APBC 2016

The Fourteenth Asia Pacific Bioinformatics Conference

San Francisco, United States, January 11th-13th 2016

http://www.sfasa.org/apbc2016/apbc2016.html

Welcome

The Asia Pacific Bioinformatics Conference (APBC) is a leading conference in the Bioinformatics community and has grown rapidly since its inception in 2003. The goal of the annual conference series is to enable high quality interaction on bioinformatics research.

The past APBC conferences were held in:

1. Adelaide, Australia. Feb 4-7, 2003.
2. Dunedin, New Zealand. Jan 18-22, 2004.
3. Singapore. Jan 17-21, 2005.
4. Taipei, Taiwan. Feb 13-16, 2006.
5. Hong Kong. Jan 15-17, 2007.
6. Kyoto, Japan. Jan 14-17, 2008.
7. Beijing, China. Jan 13-16, 2009.
8. Bangalore, India. Jan 18-21, 2010.
9. Incheon, Korea. Jan 11-14, 2011.
10. Melbourne, Australia. Jan 17-19, 2012.
11. Vancouver, BC, Canada. Jan 21-23, 2013.
12. Shanghai, China. Jan 17-19, 2014.
13. HsinChu, Taiwan. Jan 21-23, 2015.

The Fourteenth Asia-Pacific Bioinformatics Conference is held in San Francisco, USA. It is the first time APBC being held in USA. The aim of this conference is to bring together researchers, academics, and industrial practitioners. APBC2016 invites high quality original full papers on any topic related to Bioinformatics and Computational Biology. The submitted papers must have not been published or under the consideration for publication in any other journal or conference with formal proceedings. All accepted papers will have to be presented by one of the authors at the conference. Accepted papers will be invited to be published in IEEE/ACM TCCB, BMC Bioinformatics, BMC Genomics, BMC Systems Biology, and Computational Biology and Chemistry.

Thanks very much to the Organization Chairs: General Co-Chairs Jing Huang (Veracyte, USA), Ying Lu (Stanford University, USA) and Philip Bourne (NIH/OD, USA); PC Co-Chairs Lu Tian (Stanford University, USA), Jijun Tang (University of South Carolina, USA), Phoebe Chen (La Trobe University, Australia) and Local Organization Chairs Jing Huang (Veracyte, USA), Hua Tang (Stanford University, USA) and Haiyan Huang (University of California, Berkeley, USA). We would also like to thank all local Organizing Committee members behind the scenes, especially, Ruixiao Lu, Bin Chen; Sammi Alsammi; Xin Zhao, Peng Yang, and Zhanzhi Hu for their hard work on the website, local arrangement and many other miscellaneous tasks. We would also like to thank the program committee members and steering committee members for their efficient work on review process and providing useful and detailed feedback to the authors. Thanks to all the authors for their high quality submissions.

Yours Sincerely,

Phoebe Chen
APBC Steering Committee Chair
Jan 2016

Committees

Steering Committee:

Phoebe Chen (Chair), Ph.D. (La Trobe University, Australia)
Sang Yup Lee, (KAIST, Korea)
Satoru Miyano, (University of Tokyo, Japan)
Mark Ragan, (University of Queensland, Australia)
Limsoon Wong, (National University of Singapore, Singapore)
Michael Q. Zhang, (CHSL, USA and Tsinghua University, China)

General Co-Chairs:

Jing Huang,Ph.D. (Veracyte, USA)
Ying Lu,Ph.D. (Stanford University, USA)
Philip Bourne, Ph.D. (NIH/OD, USA)

Local Organization Committee Co-Chairs:

Jing Huang, Ph.D. (Veracyte, USA)
Hua Tang, Ph.D. (Stanford University, USA)
Haiyan Huang, Ph.D. (University of California, Berkeley, USA)

Local Organization Committee:

Bin Chen, Ph.D. (University of California, San Francisco, USA)
Ruixiao Lu, Ph.D. (Genomic Health, USA)
Ying Lu, Ph.D. (Stanford University, USA)
Rui (Sammi) Tang, Ph.D. (Amgen, USA)
Lu Tian, Ph.D. (Stanford University, USA)
Peng Yang, M.S. (Clindata Insight Inc., USA)
Xin Zhao, MD, Ph.D. (Johnson & Johnson, USA)

Program Committee Co-Chairs:

Lu Tian, Ph.D. Stanford University, USA)
Jijun Tang, Ph.D. (University of South Carolina, USA)
Phoebe Chen, Ph.D. (La Trobe University, Australia)

Program Committee Members

Tatsuya Akutsu, (Kyoto University, Japan)
Max Alekseyev, (George Washington University, USA)
Gil Alterovitz, (Haravrd/MIT, USA)
Daniel Brown, (University of Waterloo, Canada)
Vladimir Brusic, (Nazarbayev University, Kazakhstan)
Dongbo Bu, (Chinese Academy of Sciences,China)
Janusz Bujnicki, (International Institute of Molecular and Cell Biology,Poland)
Pritam Chanda, (Johns Hopkins University, USA)
Tzu-Hao Chang, (Taipei Medical University, Taiwan)
Kun-Mao Chao, (National Taiwan University, Taiwan)
Yi-Ping Phoebe Chen, (La Trobe University, Australia)
Jake Yue Chen, (Indiana University School of Informatics,USA)
Chao Cheng, (Dartmouth College, USA)
Lenore Cowen, (Tufts University, USA)
Minghua Deng, (Peking University, China)
Nadia El-Mabrouk, (University of Montreal, Canada)
Liliana Florea, (Johns Hopkins University,USA)
Kyungsook Han, (Inha University, Korea)
Eric Ho, (Lafayette College, USA)
Jun Huan, (University of Kansas, USA)
Hsien-Da Huang, (National Chiao Tung University, Taiwan)
Heng Huang,(Univ. of Texas at Arlington, USA)
Hsuan-Cheng Huang, (National Yang-Ming University, Taiwan)
Haiyan Huang, (University of California, Berkeley, USA)
Hua Tang, Ph.D. (Stanford University, USA)
Jenn-Kang Hwang, (National Chiao Tung University, Taiwan)
Rui Jiang, (Tsinghua University, China)
Sun Kim, (Seoul National University,South Korea)
Mehmet Koyuturk, (Case Western Reserve University, USA)
Tzong-Yi Lee, (Yuan Ze University,Taiwan)
Doheon Lee, (KAIST, Korea)
Hongxing Lei, (Beijing Institute of Genomics, China)
Juan Liu, (Wuhan University, China)
Xinghua Lu, (University of Pittsbutgh, USA)
Bin Ma, (University of Waterloo, Canada)
Hiroshi Mamitsuka, (Kyoto University, Japan)
Ion Mandoiu, (University of Connecticut, USA)
Satoru Miyano, (University of Tokyo, Japan)
Bernard Moret, (EPFL, Switzerland)
Shinichi Morishita, (University of Tokyo, Japan)
Kenta Nakai, (University of Tokyo, Japan)
Saket Navlakha, (Salk Institute for Biological Studies,USA)
Arlindo Oliveira, (IST/INESC-ID and Cadence Research Laboratories, USA)
Rob Patro, (Stony Brook University, USA)
Sourav S Bhowmick, (Nanyang Technological University, Singapore)
Cenk Sahinalp, (Simon Fraser University, Canada)
Yasubumi Sakakibara, (Keio University, Japan)
David Sankoff, (University of Ottawa, Canada)
Thomas Schlitt, (King's College London, UK)
Russell Schwartz, (Carnegie Mellon University, USA)
Rahul Singh, (San Francisco State University, USA)
Steven Skiena, (Stony Brook University, USA)
Yanni Sun, (Michigan State University, USA)
Fengzhu Sun, (University of Southern California, USA)
Wing-Kin Sung, (National University of Singapore, Singapore)
Krister Swenson, (Université de Montpellier, France)
Petrus Tang, (Chang Gung University,Taiwan)
Jijun Tang, (University of South Carolina, USA)
Lu Tian, (Stanford University, USA)
Stephen Tsui, (The Chinese University of Hong Kong, Hong Kong)
Alfonso Valencia, (CNIO,S pain)
Jerome Waldispuhl, (McGill University, Canada)
Li-san Wang, (University of Pennsylvania,U SA)
Lusheng Wang, (City University of Hong Kong, Hong Kong)
Hsei-Wei Wang, (National Yang-Ming University, Taiwan)
Ka-Chun Wong, (City University of Hong Kong, Hong Kong)
Hong Yan, (City University of Hong Kong, Hong Kong)
Jinn-Moon Yang, (National Chiao Tung University,Taiwan)
S.M. Yiu, (The University of Hong Kong,Hong Kong)
Louxin Zhang, (National University of Singapore, Singapore)
Xuegong Zhang,(Tsinghua University, China)
Xiang Zhang, (Case Western Reserve University, USA)
Qiangfeng Cliff Zhang, (Stanford University, USA)
Hongyu Zhao, (Yale University, USA)
Xing-Ming Zhao, (Tongji University, China)
Zhongming Zhao, (Vanderbilt University, USA)
Jie Zheng, (Nanyang Technological University, Singapore)
Fengfeng Zhou, (Chinese Academy of Sciences, China)
Shuigeng Zhou, (Fudan University, China)
Shanfeng Zhu, (Fudan University, China)
Christine Sinoquet, (University De Nantes, France)
Fan Zhang, (Google, US)
Fei Guo, (Tianjin University, China)

Local Sponsoring Organization

DahShu is the sole legal entity organizing APBC 2016. It is a nonprofit corporation which is founded to promote the research and education in data science. Planned activities of DahShu include hosting scientific conferences and educational seminars to promote learning and knowledge sharing, as well as to showcase the latest innovations and achievements in data science and bioinformatics field.

DahShu has been granted IRS federal tax exempt status under Internal Revenue Code section 501(c)(3).

To learn more about our organization, please visit our website:
www.dahshu.org

Conference tips from DahShu:

1. The conference center provides a free wifi hotspot "**Paid Pass**" and the password is "**APBC2016**".
2. Salon F serves keynote speeches as well as one parallel session. Salon G-J are joined together to serve the other parallel session. Salon G and Salon H serve short courses. Salons B,C and E are joined together to serve Sponsor Exhibits, Poster Exhibits and Luncheon. **Salon A is reserved for all attendees to use as a focus room.**
3. This printed program book only includes following information: Welcome Message, Committees, Sponsors, Conference Venue and Floor Plan, Profiles of Keynote Speakers and Conference Schedule. **For more details such as accepted papers and posters, please check the electronic version of the program book in the USB disk**.
4. Blank pages are attached to the program book to be used as a note pad.

Conference Venue and Floor Plan

Conference Location: South San Francisco Conference Center
Address: 255 South Airport Boulevard, South San Francisco, CA 94080

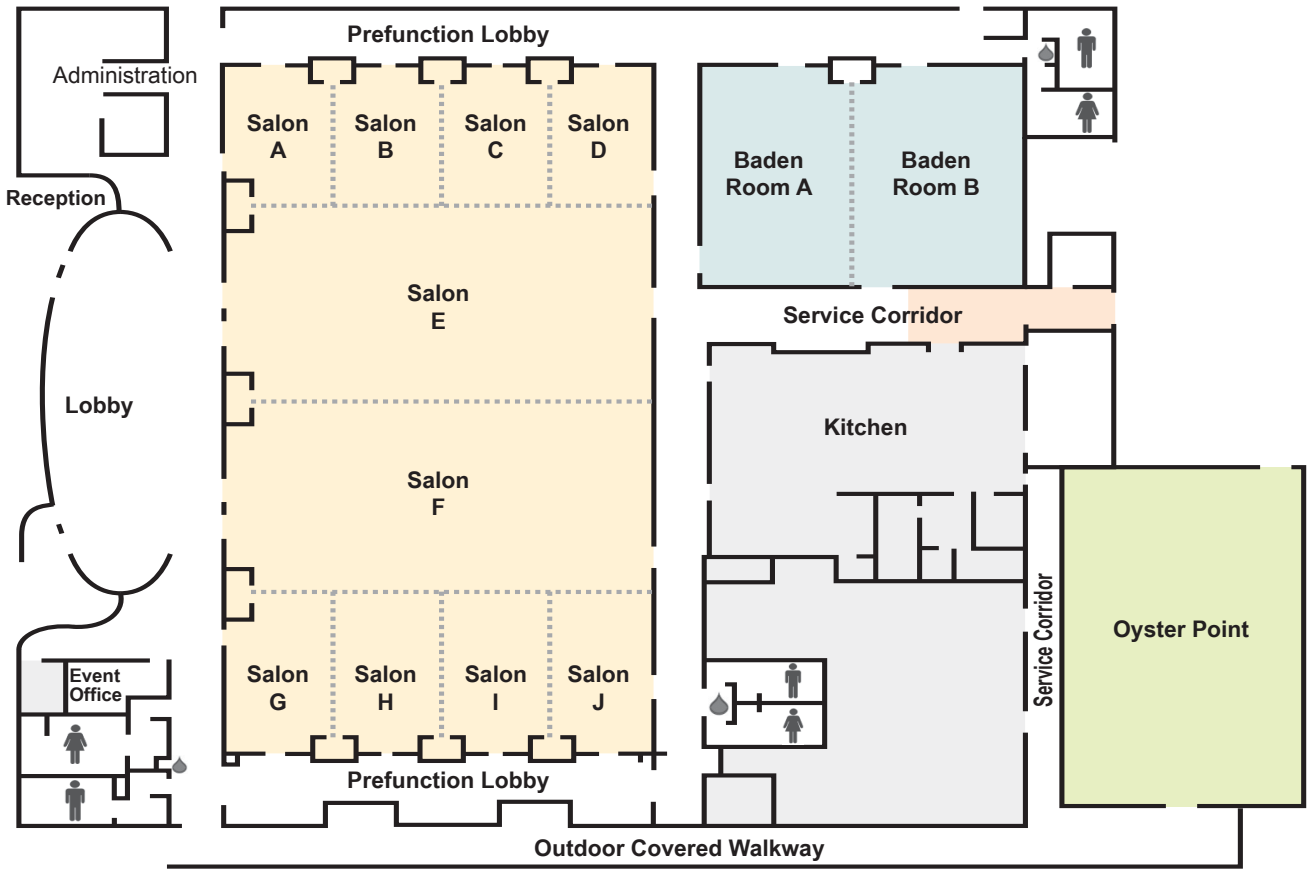
From the South (San Jose)

Take Highway 101 north to the South Airport Boulevard exit #424 (which is two miles north of the San Francisco International Airport). At the first stop light, drive straight across the intersection and directly into the Holiday Inn parking lot. The South San Francisco Conference Center is on the left.

