Institutes of Advanced Technology	China
Zhang, Zhihua	Beijing Institute of Genomics, CAS	China
Zhang, Zhizhuo	National University of Singapore	Singapore
Zhang, Ziding	China Agriculturial University	China
Zhao, Fangqing	Beijing Institutes of Life Science, CAS	China
Zhao, Guoping	Chinese National Human Genome Center at Shanghai	China
Zhao, Hui	Beijing Institutes of Life Science, CAS	China
Zhao, Junfei	Academy of Mathematics and Systems Science, CAS	China
Zhao, Junsuo	Xiangtan University	China
Zhao, Kaiyong	Hong Kong Baptist University	Hong Kong
Zhao, Miaomiao	Shenzhen Institutes of Advanced Technology	China
Zhao, Yanlin	Chinese Center for Disease Control and Prevention	China
Zhao, Yi	Institute of Computing Technology, CAS	China
Zhao, Yue	Tsinghua University	China
Zhao, Zhen	Tsinghua University	China
Zheng, Chunfang	University of Ottawa	Canada
Zheng, Huajun	Chinese National Human Genome Center at Shanghai	China
Zheng, Jie	Nanyang Technological University	Singapore
Zheng, Yu	National University of Singapore	Singapore
Zhong, Qiaoyong	CAS-MPG Partner Institute	China
Zhong, Sheng	University of California, San Diego	USA
Zhou, Da	Tsinghua University	China
Zhou, Dongrui	Southeast University	China
Zhou, Fengfeng	Shenzhen Institutes of Advanced Technology, CAS	China
Zhou, Jasmine Xianghong	University of Southern Canlifornia	USA
Zhou, Mi	Huazhong University of Science and Technology	China
Zhou, Pei	Duke University	USA
Zhou, Shuigeng	Fudan University	China
Zhou, Tao	Nanjing medical university	China
Zhou, Tong	Tsinghua University	China
Zhou, Xiaobo	Weill Medical College of Cornell University	USA
Zhou, Xueya	Tsinghua University	China
Zhou, Yang	Chinese Center for Disease Control and Prevention	China
Zhou, Zuomin	Nanjing medical university	China
Zhu, Binhai	Montana State University	USA
Zhu, Daming	Shandong University	China
Zhu, Meifang	Xiamen University	China
Zhu, Mu	University of Waterloo	Canada
Zhu, Peijun	Dalian University of Technology	China
Zhu, Yangyong	Fudan University	China
Zhu, Yihua	Nanjing Agricultural University	China
Zhu, Yongqiang	Chinese National Human Genome Center at Shanghai	China
Zhuo, Ying	Institute of Microbiology, CAS	China
Zinovyev, Andrei	Institut Curie	France
Zou, Wei	Beijing Institute of Genomics, CAS	China

The RECOMB Chronology

#	Year	Dates and Location	Hosting Institution	Program Chair	Conference Chair
1	1997	Jan 20-23, Santa Fee, NM, USA	Sandia National Lab	Michael Waterman	Sorin Istrail
2	1998	Mar 22-25, New York, NY, USA	Mt. Sinai School of Medicine	Pavel Pevzner	Gary Benson
3	1999	Apr 22-25, Lyon, France	INRIA	Sorin Istrail	Mireille Regnier
4	2000	Apr 8-11, Tykyo, Japan	University of Tokyo	Ron Shamir	Satoru Miyano
5	2001	Apr 22-25, Montreal, Canada	Université de Montreal	Thomas Lengauer	David Sankoff
6	2002	Apr 18-21, Washington, DC, USA	Celera	Gene Myers	Sridhar Hannehalli
7	2003	Apr 10-13, Berlin, Germany	German Federal Ministry for Education and Research	Webb Miller	Martin Vingron
8	2004	Mar 14-18, San Diego, CA, USA	University of California San Diego	Dan Gusfield	Phillip E. Bourne
9	2005	May 14-18, Boston, MA, USA	Broad Institute of MIT and Harvard	Satoru Miyano	Jill P. Mesirow and Simon Kasif
10	2006	Apr 2-5, Venice, Italy	University of Padova	Alberto Apostolico	Concettina Guerra
11	2007	Apr 21-25, San Francisco, CA, USA	QB3	Terry Speed	Sandrine Dudoit
12	2008	Mar 30- Apr 2, Singapore	National University of Singapore	Martin Vingron	Limsoon Wong
13	2009	May 18-21, Tucson, AZ, USA	University of Arizona	Serafim Batzoglou	John Kececioglu
14	2010	Aug 12-15, Lisbon, Portugal	INESC-ID and Instituto Superior Ténico	Bonnie Berger	Arlindo Oliveira
15	2011	Mar 28-31, Vancouver, Canada	Lab for Computational Biology, Fraser University	Vineet Bafna	S. Cenk Sahinalp
16	2012	Apr 21-24, Barcelona, Spain	Centre for Genomic Regulation (CRG)	Benny Chor	Roderic Guigó
17	2013	Apr 7-10, Beijing, China	Tsinghua University	Fengzhu Sun	Xuegong Zhang



Sponsors:

清華大學

TNList

国家自然科学基金委员会 NSFC

ISCB INTERNATIONAL SOCIETY FOR COMPUTATIONAL BIOLOGY
AFFILIATED CONFERENCE

NSF National Science Foundation
WHERE DISCOVERIES BEGIN

EMC® ISILON

中科曙光 Sugon

intel

illumina®

BioMed Central
The Open Access Publisher

MONSANTO
imagine®