From the North (San Francisco)

Take Highway 101 South to the South Airport Boulevard exit #424 in South San Francisco. Stay to the right and turn east under the freeway overpass. Make a right onto South Airport Boulevard. The South San Francisco Conference Center is located on the left between the Citi Garden Hotel and the Holiday Inn.

Parking is complimentary.



Food and Beverages

Food and Beverages Provided by the Conference
For all registered attendees, breakfast, lunch and all coffee breaks throughout the conference are complimentary. In addition, we will have a reception between 4:30-6pm on Monday the 11th and a banquet dinner between 6:30-8:30pm on Tuesday the 12th that are free and complimentary to all registered attendees. Vegetarian options are provided in each meal. For more details please check out the program section of this brochure.

The only thing you need to pay by yourself at the conference center is the beverages served from the bar during the Monday reception and the Tuesday dinner banquet. It is a full cash bar serving a wide variety of alcoholic and non-alcoholic beverages. Get your money ready and have some fun!!

Alternatives
All aforementioned food and beverages are western-style food. If you prefer Asian style food, which you will need to pay by yourself, there are abundant options around the conference center. More details please see here: http://www.yelp.com/search?find_desc=asian+restaurant&find_loc=south+san+francisco

The closest one is this Asian Buffet restaurant which is very reasonably priced and literally next door to the conference center https://www.zomato.com/south-san-francisco-ca/king-sun-buffet-south-san-francisco

Dairy Products and Lactose Intolerance
Unlike some East Asian Countries where dairy products have been pre-processed and the lactose content largely removed, in the US most dairy products (milk, yogurt, cream, cheese, and food prepared with these ingredients) could potentially contain substantial amount of lactose. If you are lactose intolerant, which is quite common among the Asian population (https://en.wikipedia.org/wiki/Lactose_intolerance), consume large amount of diary product may cause moderate to severe discomfort in the digestive system. Please consume with caution. There are over the counter medications that could help as well, please consult your physician for more details.

Sponsors

Genentech

CLI ventures

Gilead Sciences Inc

Clindata Insight Inc

Bina Technologies

Taylor & Francis Group

Sponsors

San Francisco Bay Area Chapter of the American Statistical Association

Peking University Alumni Association of Northern California

The International Chinese Statistical Association

Korean International Statistical Society

American Statistical Association

Keynote Speakers

Atul Butte, MD, PhD

Director, Institute for Computational Health Sciences
Professor of Pediatrics
University of California, San Francisco

Translating a Trillion Points of Data into Therapies, Diagnostics, and New Insights into Disease

Abstract: There is an urgent need to take what we have learned in our new “genome era” and use it to create a new system of precision medicine, delivering the best preventative or therapeutic intervention at the right time, for the right patients. Dr. Butte's lab at the University of California, San Francisco builds and applies tools that convert trillions of points of molecular, clinical, and epidemiological data -- measured by researchers and clinicians over the past decade and now commonly termed “big data” -- into diagnostics, therapeutics, and new insights into disease. Dr. Butte, a computer scientist and pediatrician, will highlight how publicly-available molecular measurements can be used to find new uses for drugs including drugs for inflammatory bowel disease and cancer, discovering new diagnostics include blood tests for complications during pregnancy, and how the next generation of biotech companies might even start in your garage.

Biography for Atul Butte

Atul Butte, MD, PhD is the inaugural Director of the Institute of Computational Health Sciences (ICHS) at the University of California, San Francisco, and a Professor of Pediatrics. Dr. Butte is also the Executive Director for Clinical Informatics across the six University of California Medical Schools and Medical Centers. Dr. Butte trained in Computer Science at Brown University, worked as a software engineer at Apple and Microsoft, received his MD at Brown University, trained in Pediatrics and Pediatric Endocrinology at Children's Hospital Boston, then received his PhD from Harvard Medical School and MIT. Dr. Butte has authored nearly 200 publications, with research repeatedly featured in Wired Magazine, the New York Times, and the Wall Street Journal. In 2015, Dr. Butte was elected into the National Academy of Medicine (formally known as the Institute of Medicine). In 2013, Dr. Butte was recognized by the White House as an Open Science Champion of Change for promoting science through publicly available data. Other recent awards include the 2014 E. Mead Johnson Award for Research in Pediatrics, 2013 induction into the American Society for Clinical Investigation, and the 2011 National Human Genome Research Institute Genomic Advance of the Month. Dr. Butte is also a founder of three investor-backed data-driven companies: Personalis, providing clinical interpretation of whole genome sequences, Carmenta (acquired by Progenity), discovering diagnostics for pregnancy complications, and NuMedii, finding new uses for drugs through open molecular data. Dr. Butte is also the principal investigator of the California Initiative to Advance Precision Medicine, and the principal investigator for ImmPort, the clinical and molecular data repository for the National Institute of Allergy and Infectious Diseases.

Keynote Speakers

Robert Gentleman, PhD

VP of Computational Biology
23andMe, Inc

Using GWAS and PheWAS for Drug Discovery

Abstract: Finding good, effective therapeutics is one of the major challenges facing health care. Strategies based on human genetics are starting to show real promise. I will discuss some of the challenges and opportunities. Combining PheWAS with GWAS can help improve target selection and identify potential concerns for adverse events.

Biography for Robert Gentleman

Robert Gentleman is the co-creator of the R language and he initiated the Bioconductor project. He has a long standing interest in scientific computing. After more than 20 years in academia he joined Genentech in 2009 and more recently moved to 23AndMe in April 2015.

Keynote Speakers

Rasmus Nielsen, PhD

Professor in Integrative Biology and Statistics
Director of the Center for Computational Biology
University of California, Berkeley

Finding the Footprints of Evolutional Adaptation in the Human Genome

Abstract: Evolutionary analyses of genomic data can provide important information regarding functional relationships, and we have developed computational methods for identifying positions in a genome targeted by selection. In this talk I will give two examples of the application of these methods. The first example concerns physiological adaptation to the hypoxic environment of the high-altitude plateau of Tibet. We have previously shown that Tibetans harbor genetic variants in two genes, EPAS1 and EGLN1, that affect hemoglobin production. Recently, we were able to demonstrate that the adaptive EPAS1 haplotype was transferred into humans by introgression from Denisovans. I will discuss recent progress on understanding the process of adaptive introgression in humans and its role in altitude adaptation. The second example is adaptation of the indigenous people of Greenland, the Inuit, to life in the Arctic, including low temperatures and a diet based primarily on fish and marine mammals and rich in ω-3 polyunsaturated fatty acids (PUFAs). Studies of Inuit have been used to argue for the benefits of a high dietary intake of ω-3 PUFAs. We recently performed the first scan of Inuit genomes for signatures of adaptation and found extreme signals in several loci, relating to metabolism of fatty acids, particularly PUFAs. Using association mapping, we show that the selected alleles have strong effects on a number of health-related phenotypes, and we replicate the findings in Europeans. Our results show that Inuit have unique physiological adaptions to life in the artic, in particular a diet rich in ω-3 PUFAs.

Biography for Rasmus Nielsen

Rasmus Nielsen's work is on development and application of statistical and population genetic methods in genomics. Much of his current research concerns statistical analyses of next-generation sequencing data, both in the context of medical genetics and population genetics. Methods developed by Nielsen include phylogeny based methods for detecting positive selection implemented in the popular program package PAML, for inferring demographic histories implemented in the IM and IMa programs, for detecting selective sweeps implemented in SweepFinder, and for analyzing Next Generation Sequencing (NGS) data implemented in ANGSD. He is a Professor of Computational Biology at UC Berkeley in the departments of Integrative Biology and of Statistics, and also serves as the director of the Center for Computational Biology at UC Berkeley. He has previously held faculty positions at Cornell University and at University of Copenhagen. He received his PhD from UC Berkeley in 1998.

Keynote Speakers

Wing Hung Wong, PhD

Stephen R. Pierce Family Goldman Sachs Professor in Science and Human Health
Department of Statistics and Department of Health Research Policy
Stanford University

Interpretation of Variants

Abstract: Every genome carries millions of genetic variants. Many of these variants are expected to have profound implications on health and diseases. Currently we have reasonable confidence in interpreting some of the variants that affect gene-coding regions of the genome. However, the overwhelming majority of the variants are located in non-coding parts of the genome and current methods to interpret such variants are woefully inadequate. In this talk I will discuss strategies that may help to close this gap in interpretation. Advances in this direction will be critical for the routine incorporation of genome sequence information to support health care decisions.

Biography for Wing Hung Wong

Dr. Wong is the Stephen R. Pierce Family Goldman Sachs Professor in Science and Human Health at Stanford University, and is a member of the National Academy of Sciences of the USA, Academia Sinica and Academy of Sciences of Hong Kong. Professor Wong's current research is motivated by problems from personalized medicine and systems biology. He is developing statistical methods and computational solutions to these problems. In the past his group has contributed a number of widely used bioinformatics tools, and technologies from his group had led to the formation of the several companies in the space of genomics data analysis and personalized prognostics.

Keynote Speakers

Bin Yu, PhD

Chancellor's Professor, Department of Statistics and
Electrical Engineering & Computer Science
University of California, Berkeley
Member, National Academy of Sciences

The Multi-Facets of a Data Science Project to Answer: How Are Organs Formed?

Abstract: Genome wide data reveal an intricate landscape where gene actions and interactions in diverse spatial areas are common both during development and in normal and abnormal tissues. Understanding local gene networks is thus key to developing treatments for human diseases. Given the size and complexity of recently available systematic spatial data, defining the biologically relevant spatial areas and modeling the corresponding local biological networks present an exciting and on-going challenge. It requires the integration of biology, statistics and computer science; that is, it requires data science. In this talk, I present results from a current project co-led by biologist Erwin Frise from Lawrence Berkeley National Lab (LBNL) to answer the fundamental systems biology question in the talk title. My group (Siqi Wu, Antony Joseph, Karl Kumbier) collaborates with Dr. Erwin and other biologists (Ann Hommands) of Celniker's Lab at LBNL that generate the Drosophila spatial expression embryonic image data. We leverage our group's prior research experience from computational neuroscience to use appropriate ideas of statistical machine learning in order to create a novel image representation decomposing spatial data into building blocks (or principal patterns). These principal patterns provide an innovative and biologically meaningful approach for the interpretation and analysis of large complex spatial data. They are the basis for constructing local gene networks, and we have been able to reproduce 11 out of the 12 links in a well-known gap-gene network of 6 genes. I will also present results from knock-out experiments in the Celniker Lab to validate our predictions on gene-gene interactions. Moreover, to understand the decomposition algorithm of images, we have derived sufficient and almost necessary conditions for local identifiability of the algorithm in the noiseless and complete case. Finally, we are collaborating with Dr. Wei Xue from Tsinghua Univ to devise a scalable open software package to manage the acquisition and computation of imaged data, designed in a manner that will be usable by biologists and expandable by developers.

Biography for Bin Yu

Bin Yu is Chancellor's Professor in the Departments of Statistics and of Electrical Engineering & Computer Science at the University of California at Berkeley. Her current research interests focus on statistics and machine learning theory, methodologies, and algorithms for solving high-dimensional data problems. Her group is engaged in interdisciplinary research with scientists from genomics, neuroscience, and remote sensing. She obtained her B.S. degree in Mathematics from Peking University in 1984, her M.A. and Ph.D. degrees in Statistics from the University of California at Berkeley in 1987 and 1990, respectively. She held faculty positions at the Univ of Wisconsin-Madison and Yale University and was a Member of Technical Staff at Bell Labs, Lucent. She was Chair of Department of Statistics at UC Berkeley from 2009 to 2012, and is a founding co-director of the Microsoft Lab on Statistics and Information Technology at Peking University, China. She is Member of the U.S. National Academy of Sciences and Fellow of the American Academy of Arts and Sciences. She was a Guggenheim Fellow in 2006, an Invited Speaker at ICIAM in 2011, and the Tukey Memorial Lecturer of the Bernoulli Society in 2012. She was President of IMS (Institute of Mathematical Statistics) in 2013-2014, and will be the Rietz Lecturer of IMS in 2016.

Banquet Speaker

Neil Risch, PhD

Lamond Family Foundation Distinguished Professor in Human Genetics
Professor of Epidemiology & Biostatistics
Director, Institute for Human Genetics
Member, National Academy of Medicine of the USA

Dr. Risch focuses on the development and application of statistical methods to address problems in human population genetics and genetic epidemiology. This has involved numerous projects using linkage analysis and positional cloning to identify novel disease genes, such as the genes causing hemochromatosis and torsion dystonia, as well as methodology for dissection of genetically complex traits including autism, hypertension, and multiple sclerosis. He has also spearheaded the approach of genome-wide association studies, the recent mainstay of human genetic analysis, and developed with investigators at Kaiser Permanente Northern California Division of Research a large genetic epidemiology research cohort on aging.

Program (Sunday, January 10th)

9:00 - 12:30	**Course1**: ENCODE Portal and Uniform Processing Pipelines : Open Web and Programatic Access to ENCODE Data, Metadata, and Software Pipelines **Instructor**: J.Michael Cherry Eurie Hong J.Seth Strattan Benjamin Hitz (Stanford University) **Place**: Salon G	**Course 2**: Machine Learning and Disease Taxonomy Study with Genomics Data for Precision Medicine **Instructor**: Rui Jiang Xuegong Zhang (Tsinghua University) **Place**: Salon H-I
14:00 - 17:30	**Course 3**: Modern Statistical Methods for Big Data and Big Questions in Genomics **Instructor**: Sandrine Dudoit Haiyan Huang Elizabeth Purdom (UC Berkeley) **Place:** Salon H-I	**Course 4**: The Roadmap Epigenomics Project: Data Access, Visualization and Integrative Analysis of 127 Human Epigenomes **Instructor**: Anshul Kundaje (Stanford University) **Place:** Salon G

For people who are attending both the morning and afternoon session, lunch will be provided

Program (Monday, January 11th)

7:45 - 8:30	**Breakfast (Lobby)**	
8:30 - 9:00	**Opening Remarks (Salon F)**	
9:00 - 10:00	**Keynote speech**: Translating a Trillion Points of Data into Therapies, Diagnostics, and New Insights into Disease **Speaker**: Dr. Atul Butte **Chair**: Bin Chen **Place**: Salon F	
10:00 - 10:25	**Coffee Break (Lobby)**	
parallel session	**Molecular Diagnosis With Genomic Data** **Chair:** Hyojung Paik **Place:** Salon F	**Recent Statistical Advancement in Network Based Analysis** **Chair:** Jingshu Wang **Place:** Salon G-J
10:25 - 10:50	**Gene Expression Profiling Identifies Candidate Biomarkers for Active and Latent Tuberculosis** **(Tzu-Ya Weng et al)**	**SUMONA: A Supervised Method for Optimizing Network Alignment** **(Tolga Can et al)**
10:50 - 11:15	**The Diagnostic Application of RNA-Sequencing in Patients with Thyroid Cancer: an Analysis of 851Variants and 133 Fusions in 524 Genes** **(Giulia C. Kennedy et al)**	**Characterizing Mutation-Expression Network Relationships in Multiple Cancers** **(Shila Ghazanfar et al)**
11:15 - 11:40	**Meta-Analysis of Sex Differences in Gene Expression in Schizophrenia** **(Hui Lu et al)**	**A Network based Covariance Test for Detecting Network Multivariate eQTL in Saccharomyces Cerevisiae** **(Minghua Deng et al)**
11:40 - 12:05	**Characterizing Redescriptions using Persistent Homology to Isolate Genetic Pathways Contributing to Pathogenesis** **(Daniel E Plattet al)**	**Generalized Logic Model Based on Network Topology Able to Capture the Dynamical Trends of Celluar Signaling Pathways** **(Jie Zheng et al)**
12:05 - 13:05	**Lunch Break (Hot Buffet Style Lunch will be served in Salon E)**	
13:05 - 14:05	**Keynote speech**: The Multi-Facets of a Data Science Project to Answer: How Are Organs Formed? **Speaker**: Dr. Bin Yu **Chair**: Haiyan Huang **Place**: Salon F	
14:05 - 14:30	**Coffee Break (Lobby)**	
parallel session	**Genetic Evolution:Recent advancement** **Chair:** Jijun Tang **Place:** Salon G-J	**miRNA Analysis** **Chair:** Rui (Sammi) Tang **Place:** Salon F
14:30 - 14:55	**Computational of Hybridization Networks on Realistic Phylogenetic Trees** **(Daniel H. Huson et al)**	The Modularity and Dynamicity of miRNA-mRNA Interactions in High-Grade Serous Ovarian Carcinomas and the Prognostic Implication (Kun Zhang et al)
14:55 - 15:20	**Algorithms for Pedigree Comparison** **(Lusheng Wang et al)**	**Identification of miRNA-mRNA Regulatory Modules by Exploring Collective Group Relationships** **(Jiuyong Li et al)**
15:20 - 15:45	**Genomic Duplication Problems for Unrooted Gene Trees** **(Jaroslaw Paszek et al)**	**c-Myc and Viral Cofactor Kaposin B Co-operate to Elicit Angiogenesis through Modulating miRNome Traits of Endothelial Cells** **(Cheng-Chung Cheng et al)**
15:45 - 16:10	**DTL-RnB: Algorithms and Tools for Summarizing the Space of DTL Reconciliations** **(W. Ma et al)**	**Identification of microRNA Precursor based on Gapped N-tuple Structure Status Composition** **(Bin Liu et al)**
16:10 - 16:35	**Mapping the Genomic Architecture of Adaptive Traits with Interspecific Introgressive Origin:A Coalescent-Based Approach** **(Kevin J Liu et al)**	**Missing Value Imputation for microRNA Expression Data by using a GO-based Functional Similarity Measure** **(Yang Yang et al)**
16:35 - 18:00	Reception & Poster Session I/Sponsor Exhibit (Complimentary Hors d'oeuvre will be served. A cash bar is available) **Place** : Salon E	

Program (Tuesday, January 12th)

7:45 - 8:30	Breakfast (Lobby)	
8:30 - 9:30	**Keynote speech**: Finding the Footprints of Evolutional Adaptation in the Human Genome **Speaker**: Dr. Rasmus Nielsen **Chair:** Jijun Tang **Place**: Salon F	
9:30 - 9:50	Coffee Break (Lobby)	
parallel session	**Protein Function and Statistical Methodology Development** **Chair:** Ilana.Belitskaya-Levy **Place:** Salon G-J	**Panel Discussion for Career Development Presented by Peking University Alumni** **Chair:** Ruixiao Lu **Place:** Salon F
9:50 - 10:15	Protein Inference: A Protein Quantification Perspective (Zengyou He et al)	Panelist: Dr. Fengzhu Sun
10:15 - 10:40	Predicting the Absorption Potential of Chemical Compounds through Deep-Learning Approach (Moonshik Shin et al)	Panelist: Dr. Jing Huang
10:40 - 11:05	Multi-Instance Multi-Label Distance Metric Learning for Genome-Wide Protein Function Prediction (Qingyao Wu et al)	Panelist: Dr. Jie Peng
11:05 - 11:20	Coffee Break (Lobby)	
11:20 - 11:45	Power Estimation and Sample Size Determination for Replication Studies of Genome-Wide Association Studies (Weichuan Yu et al)	Panelist: Dr. Xinmin Zhang
11:45 - 12:10	Protein-protein Interface Residues Share Similar Hexagon Neighborhood Conformations (Lusheng Wang et al)	Panelist: Dr. Minghua Deng
12:10 - 13:30	Lunch Break (Hot Buffet Style Lunch will be served at Salon E)	
parallel session	**Optimize Genomic Information in Sequencing Data** **Chair:** Kai Wang **Place:** Salon F	**Protein Function** **Chair:** Jingyi Jessica Li **Place:** Salon G-J
13:30 - 13:55	Identifying Micro-Inversions using High-Throughput Sequencing Reads (Huaiqiu Zhu et al)	SOHSite: Incorporating Evolutionary Information and Physicochemical Properties to Identify Protein S-sulfenylation Sites (Tzong-Yi Lee et al)
13:55 - 14:20	Locating Rearrangement Events in a Phylogeny based on Highly Fragmented Assemblies (David Sankoff et al)	PredRSA: A Gradient Boosted Regression Trees Approach for Predicting Protein Solvent Accessibility (Lei Deng et al)
14:20 - 14:45	Codon Context Optimization in Synthetic Gene Design (Georgios Papamichail et al)	A New Scheme to Discover Functional Associations and Regulatory Networks of Protein Ubiquitination (Shun-Long Weng et al)
14:45 - 15:10	A Maximum-likelihood Approach for Building Cell-Type Trees by Lifting (Bernard M.E. Moret et al)	Incorporating Two-Layered Machine Learning Methods with Substrate Motifs to Identify Lysine Ubiquitination Sites (Shun-Long Weng et al)
15:10 - 15:30	Coffee Break (Lobby)	
parallel session	**Algorithm Development and Machine Learning in Drug Discovery and Precision Medicine** **Chair:** Hui Yang **Place:** Salon F	**Metagenome and Trans-omits** **Chair:** John Lin **Place:** Salon G-J
15:30 - 15:55	Drug Repositioning Discovery for Non-Small Cell Lung Cancer by Using Machine Learning Algorithms and Topological Graph Theory (Ka-Lok Ng et al)	Comprehensive Prediction of lncRNA–RNA Interactions in Human Transcriptome (Michiaki Hamada et al)
15:55 - 16:20	PDOD: Prediction of Drugs Having Opposite Effects on Disease Genes in a Directed Network (Doheon Lee et al)	Computational Prediction of CRISPR Cassettes in Gut Metagenome Samples from Chinese Type-2 Diabetic Patients and Healthy Controls (Xuegong Zhang et al)
16:20 - 16:45	Algorithmic Mapping and Characterization of the Drug-Induced Phenotypic-Response Space of Parasites Causing Schistosomiasis (Rahul Singh et al)	Learning a Hierarchical Representation of the Yeast Transcriptomic Machinery using an Autoencoder Model (Xinghua Lu et al)
16:45 - 17:10	Inference of Domain-Disease Associations from Domain-Protein, Protein-Disease and Disease-Disease Relationships (Fengzhu Sun et al)	Transcriptome Sequencing Based Annotation and Homologous Evidence Based Scaffolding of Anguilla Japonica Draft Genome (Chung-Der Hsiao et al)
17:10 - 18:30	Poster Session II/Sponsor Exhibit (Salon E)	
18:30 - 20:30	On-site Banquet with Dinner Speech (open and complimentary to all registered attendees, Salon F) **Speaker**: Dr. Neil Risch	

Program (Wednesday, January 13th)

8:00 - 9:00	Breakfast (Lobby)	
9:00 - 10:00	**Keynote speech**: Using GWAS and PheWAS for Drug Discovery **Speaker**: Dr. Robert Gentleman **Chair**: Hua Tang **Place**: Salon F	
10:00 - 10:20	Coffee Break (Lobby)	
parallel session	**Epigenome and RNA-seq Data Analysis** **Chair:** Dvir Aran **Place:** Salon F	**Data Mining and Feature Extraction in Biomedical Applications** **Chair:** Min Yi **Place:** Salon G-J
10:20 - 10:45	Genome Reconstruction with ShRec3D+ and Hi-C data (Xiaodan Li et al)	A Semi-parametric Statistical Model for Intergrating Gene Expression Profiles across Different Platforms (Qunhua Li et al)
10:45 - 11:10	A Full Bayesian Partition Model for Identifying Hypo- and Hyper-methylated Loci from Single Nucleotide Resolution Sequencing Data (Jing Qiu et al)	Weakly Supervised Learning of Biomedical Information Extraction from Curated Data (Chun-Nam Hsu et al)
11:10 - 11:35	A Tale of Two Gene Sets:Low and High Variability in Single Cell RNA-seq data (Anagha Joshi et al)	hc-OTU: A Fast Homopolymer Compaction-Based Operation Taxonomic Unit Clustering Algorithm (Sungroh Yoon et al)
11:35 - 12:00		Medoidshift Clustering Applied to Genomic Bulk Tumor data (Russell Schwartz et al)
12:00 - 13:00	Lunch Break (Boxed sandwich lunch will be served outside Salon F)	
parallel session	**Epigenome** **Chair:** Hui Shen **Place:** Salon G-J	**RNA-Seq Data Analysis** **Chair:** Ruibin Xi **Place:** Salon F
13:00 - 13:25	Predicting Transcription Factor Site Occupancy using DNA Sequence Intrinsic and Cell-Type Specific Chromatin Features (Philipp Bucher et al)	Bayesian Method for Estimate Allele-Specific Expression based RNA-Seq (Masao Nagasaki et al)
13:25 - 13:50	Epigenome Overlap Measure(EpOM) for Comparing Tissue/Cell Types based on Chromatin States (Jingyi Jessica Li et al)	A Non-negative Matrix Factorization based Preselection Procedure for More Accurate Isoform Discovery form RNA-Seq data (Jingyi Jessica Li et al)
13:50 - 14:15	MOCCS: Clarifying DNA-binding Motif Ambiguity using ChIP-Seq data (Haruka Ozaki et al)	RDDpred: A Condition-Specific RNA-Editing Prediction Model from RNA-Seq data (Sun Kim et al)
14:15 - 14:30	Coffee Break (Lobby)	
14:30 - 15:30	**Keynote speech**: Interpretation of Variants **Speaker**: Dr. Wing Hung Wong **Chair**: Lu Tian **Place**: Salon F	
15:30 - 16:00	Closing Remarks Award Ceremony and Raffle Drawing (**Place: Salon F**)	

Accepted Papers

1. Weakly supervised learning of biomedical information extraction from curated data (Suvir Jain, Kashyap Tumkur, Tsung-Ting Kuo, Shitij Bhargava, Gordon Lin and Chun-Nam Hsu) 21

2. Drug repositioning for non-small cell lung cancer by using machine learning algorithms and topological graph theory (Chien-Hung Huang, Peter Mu-Hsin Chang, Chia-Wei Hsu, Chi-Ying F Huang, Ka-Lok Ng).....22

3. Gene expression profiling identifies candidate biomarkers for active and latent tuberculosis (Shih-Wei Lee, Lawrence Shih-Hsin Wu, Guan-Mau Huang, Kai-Yao Huang, Tzong-Yi Lee, Julia Tzu-Ya Weng).....23

4. Predicting transcription factor site occupancy using DNA sequence intrinsic and cell-type specific chromatin features (Sunil Kumar, Philipp Bucher).....24

5. A semi-parametric statistical model for integrating gene expression profiles across different platforms (Yafei Lyu, Qunhua Li) 25

6. The diagnostic application of RNA sequencing in patients with thyroid cancer: an analysis of 851 variants and 133 fusions in 524 genes (Moraima Pagan, Richard T. Kloos, Chu-Fang Lin, Kevin J. Travers, Hajime Matsuzaki, Ed Y. Tom, Su Yeon Kim, Mei G. Wong, Andrew C. Stewart, Jing Huang, P. Sean Walsh, Robert J.Monroe and Giulia C.Kennedy) 26

7. A full bayesian partition model for identifying hypo- and hyper-methylated loci from single nucleotide resolution sequencing data (Henan Wang, Chong He, Garima Kushwaha, Dong Xu and Jing Qiu) 27

8. PredRSA: a Gradient Boosted Regression Trees approach for predicting protein solvent accessibility (Chao Fan, Diwei Liu, Rui Huang, Zhigang Chen and Lei Deng) 28

9. Learning a hierarchical representation of the yeast transcriptomic machinery using an autoencoder model (Lujia Chen, Chunhui Cai, Vicky Chen, and Xinghua Lu).....29

10. Missing value imputation for microRNA expression data by using a GO-based similarity measure (Yang Yang, Zhuangdi Xu, Dandan Song) 30

11. Locating rearrangement events in a phylogeny based on highly fragmented assemblies (Chunfang Zheng and David Sankoff).....31

12. A Bayesian approach for estimating allele-specific expression from RNA-Seq data with diploid genomes (Naoki Nariai, Kaname Kojima, Takahiro Mimori, Yosuke Kawai and Masao Nagasaki).....32

13. Power estimation and sample size determination for replication studies of genome-wide association studies (Wei Jiang and Weichuan Yu).....33

14. A maximum-likelihood approach for building cell-type trees by lifting (Nishanth Ulhas Nair, Laura Hunter, Mingfu Shao,Paulina Grnarova, Yu Lin, Philipp Bucher and Bernard M.E. Moret) 34

15. Identification of miRNA-mRNA regulatory modules by exploring collective group relationships (S.M. Masud Karim, Lin Liu, Thuc Duy Le and Jiuyong Li) 35

16. Mapping the genomic architecture of adaptive traits with interspecific introgressive origin: A coalescent-based approach (Hussein A Hejase and Kevin J Liu) 36

17. SOHSite: incorporating evolutionary information and physicochemical properties to identify protein S- sulfenylation sites (Van-Minh Bui, Shun-Long Weng, Cheng-Tsung Lu, Tzu-Hao Chang, Julia Tzu-Ya Weng, and Tzong-Yi Lee) 37

18. Epigenome Overlap Measure (EPOM) for comparing tissue/cell types based on chromatin states (Wei Vivian Li, Zahra S. Razaee and Jingyi Jessica Li) 38

Accepted Papers

19. Medoidshift clustering applied to genomic bulk tumor data (Theodore Roman, Lu Xie and Russell Schwartz) 39

20. RDDpred: A condition-specific RNA-editing prediction model from RNA-seq data (Min-su Kim, Benjamin Hur and Sun Kim).....40

21. NMFP: a non-negative matrix factorization based preselection method to increase accuracy of identifying mRNA isoforms from RNA-seq data (Yuting Ye and Jingyi Jessica Li).....41

22. Identifying micro-inversions using high-throughput sequencing reads (Feifei He, Yang Li, Yu-Hang Tang, Jian Ma and Huaiqiu Zhu).....42

23. Comprehensive prediction of lncRNA–RNA interactions in human transcriptome (Goro Terai, Junichi Iwakiri, Tomoshi Kameda, Michiaki Hamada and Kiyoshi Asai) 43

24. Genomic duplication problems for unrooted gene trees (Jaroslaw Paszek and Pawel G´orecki) 44

25. Transcriptome sequencing based annotation and homologous evidence based scaffolding of Anguilla japonica draft genome (Yu-Chen Liu, Sheng-Da Hsu, Chih-Hung Chou, Wei-Yun Huang, Yu-Hung Chen,Chia-Yu Liu, Guan-Jay Lyu, Shao-Zhen Huang, Sergey Aganezov, Max A. Alekseyev, Chung-Der Hsiao and Hsien-Da Huang).....45

26. c-Myc and viral cofactor Kaposin B co-operate to elicit angiogenesisthrough modulating miRNome traits of endothelial cells (Hsin-Chuan Chang, Tsung-Han Hsieh, Yi-Wei Lee, Cheng-Fong Tsai, Ya-Ni Tsai, Cheng-Chung Cheng and Hsei-Wei Wang).....46

27. Prediction of drugs having opposite effects on disease genes in a directed network (Hasun Yu, Sungji Choo, Junseok Park, Jinmyung Jung, Yeeok Kang, Doheon Lee).....47

28. A new scheme to discover functional associations and regulatory networks of E3 ubiquitin ligases (Kai-Yao Huang, Julia Tzu-Ya Weng, Tzong-Yi Lee and Shun-Long Weng).....48

29. A network based covariance test for detecting network multivariate eQTL in saccharomyces cerevisiae (Huili Yuan, Zhenye Li, Nelson L. S. Tang and Minghua Deng) 49

30. Inference of domain-disease associations from domain-protein, protein-disease and disease-disease relationships (Wangshu Zhang, Marcelo P. Coba, Fengzhu Sun) 50

31. UbiSite: incorporating two-layered machine learning method with substrate motifs to predict ubiquitin-conjugation site on lysines (Chien-Hsun Huang, Min-Gang Su, Hui-Ju Kao, Jhih-Hua Jhong, Shun-Long Weng and Tzong-Yi Lee) 51

32. Computational prediction of CRISPR cassettes in gut metagenome samples from Chinese type-2 diabetic patients and healthy controls (Tatiana C. Mangericao, Zhanhao Peng, Xuegong Zhang) 52

33. Generalized logical model based on network topology to capture the dynamical trends of cellular signaling pathways (Fan Zhang, Haoting Chen, Li Na Zhao, Hui Liu, Teresa M. Przytycka and Jie Zheng) 53

34. Meta-analysis of sex differences in gene expression in schizophrenia (Wenyi Qin, Cong Liu, Monsheel Sodhi, Hui Lu).....54

35. Characterizing redescriptions using persistent homology to isolate genetic pathways contributing to pathogenesis (Daniel E Platt, Saugata Basu, Pierre A Zalloua and Laxmi Parida).....55

36. The Modularity and Dynamicity of miRNA-mRNA Interactions in High-Grade Serous Ovarian Carcinomas and the Prognostic Implication (Wensheng Zhang, Andrea Edwards, Wei Fan, Erik K. Flemington and Kun Zhang).....56

Accepted Papers

37. Protein Inference: A Protein Quantification Perspective (Zengyou He, Ting Huang, Xiaoqing Liu, Peijun Zhu, Ben Teng, Shengchun Deng) 57

38. Identification of microRNA precursor based on gapped n-tuple structure status composition kernel (Bin Liu, Longyun Fang) 58

39. Multi-Instance Multi-Label Distance Metric Learning for Genome-Wide Protein Function Prediction (Yonghui Xu, Huaqing Min, Hengjie Song, Qingyao Wu) 59

40. Protein-Protein Interface Residues Share Similar Hexagon Neighborhood Conformations (Fei Guo, Yijie Ding, Shuai Cheng Li, Lusheng Wang).....60

41. SUMONA: A Supervised Method for Optimizing Network Alignment (Erhun Giray Tuncay and Tolga Can) 61

42. Gene expression variability in mammalian embryonic stem cells using single cell RNA-seq data (Anna Mantsoki, Guillaume Devailly, Anagha Joshi) 62

43. MOCCS: clarifying DNA-binding motif ambiguity using ChIP-Seq data (Haruka Ozaki, Wataru Iwasaki) 63

44. Characterizing mutation-expression network relationships in multiple cancers (Shila Ghazanfar, Yee Hwa Yang).....64

45. Autumn Algorithm – Computation of Hybridization Networks for Realistic Phylogenetic Trees (Daniel H. Huson and Simone Linz) 65

46. DTL-RnB: Algorithms and Tools for Summarizing the Space of DTL Reconciliations (W. Ma ,D. Smirnov, J. Forma ,A.Schweickart, C. Slocum, S. Srinivasan and R. Libeskind-Hadas) 66

47. Algorithms for Pedigree Comparison (Zhi-Zhong Chen, Qilong Feng, Chao Shen, Jianxin Wang, Lusheng Wang) 67

48. Predicting the Absorption Potential ofChemical Compounds through a DeepLearning Approach (Moonshik Shin, Donjin Jang, Hojung Nam, Kwang Hyung Lee, and Doheon Lee) 68

49. hc-OTU: A Fast and Accurate Method forClustering Operational Taxonomic Unit based onHomopolymer Compaction (Seunghyun Park, Hyun-soo Choi, Byunghan Lee, Jongsik Chun, Joong-Ho Won and Sungroh Yoon).....69

50. Codon Context Optimization inSynthetic Gene Design (Dimitris Papamichail, Hongmei Liu, Vitor Machado, Nathan Gould, J. Robert Coleman, Georgios Papamichail) 70

51. 3D genome reconstruction with ShRec3D+and Hi-C data (Jiangeng Li, Wei Zhang, Xiaodan Li) 71

52. Algorithmic Mapping and Characterization of the Drug-Induced Phenotypic-Response Space of Parasites Causing Schistosomiasis (Rahul Singh, Rachel Beasley, Thavy Long, and Conor R. Caffrey).....72

Weakly supervised learning of biomedical information extraction from curated data

Suvir Jain[1†], Kashyap Tumkur[1†], Tsung-Ting Kuo[2], Shitij Bhargava[1], Gordon Lin[1] and Chun-Nam Hsu[2]

[1]Department of Computer Science and Engineering, Jacobs School of Engineering, University of California, San Diego, 9500 Gilman Drive, 92093 La Jolla, United States.

[2]Department of Biomedical Informatics, School of Medicine, University of California, San Diego, 9500 Gilman Drive, 92093 La Jolla, United States.

†Equal contributor Correspondence: chunnan@ucsd.edu

Background:

Numerous publicly available biomedical databases derive data by curating from literatures. The curated data can be useful as training examples for information extraction, but curated data usually lack the exact mentions and their locations in the text required for supervised machine learning. This paper describes a general approach to information extraction using curated data as training examples. The idea is to formulate the problem as cost-sensitive learning from noisy labels, where the cost is estimated by a committee of weak classifiers that consider both curated data and the text.

Results:

We test the idea on two information extraction tasks of Genome-Wide Association Studies (GWAS). The first task is to extract target phenotypes (diseases or traits) of a study and the second is to extract ethnicity backgrounds of study subjects for different stages (initial or replication). Experimental results show that our approach can achieve 87% of Precision-at-2 (P@2) for disease/trait extraction, and 0.83 of F1-Score for stage-ethnicity extraction, both outperforming their cost-insensitive baseline counterparts.

Conclusions:

The results show that curated biomedical databases can potentially be reused as training examples to train information extractors without expert annotation or refinement, opening an unprecedented opportunity of using “big data” in biomedical text mining.

Keywords:

Biomedical text mining; Natural language processing; Information extraction; Database curation; Machine learning

Drug repositioning for non-small cell lung cancer by using machine learning algorithms and topological graph theory

Chien-Hung Huang[1], Peter Mu-Hsin Chang[2], Chia-Wei Hsu[1], Chi-Ying F Huang[3], Ka-Lok Ng[4,5]§
[1]Department of Computer Science and Information Engineering, National Formosa University, Hu-Wei, Taiwan 63205.
[2] Division of Hematology and Oncology, Department of Medicine, Taipei Veterans General Hospital; Faculty of Medicine, National Yang Ming University, Taiwan 112;
[3]Institute of Biopharmaceutical Sciences, National Yang-Ming University, 155, Sec. 2, Linong Street, Taipei, Taiwan 112;
[4]Department of Bioinformatics and Medical Engineering, Asia University, Taichung, Taiwan 41354.
[5]Department of Medical Research, China Medical University Hospital, China Medical University, Taichung, Taiwan 40402
Correspondence: ppiddi@gmail.com

Background
Non-small cell lung cancer (NSCLC) is one of the leading causes of death globally, and research into NSCLC has been accumulating steadily over several years. Drug repositioning is the current trend in the pharmaceutical industry for identifying potential new uses for existing drugs and ac- celerating the development process of drugs, as well as reducing side effects.

Results
This work integrates two approaches - machine learning algorithms and topological parameter- based classification - to develop a novel pipeline of drug repositioning to analyze four lung can- cer microarray datasets, enriched biological processes, potential therapeutic drugs and targeted genes for NSCLC treatments. A total of 7(8) and 11(12) promising drugs (targeted genes) were discovered for treating early- and late-stage NSCLC, respectively. The effectiveness of these drugs is supported by the literature, experimentally determined in-vitro IC_{50} and clinical trials. This work provides better drug prediction accuracy than competitive research according to IC_{50} measurements.

Conclusions
With the novel pipeline of drug repositioning, the discovery of enriched pathways and potential drugs related to NSCLC can provide insight into the key regulators of tumorigenesis and the treatment of NSCLC. Based on the verified effectiveness of the targeted drugs predicted by this pipeline, we suggest that our drug-finding pipeline is effective for repositioning drugs.

Gene expression profiling identifies candidate biomarkers for active and latent tuberculosis

Shih-Wei Lee[1,2*], Lawrence Shih-Hsin Wu[3*], Guan-Mau Huang[4], Kai-Yao Huang[4], Tzong-Yi Lee[4,5], Julia Tzu-Ya Weng[4,5§]
[1]Taoyuan General Hospital, Ministry of Health and Welfare, Taoyuan, Taiwan
[2]Department of Life Sciences, National Central University, Taoyuan, Taiwan
[3]Institute of Medical Sciences, Tzu Chi University, Hualien, Taiwan
[4]Department of Computer Science and Engineering, Yuan Ze University, Taoyuan, Taiwan
[5]Innovation Center for Big Data and Digital Convergence, Yuan Ze University, Taoyuan, Taiwan
*These authors contributed equally to this work
§Corresponding author:julweng@saturn.yzu.edu.tw

Background
Tuberculosis (TB) is a serious infectious disease in that 90% of those latently infected with Mycobacterium tuberculosis present no symptoms, but possess a 10% lifetime chance of developing active TB. To prevent the spread of the disease, early diagnosis is crucial. However, current methods of detection require improvement in sensitivity, efficiency or specificity. In the present study, we conducted a microarray experiment, comparing the gene expression profiles in the peripheral blood mononuclear cells among individuals with active TB, latent infection, and healthy conditions in a Taiwanese population.

Results
Bioinformatics analysis revealed that most of the differentially expressed genes belonged to immune responses, inflammation pathways, and cell cycle control. Subsequent RT-PCR validation identified four differentially expressed genes, NEMF, ASUN, DHX29, and PTPRC, as potential biomarkers for the detection of active and latent TB infections. Receiver operating characteristic analysis showed that the expression level of PTPRC may discriminate active TB patients from healthy individuals,while ASUN could differentiate between the latent state of TB infection and healthy condition. In contrast, DHX29 may be used to identify latently infected individuals among active TB patients or healthy individuals. To test the concept of using these biomarkers as diagnostic support, we constructed classification models using these candidate biomarkers and found the Naïve Bayes-based model built with ASUN, DHX29, and PTPRC to yield the best performance.

Conclusions
Our study demonstrated that gene expression profiles in the blood can be used to identify not only active TB patients, but also to differentiate latently infected patients from their healthy counterparts. Validation of the constructed computational model in a larger sample size would confirm the reliability of the biomarkers and facilitate the development of a cost-effective and sensitive molecular diagnostic platform for TB. Keywords: tuberculosis, latent infection, gene expression, biomarker

Predicting transcription factor site occupancy using DNA sequence intrinsic and cell-type specific chromatin features

Sunil Kumar[1,2], Philipp Bucher[1,2§]
[1]Swiss Institute for Experimental Cancer Research (ISREC), School of Life Sciences, EPFL, Station 15, Lausanne CH-1015, Switzerland;
[2]Swiss Institute of Bioinformatics (SIB), EPFL, Station 15, Lausanne CH-1015, Switzerland
Corresponding author : philipp.bucher@epfl.ch

Background
Understanding the mechanisms by which transcription factors (TF) are recruited to their physiological target sites is crucial for understanding gene regulation. DNA sequence intrinsic features such as predicted binding affinity are often not very effective in predicting in vivo site occupancy and in any case could not explain cell- type specific binding events. Recent reports show that chromatin accessibility, nucleosome occupancy and specific histone post-translational modifications greatly influence TF site occupancy in vivo. In this work, we use machine-learning methods to build predictive models and assess the relative importance of different sequence- intrinsic and chromatin features in the TF-to-target-site recruitment process.

Methods
Our study primarily relies on recent data published by the ENCODE consortium. Five dissimilar TFs assayed in multiple cell-types were selected as examples: CTCF, JunD, REST, GABP and USF2. We used two types of candidate target sites: (a) predicted sites obtained by scanning the whole genome with a position weight matrix, and (b) cell-type specific peak lists provided by ENCODE. Quantitative in vivo occupancy levels in different cell-types were based on ChIP-seq data for the corresponding TFs. In parallel, we computed a number of associated sequence-intrinsic and experimental features (histone modification, DNase I hypersensitivity, etc.) for each site. Machine learning algorithms were then used in a binary classification and regression framework to predict site occupancy and binding strength, for the purpose of assessing the relative importance of different contextual features.

Results
We observed striking differences in the feature importance rankings between the five factors tested. PWM-scores were amongst the most important features only for CTCF and REST but of little value for JunD and USF2. Chromatin accessibility and active histone marks are potent predictors for all factors except REST. Structural DNA parameters, repressive and gene body associated histone marks are generally of little or no predictive value.

Conclusions
We define a general and extensible computational framework for analyzing the importance of various DNA-intrinsic and chromatin-associated features in determining cell-type specific TF binding to target sites. The application of our methodology to ENCODE data has led to new insights on transcription regulatory processes and may serve as example for future studies encompassing even larger datasets.

A semi-parametric statistical model for integrating gene expression profiles across different platforms

Yafei Lyu[1], Qunhua Li[2§]
[1]Huck Institute of Life Science, Pennsylvania State University, State College, PA 16801, USA
[2]Department of Statistics, Pennsylvania State University, State College, PA 16801, USA
[§]Corresponding author:qunhua.li@psu.edu

Background
Determining differentially expressed genes (DEGs) between biological samples is the key to understand how genotype gives rise to phenotype. RNA-seq and microarray are two main technologies for profiling gene expression levels. However, considerable discrepancy has been found between the DEGs detected using the two technologies. Integration data across these two platforms has the potential to improve the power and reliability of DEG detection.

Methods
We propose a rank-based semi-parametric model to determine DEGs using information across different sources and apply it to the integration of RNA-seq and microarray data. By incorporating both the significance of differential expression and the consistency across platforms, our method effectively detects DEGs with moderate but consistent signals. We demonstrate the effectiveness of our method using simulation studies, MAQC/SEQC data and a synthetic microRNA dataset.

Conclusions
Our integration method is not only robust to noise and heterogeneity in the data, but also adaptive to the structure of data. In our simulations and real data studies, our approach shows a higher discriminate power and identifies more biologically relevant DEGs than eBayes, DEseq and some commonly used meta-analysis methods.

The diagnostic application of RNA sequencing in patients with thyroid cancer: an analysis of 851 variants and 133 fusions in 524 genes

[1]Moraima Pagan, [1]Richard T. Kloos,[1] Chu-Fang Lin, [1]Kevin J. Travers, [1]Hajime Matsuzaki, [1]Ed Y. Tom, [1]Su Yeon Kim, [1]Mei G. Wong, [1]Andrew C. Stewart, [1]Jing Huang, [1]P. Sean Walsh, [1]Robert J.Monroe and [1]Giulia C.Kennedy
[1]Veracyte, Inc., South San Francisco, CA
Corresponding author : giulia@veracyte.com

Background
Thyroid carcinomas are known to harbor oncogenic driver mutations and advances in sequencing technology now allow the detection of these in fine needle aspiration biopsies (FNA). Recent work by The Cancer Genome Atlas (TCGA) Research Network has expanded the number of genetic alterations detected in papillary thyroid carcinomas (PTC). We sought to investigate the prevalence of these and other genetic alterations in diverse subtypes of thyroid nodules beyond PTC, including a variety of samples with benign histopathology. This is the first clinical evaluation of a large panel of TCGA-reported genomic alterations in thyroid FNAs.

Results
In FNAs, genetic alterations were detected in 19/44 malignant samples (43% sensitivity) and in 7/44 histopathology benign samples (84% specificity). Overall, after adding a cohort of tissue samples, 38/76 (50%) of histopathology malignant samples were found to harbor a genetic alteration, while 15/75 (20%) of benign samples were also mutated. The most frequently mutated malignant subtypes were medullary thyroid carcinoma (9/12, 75%) and PTC (14/30, 47%). Additionally, follicular adenoma, a benign subtype of thyroid neoplasm, was also found to harbor mutations (12/29, 41%). Frequently mutated genes in malignant samples included BRAF (20/76, 26%) and RAS (9/76, 12%). Of the TSHR variants detected, (6/7, 86%) were in benign nodules. In a direct comparison of the same FNA also tested by an RNA-based gene expression classifier (GEC), the sensitivity of genetic alterations alone was 42%, compared to the 91% sensitivity achieved by the GEC. The specificity based only on genetic alterations was 84%, compared to 77% specificity with the GEC.

Conclusions
While the genomic landscape of all thyroid neoplasm subtypes will inevitably be elucidated, caution should be used in the early adoption of published mutations as the sole predictor of malignancy in thyroid. The largest set of such mutations known to date detects only a portion of thyroid carcinomas in preoperative FNAs in our cohort and thus is not sufficient to rule out cancer. Due to the finding that variants are also found in benign nodules, testing only GEC suspicious nodules may be helpful in avoiding false positives and altering the extent of treatment when selected mutations are found.

A full bayesian partition model for identifying hypo- and hyper-methylated loci from single nucleotide resolution sequencing data

Henan Wang[1], Chong He[1], Garima Kushwaha[2] , Dong Xu[2]and Jing Qiu[3*]
[1]Department of Statistics, University of Missouri, Columbia, Missouri, USA.
[2]Department of Computer Science and Informatics Institute, University of Missouri, Columbia, USA
[3]Department of Applied Economics and Statistics, University of Delaware, Newark, DE, USA.
*Correspondence: qiujing@udel.edu

Background:
DNA methylation is an epigenetic modification that plays important roles on gene regulation. Study of whole-genome bisulfite sequencing and reduced representation bisulfite sequencing brings the availability of DNA methylation at single CpG resolution. The main interest of study on DNA methylation data is to test the methylation difference under two conditions of biological samples. However, the high cost and complexity of this sequencing experiment limits the number of biological replicates, which brings challenges to the development of statistical methods.

Results:
Bayesian modeling is well known to be able to borrow strength across the genome, and hence is a powerful tool for high-dimensional- low-sample- size data. In order to provide accurate identification of methylation loci, especially for low coverage data, we propose a full Bayesian partition model to detect differentially methylated loci under two conditions of scientific study. Since hypo-methylation and hyper-methylation have distinct biological implication, it is desirable to differentiate these two types of differential methylation. The advantage of our Bayesian model is that it can produce one-step output of each locus being either equal-, hypo- or hyper-methylated locus without further post-hoc analysis. An R package named as MethyBayes implementing the proposed full Bayesian partition model will be submitted to the bioconductor website upon publication of the manuscript.

Conclusions:
The proposed full Bayesian partition model outperforms existing methods in terms of power while maintaining a low false discovery rate based on simulation studies and real data analysis including bioinformatics analysis.

PredRSA: a Gradient Boosted Regression Trees approach for predicting protein solvent accessibility

Chao Fan[1], Diwei Liu[1], Rui Huang[1], Zhigang Chen[1] and Lei Deng[1,2*]
[1]School of Software, Central South University, No.22 Shaoshan South Road, 410075 Changsha, China
[2]Shanghai Key Laboratory of Intelligent Information Processing, No.220 Handan Road, 200433 Shanghai, China.
* Correspondence: leideng@csu.edu.cn

Background:
Protein solvent accessibility prediction is a pivotal intermediate step towards modeling protein tertiary structures directly from one-dimensional sequences. It also plays an important part in identifying protein folds and domains. Although some methods have been presented to the protein solvent accessibility prediction in recent years, the performance is far from satisfactory. In this work, we propose PredRSA, a computational method that can accurately predict relative solvent accessible surface area (RSA) of residues by exploring various local and global sequence features which have been observed to be associated with solvent accessibility. Based on these features, a novel and efficient approach, Gradient Boosted Regression Trees (GBRT), is first adopted to predict RSA.

Results:
Experimental results obtained from 5-fold cross-validation based on the Manesh-215 dataset show that the mean absolute error (MAE) and the Pearson correlation coefficient (PCC) of PredRSA are 9.0% and 0.75, respectively, which are better than that of the existing methods. Moreover, we evaluate the performance of PredRSA using an independent test set of 68 proteins. Compared with the state-of-the-art approaches (SPINE-X and ASAquick), PredRSA achieves a significant improvement on the prediction quality.

Conclusions:
Our experimental results show that the Gradient Boosted Regression Trees algorithm and the novel feature combination are quite effective in relative solvent accessibility prediction. The proposed PredRSA method could be useful in assisting the prediction of protein structures by applying the predicted RSA as useful restraints.

Learning a hierarchical representation of the yeast transcriptomic machinery using an autoencoder model

Lujia Chen[1], Chunhui Cai[1], Vicky Chen[1], and Xinghua Lu[1*]
[1]Department of Biomedical Informatics, University of Pittsburgh, 5607 Baum Blvd, Pittsburgh, PA 15237
[*] Corresponding author: Xinghua Lu, xinghua@pitt.edu

Background
A living cell has a complex, hierarchically organized signaling system that encodes and assimilates diverse environmental and intracellular signals, and it further transmits signals that control cellular responses, including a tightly controlled transcriptional program. An important and yet challenging task in systems biology is to reconstruct cellular signaling system in a data-driven manner. In this study, we investigate the utility of deep hierarchical neural networks in learning and representing the hierarchical organization of yeast transcriptomic machinery.

Results
We have designed a sparse autoencoder model consisting of a layer of observed variables and 4 layers of hidden variables. We applied the model to over a thousand of yeast microarrays to learn the encoding system of yeast transcriptomic machinery. After model selection, we evaluated whether the trained models captured biologically sensible information. We show that the latent variables in the first hidden layer correctly captured the signals of yeast transcription factors (TFs), obtaining a close to one-to-one mapping between latent variables and TFs. We further show that genes regulated by latent variables at higher hidden layers are often involved in a common biological process, and the hierarchical relationships between latent variables conform to existing knowledge. Finally, we show that information captured by the latent variables provide more abstract and concise representations of each microarray, enabling the identification of better separated clusters in comparison to gene-based representation.

Conclusions
Contemporary deep hierarchical latent variable models, such as the autoencoder, can be used to partially recover the organization of transcriptomic machinery.

Missing value imputation for microRNA expression data by using a GO-based similarity measure

Yang Yang[1,2*], Zhuangdi Xu[1], Dandan Song[3]

[1]Department of Computer Science and Engineering, Shanghai Jiao Tong University, 800 Dongchuan Rd., Shanghai 200240, China

[2]Key Laboratory of Shanghai Education Commission for Intelligent Interaction and Cognitive Engineering, Shanghai 200240, China

[3]School of Computer Science and Technology, Beijing Institute of Technology, Beijing 100081, China

*Corresponding author :Yang Yang - yangyang@cs.sjtu.edu.cn

Background

Missing values are commonly present in microarray data profiles. Instead of discarding genes or samples with incomplete expression level, missing values need to be properly imputed for accurate data analysis. The imputation methods can be roughly categorized as expression level-based and domain knowledge-based. The first type of methods only rely on expression data without the help of external data sources, while the second type incorporates available domain knowledge into expression data to improve imputation accuracy. In recent years, microRNA (miRNA) microarray has been largely developed and used for identifying miRNA biomarkers in complex human disease studies. Similar to mRNA profiles, miRNA expression profiles with missing values can be treated with the existing imputation methods. However, the domain knowledge-based methods are hard to be applied due to the lack of direct functional annotation for miRNAs. With the rapid accumulation of miRNA microarray data, it is increasingly needed to develop domain knowledge-based imputation algorithms specific to miRNA expression profiles to improve the quality of miRNA data analysis.

Results

We connect miRNAs with domain knowledge of Gene Ontology (GO) via their target genes, and define miRNA functional similarity based on the semantic similarity of GO terms in GO graphs. A new measure combining miRNA functional similarity and expression similarity is used in the imputation of missing values. The new measure is tested on two miRNA microarray datasets from breast cancer research and achieves improved performance compared with the expression-based method on both datasets.

Conclusions

The experimental results demonstrate that the biological domain knowledge can benefit the estimation of missing values in miRNA profiles as well as mRNA profiles. Especially, functional similarity defined by GO terms annotated for the target genes of miRNAs can be useful complementary information for the expression-based method to improve the imputation accuracy of miRNA array data. Our method and data are available to the public upon request.

Locating rearrangement events in a phylogeny based on highly fragmented assemblies

Chunfang Zheng[1] and David Sankoff[2*]
[1]Department of Mathematics and Statistics, University of Ottawa, 585 King Edward Avenue, Ottawa, Canada, K1N 6N5.
[2]Department of Mathematics and Statistics, University of Ottawa, 585 King Edward Avenue, Ottawa, Canada, K1N 6N5.
*Correspondence: sankoff@uottawa.ca

Background
The inference of genome rearrangement operations requires complete genome assemblies as input data, since a rearrangement can involve an arbitrarily large proportion of one or more chromosomes. Most genome sequence projects, especially those on non-model organisms for which no physical map exists, produce very fragmented assembles, so that a rearranged fragment may be impossible to identify because its two endpoints are on different scaffolds. However, breakpoints are easily identified, as long as they do not coincide with scaffold ends. For the phylogenetic context, in comparing a fragmented assembly with a number of complete assemblies, certain combinatorial constraints on breakpoints can be derived. We ask to what extent we can use breakpoint data between a fragmented genome and a number of complete genomes to recover all the arrangements in a phylogeny.

Results
We simulate genomic evolution via chromosomal inversion, fragmenting one of the genomes into a large number of scaffolds to represent the incompleteness of assembly. We identify all the breakpoints between this genome and the remainder. We devise an algorithm which takes these breakpoints into account in trying to determine on which branch of the phylogeny a rearrangement event occurred. We present an analysis of the dependence of recovery rates on scaffold size and rearrangement rate, and show that the true tree, the one on which the rearrangement simulation was performed, tends to be most parsimonious in estimating the number of true events inferred.

Conclusions
It is somewhat surprising that the breakpoints identified just between the fragmented genome and each of the others suffice to recover most of the rearrangements produced by the simulations. This holds even in parts of the phylogeny disjoint from the lineage of the fragmented genome.

A Bayesian approach for estimating allele-specific expression from RNA-Seq data with diploid genomes

Naoki Nariai[1,2], Kaname Kojima[2], Takahiro Mimori[2], Yosuke Kawai[2] and Masao Nagasaki[2*]

[1]Present address: Institute for Genomic Medicine, University of California, San Diego, 9500 Gilman Drive, La Jolla, California, 92093 USA.

[2]Department of Integrative Genomics, Tohoku Medical Megabank Organization, Tohoku University, 2-1 Seiryo-machi, Aoba-ku, Sendai, Miyagi, 980-8575 Japan.

*Correspondence: nagasaki@megabank.tohoku.ac.jp

Background

RNA-sequencing (RNA-Seq) has become a popular tool for transcriptome profiling in mammals. However, accurate estimation of allele-specific expression (ASE) based on alignments of reads to the reference genome is challenging, because it contains only one allele on a mosaic haploid genome. Even with the information of diploid genome sequences, precise alignment of reads to the correct allele is difficult because of the high-similarity between the corresponding allele sequences.

Results

We propose a Bayesian approach to estimate ASE from RNA-Seq data with diploid genome sequences. In the statistical framework, the haploid choice is modeled as a hidden variable and estimated simultaneously with isoform expression levels by variational Bayesian inference. Through the simulation data analysis, we demonstrate the effectiveness of the proposed approach in terms of identifying ASE compared to the existing approach. We also show that our approach enables better quantification of isoform expression levels compared to the existing methods, TIGAR2, RSEM and Cufflinks. In the real data analysis of the human reference lymphoblastoid cell line GM12878, some autosomal genes were identified as ASE genes, and skewed paternal X-chromosome inactivation in GM12878 was identified.

Conclusions

The proposed method, called ASE-TIGAR, enables accurate estimation of gene expression from RNA-Seq data in an allele-specific manner. Our results show the effectiveness of utilizing personal genomic information for accurate estimation of ASE. An implementation of our method is available at http://nagasakilab.csml.org/ase-tigar.

Power estimation and sample size determination for replication studies of genome-wide association studies

Wei Jiang and Weichuan Yu*
Department of Electronic and Computer Engineering, The Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong, China
*Correspondence: eeyu@ust.hk

Background:
Replication study is a commonly used verification method to filter out false positives in genome-wide association studies (GWAS). If an association can be confirmed in a replication study, it will have a high confidence to be true positive. To design a replication study, traditional approaches calculate power by treating replication study as another independent primary study. These approaches do not use the information given by primary study. Besides, they need to specify a minimum detectable effect size, which may be subjective. One may think to replace the minimum effect size with the observed effect sizes in the power calculation. However, this approach will make the designed replication study underpowered since we are only interested in the positive associations from the primary study and the problem of the "winner's curse" will occur.

Results:
An Empirical Bayes (EB) based method is proposed to estimate the power of replication study for each association. The corresponding credible interval is es- timated in the proposed approach. Simulation experiments show that our method is better than other plug-in based estimators in terms of overcoming the winner's curse and providing higher estimation accuracy. The coverage probability of given credible interval is well-calibrated in the simulation experiments. Weighted average method is used to estimate the average power of all underlying true associations. This is used to determine the sample size of replication study. Sample sizes are estimated on 6 diseases from Wellcome Trust Case Control Consortium (WTCC- C) using our method. They are higher than sample sizes estimated by plugging observed effect sizes in power calculation.

Conclusions:
Our new method can objectively determine replication study's sample size by using information extracted from primary study. Also the winner's curse is alleviated. Thus, it is a better choice when designing replication studies of GWAS. Availability and implementation: The R-package is available at: http:// bioinformatics.ust.hk/RPower.html.

A maximum-likelihood approach for building cell-type trees by lifting

Nishanth Ulhas Nair[1], Laura Hunter[2], Mingfu Shao[1],Paulina Grnarova[1], Yu Lin[3], Philipp Bucher[4,5†] and Bernard M.E. Moret[1,5*†]

[1]School of Computer and Communication Sciences, E´cole Polytechnique F´ed´erale de Lausanne (EPFL), EPFL IC IIF LCBB, INJ 211 (Batiment INJ), Station 14, CH-1015 Lausanne, Switzerland.

[2]Computer Science Department, Stanford University, Stanford, United States of America.

[3]Department of Computer Science and Engineering, University of California, San Diego, San Diego, United States of America.

[4]School of Life Sciences, E´cole Polytechnique F´ed´erale de Lausanne (EPFL), Lausanne, Switzerland.

[5]Swiss Institute of Bioinformatics, Lausanne, Switzerland.

Correspondence: bernard.moret@epfl.ch

Background:

In cell differentiation, a less specialized cell differentiates into a more specialized one, even though all cells in one organism have (almost) the same genome. Epigenetic factors such as histone modifications are known to play a significant role in cell differentiation. We previously introduce cell-type trees to represent the differentiation of cells into more specialized types, a representation that partakes of both ontogeny and phylogeny.

Results:

We propose a maximum-likelihood (ML) approach to build cell-type trees and show that this ML approach outperforms our earlier distance-based and parsimony-based approaches. We then study the reconstruction of ancestral cell types; since both ancestral and derived cell types can coexist in adult organisms, we propose a lifting algorithm to infer internal nodes. We present results on our lifting algorithm obtained both through simulations and on real datasets.

Conclusions:

We show that our ML-based approach outperforms previously proposed techniques such as distance-based and parsimony-based methods. We show our lifting-based approach works well on both simulated and real data.

Identification of miRNA-mRNA regulatory modules by exploring collective group relationships

S.M. Masud Karim*, Lin Liu, Thuc Duy Le and Jiuyong Li
School of Information Technology and Mathematical Sciences, University of South Australia, Mawson Lakes, SA 5095, Adelaide, Australia.
*Correspondence:masud.karim@mymail.unisa.edu.au

Background:
microRNAs (miRNAs) play an essential role in the post-transcriptional gene regulation in plants and animals. They regulate a wide range of biological processes by targeting messenger RNAs (mRNAs). Evidence suggests that miRNAs and mRNAs interact collectively in gene regulatory networks. The collective relationships between groups of miRNAs and groups of mRNAs may be more readily interpreted than those between individual miRNAs and mRNAs, and thus are useful for gaining insight into gene regulation and cell functions. Several computational approaches have been developed to discover miRNA-mRNA regulatory modules (MMRMs) with a common aim to elucidate miRNA-mRNA regulatory relationships. However, most existing methods do not consider the collective relationships between a group of miRNAs and the group of targeted mRNAs in the process of discovering MMRMs. Our aim is to develop a framework to discover MMRMs and reveal miRNA-mRNA regulatory relationships from the heterogeneous expression data based on the collective relationships.

Results:
We propose DIscovering COllective group RElationships (DICORE), an effective computational framework for revealing miRNA-mRNA regulatory relationships. We utilize the notation of collective group relationships to build the computational framework. The method computes the collaboration scores of the miRNAs and mRNAs on the basis of their interactions with mRNAs and miRNAs, respectively. Then it determines the groups of miRNAs and groups of mRNAs separately based on their respective collaboration scores. Next, it calculates the strength of the collective relationship between each pair of miRNA group and mRNA group using canonical correlation analysis, and the group pairs with significant canonical correlations are considered as the MMRMs. We applied this method to three gene expression datasets, and validated the computational discoveries.

Conclusions:
Analysis of the results demonstrates that a large portion of the regulatory relationships discovered by DICORE is consistent with the experimentally confirmed databases. Furthermore, it is observed that the top mRNAs that are regulated by the miRNAs in the identified MMRMs are highly relevant to the biological conditions of the given datasets. It is also shown that the MMRMs identified by DICORE are more biologically significant and functionally enriched.

Mapping the genomic architecture of adaptive traits with interspecific introgressive origin: A coalescent-based approach

Hussein A Hejase* and Kevin J Liu
Department of Computer Science and Engineering, Michigan State University, 428 S. Shaw Lane, 48824 East Lansing, MI, USA
*Correspondence: kjl@msu.edu

Recent studies of eukaryotes including human and Neandertal, mice, and butterflies have highlighted the major role that interspecific introgression has played in adaptive trait evolution. A common question arises in each case: what is the genomic architecture of the introgressed traits? One common approach that can be used to address this question is association mapping, which looks for genotypic markers that have significant statistical association with a trait. It is well understood that sample relatedness can be a confounding factor in association mapping studies if not properly accounted for. Introgression and other evolutionary processes (e.g., incomplete lineage sorting) typically introduce variation among local genealogies, which can also differ from global sample structure measured across all genomic loci. In contrast, state-of-the-art association mapping methods assume fixed sample relatedness across the genome, which can lead to spurious inference. We therefore propose a new association mapping method called Coal-Map, which uses coalescent-based models to capture local genealogical variation alongside global sample structure. Using simulated and empirical data reflecting a range of evolutionary scenarios, we compare the performance of Coal-Map against EIGENSTRAT, a leading association mapping method in terms of its popularity, power, and type I error control. Our empirical data makes use of hundreds of mouse genomes for which adaptive interspecific introgression has recently been described. We found that Coal-Map's performance is comparable or better than EIGENSTRAT in terms of statistical power and false positive rate. Coal-Map's performance advantage was greatest on model conditions that most closely resembled empirically observed scenarios of adaptive introgression. These conditions had: (1) causal SNPs contained in one or a few introgressed genomic loci and (2) varying rates of gene flow – from high rates to very low rates where incomplete lineage sorting dominated as a primary cause of local genealogical variation.

SOHSite: incorporating evolutionary information and physicochemical properties to identify protein S- sulfenylation sites

Van-Minh Bui[1,†], Shun-Long Weng[2,3,4,†], Cheng-Tsung Lu[1], Tzu-Hao Chang[5], Julia Tzu-Ya Weng[1,6,*] and Tzong-Yi Lee[1,6,*]
[1]Department of Computer Science and Engineering, Yuan Ze University, Taoyuan 320, Taiwan; [2]Department of Obstetrics and Gynecology, Hsinchu Mackay Memorial Hospital, Hsin-Chu 300, Taiwan;[3]Mackay Medicine, Nursing and Management College, Taipei 112, Taiwan;[4]Department of Medicine, Mackay Medical College, New Taipei City 252, Taiwan;[5]Graduate Institute of Biomedical Informatics, Taipei Medical University, Taipei 110, Taiwan;[6]Innovation Center for Big Data and Digital Convergence, Yuan Ze University, Taoyuan 320, Taiwan
†These authors contributed equally to this work.
*To whom correspondence should be addressed: francis@saturn.yzu.edu.tw julweng@saturn.yzu.edu.tw

Background
Protein S-sulfenylation is a type of post-translational modification (PTM) involving the covalent binding of a hydroxyl group to the thiol of a cysteine amino acid. Recent evidence has shown the importance of S-sulfenylation in various biological processes, including transcriptional regulation, apoptosis and cytokine signaling. Determining the specific sites of S-sulfenylation is fundamental to understanding the structures and functions of S-sulfenylated proteins. However, the current lack of reliable tools often limits researchers to use expensive and time-consuming laboratory techniques for the identification of S-sulfenylation sites. Thus, we were motivated to develop a bioinformatics method for investigating S-sulfenylation sites based on amino acid compositions and physicochemical properties

Results
In this work, physicochemical properties were utilized not only to identify S-sulfenylation sites from 1,096 experimentally verified S-sulfenylated proteins, but also to compare the effectiveness of prediction with other characteristics such as amino acid composition (AAC), amino acid pair composition (AAPC), solvent-accessible surface area (ASA), amino acid substitutionmatrix (BLOSUM62), position-specific scoring matrix (PSSM), and positional weighted matrix (PWM). Various prediction models were built using support vector machine (SVM) and evaluated by five-fold cross-validation. The model constructed from hybrid features, including PSSM and physicochemical properties, yielded the best performance with sensitivity, specificity, accuracy and MCC measurements of 0.746, 0.737, 0.738 and 0.337, respectively. The selected model also provided a promising accuracy (0.693) on an independent testing dataset. Additionally, we employed TwoSampleLogo to help discover the difference of amino acid composition among S-sulfenylation, S- glutathionylation and S-nitrosylation sites.

Conclusion
This work proposed a computational method to explore informative features and functions for protein S-sulfenylation. Evaluation by five-fold cross validation indicated that the selected features were effective in the identification of S-sulfenylation sites. Moreover, the independent testing results demonstrated that the proposed method could provide a feasible means for conducting preliminary analyses of protein S-sulfenylation. We also anticipate that the uncovered differences in amino acid composition may facilitate future studies of the extensive crosstalk among S-sulfenylation, S-glutathionylation and S-nitrosylation.

Epigenome Overlap Measure (EPOM) for comparing tissue/cell types based on chromatin states

Wei Vivian Li[1], Zahra S. Razaee[1] and Jingyi Jessica Li[1,2*]
[1]Department of Statistics, 8125 Math Sciences Bldg., University of California, Los Angeles, 90095-1554 CA, USA.
[2]Department of Human Genetics, University of California, Los Angeles, 90095-1554 CA, USA.
Correspondence: jli@stat.ucla.edu

Background
The dynamics of epigenomic marks in their relevant chromatin states regulate distinct gene expression patterns, biological functions and phenotypic variations in biological processes. The availability of high-throughput epigenomic data generated by next-generation sequencing technologies allows a data-driven approach to evaluate the similarities and differences of diverse tissue and cell types in terms of epigenomic features. While ChromImpute has allowed for the imputation of large-scale epigenomic information yielding more robust data used to capture meaningful relationships between biological samples, widely used methods such as hierarchical clustering and correlation analysis cannot adequately utilize epigenomic data to accurately reveal the distinction and grouping of different tissue and cell types.

Methods
We utilize a three-step testing procedure–ANOVA, t test and overlap test to identify tissue/cell-type-associated enhancers and promoters and to calculate a newly defined Epigenomic Overlap Measure (EPOM). EPOM results in a clear correspondence map of biological samples from different tissue and cell types through comparison of epigenomic marks evaluated in their relevant chromatin states.

Results
Correspondence maps by EPOM show strong capability in distinguishing and grouping different tissue and cell types and reveal biologically meaningful similarities between Heart and Muscle, Blood & T-cell and HSC & B-cell, Brain and Neurosphere, etc. The gene ontology enrichment analysis both supports and explains the discoveries made by EPOM and suggests that the associated enhancers and promoters demonstrate distinguishable functions across tissue and cell types. Moreover, the tissue/cell-type-associated enhancers and promoters show enrichment in the disease-related SNPs that are also associated with the corresponding tissue or cell types. This agreement suggests the potential of identifying causal genetic variants relevant to cell-type-specific diseases from our identified associated enhancers and promoters.

Conclusions
The proposed EPOM measure demonstrates superior capability in grouping and finding a clear correspondence map of biological samples from different tissue and cell types. The identified associated enhancers and promoters provide a comprehensive catalog to study distinct biological processes and disease variants in different tissue and cell types. Our results also find that the associated promoters exhibit more cell-type-specific functions than the associated enhancers do, suggesting that the non-associated promoters have more housekeeping functions than the non-associated enhancers.

Medoidshift clustering applied to genomic bulk tumor data

Theodore Roman[1,2], Lu Xie[1,2] and Russell Schwartz[1,3*]

[1]Computational Biology Department, School of Computer Science, Carnegie Mellon University, 5000 Forbes Ave, 15213 Pittsburgh, PA, USA.

[2]Joint Carnegie Mellon/University of Pittsburgh Ph.D. Program in Computational Biology, 5000 Forbes Ave, 15213 Pittsburgh, PA, USA.

[3] Department of Biological Sciences, Mellon College of Science, Carnegie Mellon University, 4400 Fifth Avenue, 15213, Pittsburgh, PA, USA.

Correspondence: russells@andrew.cmu.edu

Abstract

Despite the enormous medical impact of cancers and intensive study of their biology, detailed characterization of tumor growth and development remains elusive. This difficulty occurs in large part because of enormous heterogeneity in the molecular mechanisms of cancer progression, both tumor-to-tumor and cell-to-cell in single tumors. Advances in genomic technologies, especially at the single-cell level, are improving the situation, but these approaches are held back by limitations of the biotechnologies for gathering genomic data from heterogeneous cell populations and the computational methods for making sense of those data. One popular way to gain the advantages of whole-genome methods without the cost of single-cell genomics has been the use of computational deconvolution (unmixing) methods to reconstruct clonal heterogeneity from bulk genomic data. These methods, too, are limited by the difficulty of inferring genomic profiles of rare or subtly varying clonal subpopulations from bulk data, a problem that can be computationally reduced to that of reconstructing the geometry of point clouds of tumor samples in a genome space. Here, we present a new method to improve that reconstruction by better identifying subspaces corresponding to tumors produced from mixtures of distinct combinations of clonal subpopulations. We develop a nonparametric clustering method based on medoidshift clustering for identifying subgroups of tumors expected to correspond to distinct trajectories of evolutionary progression. We show on synthetic and real tumor copy-number data that this new method substantially improves our ability to resolve discrete tumor subgroups, a key step in the process of accurately deconvolving tumor genomic data and inferring clonal heterogeneity from bulk data.

RDDpred: A condition-specific RNA-editing prediction model from RNA-seq data

Min-su Kim[1], Benjamin Hur[1] and Sun Kim[1,2,3*]

[1]Interdisciplinary Program in Bioinformatics, Seoul National University, Seoul, Republic of Korea.

[2]Department of Computer Science and Engineering, Seoul National University, Seoul, Republic of Korea.

[3]Bioinformatics Institute, Seoul National University, Seoul, Republic of Korea.

Background

RNA-editing is an important post-transcriptional RNA sequence modification performed by two catalytic enzymes, “ADAR”(A-to-I) and “APOBEC”(C-to-U). By utilizing high-throughput sequencing technologies, the biological function of RNA-editing has been actively investigated. Currently, RNA-editing is considered to be a key regulator that controls various cellular functions, such as protein activity, alternative splicing pattern of mRNA, and substitution of miRNA targeting site. DARNED, a public RDD database, reported that there are more than 300-thousands RNA-editing sites detected in human genome(hg19). Moreover, multiple studies suggested that RNA-editing events occur in highly specific conditions. According to DARNED, 97.62% of registered editing sites were detected in a single tissue or in a specific condition, which also supports that the RNA-editing events occur condition-specifically. Since RNA-seq can capture the whole landscape of transcriptome, RNA-seq is widely used for RDD prediction. However, significant amounts of false positives or artefacts can be generated when detecting RNA-editing from RNA-seq. Since it is difficult to perform experimental validation at the whole-transcriptome scale, there should be a powerful computational tool to distinguish true RNA-editing events from artefacts.

Result

We developed RDDpred, a Random Forest RDD classifier. RDDpred reports potentially true RNA-editing events from RNA-seq data. RDDpred was tested with two publicly available RNA-editing datasets and successfully reproduced RDDs reported in the two studies (90%, 95%) while rejecting false-discoveries (NPV: 75%, 84%).

Conclusion

RDDpred automatically compiles condition-specific training examples without experimental validations and then construct a RDD classifier. As far as we know, RDDpred is the very first machine-learning based automated pipeline for RDD prediction. We believe that RDDpred will be very useful and can contribute significantly to the study of condition-specific RNA-editing. RDDpred is available at http://biohealth.snu.ac.kr/software/RDDpred.

NMFP: a non-negative matrix factorization based preselection method to increase accuracy of identifying mRNA isoforms from RNA-seq data

Yuting Ye[1] and Jingyi Jessica Li[2,3*]
[1]Division of Biostatistics, University of California, Berkeley, 94720 Berkeley, CA, US.
[2]Department of Statistics, 8125 Math Sciences Bldg., University of California, Los Angeles, 90095-1554 Los Angeles, CA, US.
[3]Department of Human Genetics, 695 Charles E. Young Drive South, University of California, Los Angeles, 90095-7088 Los Angeles, CA, US.
Correspondence: jli@stat.ucla.edu

Background
The advent of next-generation RNA sequencing (RNA-seq) has greatly advanced transcriptomic studies, including system-wide identification and quantification of mRNA isoforms under various biological conditions. A number of computational methods have been developed to systematically identify mRNA isoforms in a high-throughput manner from RNA-seq data. However, a common drawback of these methods is that their identified mRNA isoforms contain a high percentage of false positives, especially for genes with complex splicing structures, e.g., many exons and exon junctions.

Results
We have developed a preselection method called "Non-negative Matrix Factorization Preselection" (NMFP) which is designed to improve the accuracy of computational methods in identifying mRNA isoforms from RNA-seq data. We demonstrated through simulation and real data studies that NMFP can effectively shrink the search space of isoform candidates and increase the accuracy of two mainstream computational methods, Cufflinks and SLIDE, in their identification of mRNA isoforms.

Conclusion
NMFP is a useful tool to preselect mRNA isoform candidates for downstream isoform discovery methods. It can greatly reduce the number of isoform candidates while maintaining a good coverage of unknown true isoforms. Adding NMFP as an upstream step, computational methods are expected to achieve better accuracy in identifying mRNA isoforms from RNA-seq data.

Identifying micro-inversions using high-throughput sequencing reads

Feifei He[1], Yang Li[2], Yu-Hang Tang[4], Jian Ma[2,3 §] and Huaiqiu Zhu[1 §]

[1] State Key Laboratory for Turbulence and Complex Systems and Department of Biomedical Engineering, and Center for Quantitative Biology, Peking University, Beijing, 100871

[2] Department of Bioengineering, University of Illinois, Urbana, IL 61801, USA

[3] Carl R. Woese Institute for Genomic Biology, University of Illinois, Urbana, IL 61801, USA 4 Division of Applied Mathematics, Brown University, Providence, RI 02912, USA

[§]Correspondence: jianma@illinois.edu hqzhu@pku.edu.cn

Background

The identification of inversions of DNA segments shorter than read length (e.g., 100 bp), defined as micro-inversions (MIs), remains challenging for next-generation sequencing reads. It is acknowledged that MIs are important genomic variation and may play roles in causing genetic disease. However, current alignment methods are generally insensitive to detect MIs. Here we develop a novel tool, MID (Micro-Inversion Detector), to identify MIs in human genomes using next-generation sequencing reads.

Results

The algorithm of MID is designed based on a dynamic programming path-finding approach. What makes MID different from other variant detection tools is that MID can handle small MIs and multiple breakpoints within an unmapped read. Moreover, MID proves reliability in low coverage data by integrating multiple samples. Our evaluation demonstrated that MID outperforms Gustaf, which can currently detect inversions from 30 bp to 500 bp.

Conclusions

To our knowledge, MID is the first method that can efficiently and reliably identify MIs from unmapped short next-generation sequencing reads. MID is reliable on low coverage data, which is suitable for large-scale projects such as the 1000 Genomes Project (1KGP). MID identified previously unknown MIs from the 1KGP that overlap with genes and regulatory elements in the human genome. We also identified MIs in cancer cell lines from Cancer Cell Line Encyclopedia (CCLE). Therefore our tool is expected to be useful to improve the study of MIs as a type of genetic variant in the human genome. The source code can be downloaded from: http://cqb.pku.edu.cn/ZhuLab/MID.

Comprehensive prediction of lncRNA–RNA interactions in human transcriptome

Goro Terai[1†], Junichi Iwakiri[2†], Tomoshi Kameda[3], Michiaki Hamada[4,3*^] and Kiyoshi Asai[2,3^]

[1]INTEC Inc, 1-3-3 Shinsuna Koto-ku, 136-8637 Tokyo, Japan.

[2]Graduate School of Frontier Sciences, University of Tokyo, 5–1–5 Kashiwanoha, Kashiwa, 277–8562 Chiba, Japan.

[3]Biotechnology Research Institute for Drug Discovery, National Institute of Advanced Industrial Science and Technology (AIST), 2–41–6, Aomi, Koto-ku, 135–0064 Tokyo, Japan.

[4]Faculty of Science and Engineering, Waseda University, 55N–06–10, 3–4–1, Okubo Shinjuku-ku, 169–8555 Tokyo, Japan.

†Equal contributor

^Joint corresponding authors

*Correspondence: mhamada@waseda.jp

Motivation:

Recent studies have revealed that large numbers of non-coding RNAs are transcribed in humans, but only a few of them have been identified with their functions. Identification of the interaction target RNAs of the non-coding RNAs is an important step in predicting their functions. The current experimental methods to identify RNA–RNA interactions, however, are not fast enough to apply to a whole human transcriptome. Therefore, computational predictions of RNA–RNA interactions are desirable, but this is a challenging task due to the huge computational costs involved.

Results:

Here, we report comprehensive predictions of the interaction targets of lncRNAs in a whole human transcriptome for the first time. To achieve this, we developed an integrated pipeline for predicting RNA–RNA interactions on the K computer, which is one of the fastest super-computers in the world. Comparisons with experimentally-validated lncRNA–RNA interactions support the quality of the predictions. Additionally, we have developed a database that catalogs the predicted lncRNA–RNA interactions to provide fundamental information about the targets of lncRNAs.

Genomic duplication problems for unrooted gene trees

Jaroslaw Paszek* and Pawel G´orecki
University of Warsaw, Institute of Informatics, Banacha 2, 02-097 Warsaw, Poland
*Correspondence: jpaszek@mimuw.edu.pl

Background

Discovering the location of gene duplications and multiple gene duplication episodes is a fundamental issue in evolutionary molecular biology. The problem introduced by Guig´o et al. in 1996 is to map gene duplication events from a collection of rooted, binary gene family trees onto theirs corresponding rooted binary species tree in such a way that the total number of multiple gene duplication episodes is minimized. There are several models in the literature that specify how gene duplications from gene families can be interpreted as one duplication episode. However, in all duplication episode problems gene trees are rooted. This restriction limits the applicability, since unrooted gene family trees are frequently inferred by phylogenetic methods.

Results

In this article we show the first solution to the open problem of episode clustering where the input gene family trees are unrooted. In particular, by using theoretical properties of unrooted reconciliation, we show an efficient algorithm that reduces this problem into the episode clustering problems defined for rooted trees. We show theoretical properties of the reduction algorithm and evaluation of empirical datasets.

Conclusions

We provided algorithms and tools that were successfully applied to several empirical datasets. In particular, our comparative study shows that we can improve known results on genomic duplication inference from real datasets.

Transcriptome sequencing based annotation and homologous evidence based scaffolding of Anguilla japonica draft genome

Yu-Chen Liu[1], Sheng-Da Hsu[1], Chih-Hung Chou[1], Wei-Yun Huang[1], Yu-Hung Chen[1],Chia-Yu Liu[1], Guan-Jay Lyu[1], Shao-Zhen Huang[1], Sergey Aganezov[2,3], Max A. Alekseyev[2], Chung-Der Hsiao[4,*] and Hsien-Da Huang[1, 5, 6, *]
[1]Institute of Bioinformatics and Systems Biology, National Chiao Tung University, HsinChu, Taiwan
[2]Computational Biology Institute & Department of Mathematics, George Washington University
[3]Department of Higher Mathematics, ITMO University, St. Petersburg, Russia
[4]Department of Bioscience Technology, Chung Yuan Christian University, Chung-Li, Taiwan
[5]Department of Biological Science and Technology, National Chiao Tung University, HsinChu, Taiwan
[6]Department of Biomedical Science and Environmental Biology, Kaohsiung Medical University, Kaohsiung, Taiwan
correspondence email: cdhsiao@cycu.edu.tw

Background
Anguilla japonica (Japanese eel) is currently one of the most important research subjects in eastern Asia aquaculture. Enigmatic life cycle of the organism makes study of artificial reproduction extremely limited. Henceforth genomic and transcriptomic resources of eels are urgently needed to help solving the problems surrounding this organism across multiple fields. We hereby provide a reconstructed transcriptome from deep sequencing of juvenile (glass eels) whole body samples. The provided expressed sequence tags were used to annotate the currently available draft genome sequence. Homologous information derived from the annotation result was applied to improve the group of scaffolds into available linkage groups.

Results
With the transcriptome sequence data combined with publicly available expressed sequence tags evidences, 18,121 genes were structurally and functionally annotated on the draft genome. Among them, 3,921 genes were located in the 19 linkage groups. 137 scaffolds covering 13 million bases were grouped into the linkage groups in additional to the original partial linkage groups, increasing the linkage group coverage from 13% to 14%.

Conclusions
This annotation provide information of the coding regions of the genes supported by transcriptome based evidence. The derived homologous evidences pave the way for phylogenetic analysis of important genetic traits and the improvement of the genome assembly.

c-Myc and viral cofactor Kaposin B co-operate to elicit angiogenesisthrough modulating miRNome traits of endothelial cells

Hsin-Chuan Chang[1], Tsung-Han Hsieh[1], Yi-Wei Lee[2], Cheng-Fong Tsai[2], Ya-Ni Tsai[1,5] ,Cheng-Chung Cheng[3,§] and Hsei-Wei Wang[1,2,4]

[1]Institute of Microbiology and Immunology, National Yang-Ming University, Taipei,Taiwan;

[2]Institute of Biomedical Informatics, National Yang-Ming University, Taipei, Taiwan

[3]Division of Cardiology, Department of Internal Medicine, Tri-Service GeneralHospital, National Defense Medical Center, Taipei, Taiwan

[4]VGH-YM Genome Research Center, National Yang-Ming University, Taipei, Taiwan.

§Correspondence:chengcc@mail.ndmctsgh.edu.tw

Background

MicroRNAs (miRNAs) have emerged as master regulators of angiogenesis and other cancer-related events. Discovering new angiogenesis-regulating microRNAs (angiomiRs) will eventually help in developing new therapeutic strategies for tumor angiogenesis and cardiovascular diseases. Kaposi's sarcoma (KS), which is induced by the etiological infectious agent KS-associated herpesvirus (KSHV), is a peculiar neoplasm that expresses both blood and lymphatic endothelial markers and possesses extensive neovasculature. Using KSHV and its proteins as baits will be an efficient way to discover new angiomiRs in endothelial cells. Kaposin B is one of the latent viral genes and is expressed in all KSHV tumor cells. Since Kaposin B is a nuclear protein with no DNA-binding domain, it may regulate gene expression by incorporating itself into a transcription complex.

Results

We demonstrated that c-Myc and Kaposin B form a transcription complex and bind to the miR-221/-222 promoter, thereby affecting their expression and anti-angiogenic ability. By small RNA sequencing (smRNA-Seq), we revealed that 72.1% (173/240) of Kaposin B up-regulated and 46.5% (113/243) of Kaposin B down-regulated known miRNAs were regulated by c-Myc. We also found that 77 novel miRNA were up-regulated and 28 novel miRNAs were down-regulated in cells expressing both c-Myc and Kaposin B compared with cells expressing Kaposin B only. The result was confirmed by RNA-IP-seq data.

Conclusions

Our study identifies known and novel c-Myc-regulated microRNAs and reveals that a c-Myc-oriented program is coordinated by Kaposin B in KSHV-cells.

Prediction of drugs having opposite effects on disease genes in a directed network

Hasun Yu[1,2], Sungji Choo[1,2], Junseok Park[1,2], Jinmyung Jung[1,2], Yeeok Kang[1,2], Doheon Lee[1,2§]
[1]Department of Bio and Brain Engineering, KAIST, 291 Daehak-ro, Yuseong-gu, Daejeon, 305- 701, Republic of Korea
[2]Bio-Synergy Research Center, 291 Daehak-ro, Yuseong-gu, Daejeon, 305- 701, Daejeon, Republic of Korea
[§]Correspondence: dhlee@biosoft.kaist.ac.kr

Background
Developing novel uses of approved drugs, called drug repositioning, can reduce costs and times in traditional drug development. Network-based approaches have presented promising results in this field. However, even though various types of interactions such as activation or inhibition exist in drug-target interactions and molecular pathways, most of previous network-based studies disregarded this information.

Methods
We developed a novel computational method, Prediction of Drugs having Opposite effects on Disease genes (PDOD), for identifying drugs having opposite effects on altered states of disease genes. PDOD utilized drug-drug target interactions with 'effect type', an integrated directed molecular network with 'effect type' and 'effect direction', and disease genes with regulated states in diseases. With this information, we proposed a scoring function to discover drugs likely to restore altered states of disease genes using the path from a drug to a disease through the drug-drug target interactions, shortest paths from drug targets to disease genes in molecular pathways, and disease gene-disease associations.

Results
We collected drug-drug target interactions, molecular pathways, and disease genes with their regulated states in the diseases. PDOD is applied to 898 drugs with known drug-drug target interactions and nine diseases. We compared performance of PDOD for predicting known therapeutic drug-disease associations with the previous methods. PDOD outperformed other previous approaches which do not exploit directional information in molecular network. In addition, we provide a simple web service that researchers can submit genes of interest with their altered states and will obtain drugs seeming to have opposite effects on altered states of input genes at http://gto.kaist.ac.kr/pdod/index.php/main.

Conclusions
Our results showed that 'effect type' and 'effect direction' information in the network based approaches can be utilized to identify drugs having opposite effects on diseases. Our study can offer a novel insight into the field of network-based drug repositioning.

A new scheme to discover functional associations and regulatory networks of E3 ubiquitin ligases

Kai-Yao Huang[1], Julia Tzu-Ya Weng[1,2], Tzong-Yi Lee[1,2,*] and Shun-Long Weng[3,4,5,*]
[1]Department of Computer Science and Engineering, Yuan Ze University, Taoyuan 320, Taiwan
[2]Innovation Center for Big Data and Digital Convergence, Yuan Ze University, Taoyuan 320, Taiwan
[3]Department of Obstetrics and Gynecology, Hsinchu Mackay Memorial Hospital, Hsin-Chu 300,Taiwan
[4]Mackay Medicine, Nursing and Management College, Taipei 112, Taiwan
[5]Department of Medicine, Mackay Medical College, New Taipei City 252, Taiwan
*Correspondence:francis@saturn.yzu.edu.tw 4467@mmh.org.tw

Background:
Protein ubiquitination catalyzed by E3 ubiquitin ligases play important modulatory roles in various biological processes. With the emergence of high-throughput mass spectrometry technology, the proteomics research community embraced the development of numerous experimental methods for the determination of ubiquitination sites. The result is an accumulation of ubiquitinome data, coupled with a lack of available resources for investigating the regulatory networks among E3 ligases and ubiquitinated proteins. In this study, by integrating existing ubiquitinome data, experimentally validated E3 ligases and established protein-protein interactions, we have devised a strategy to construct a comprehensive map of protein ubiquitination networks.

Results:
In total, 41,392 experimentally verified ubiquitination sites from 12,786 ubiquitinated proteins of humans have been obtained for this study. Additional 494 E3 ligases along with 1220 functional annotations and 28588 protein domains were manually curated. To characterize the regulatory networks among E3 ligases and ubiquitinated proteins, a well-established network viewer was utilized for the exploration of ubiquitination networks from 40892 protein-protein interactions. The effectiveness of the proposed approach was demonstrated in a case study examining E3 ligases involved in the ubiquitination of tumor suppressor p53. In addition to Mdm2, a known regulator of p53, the investigation also revealed other potential E3 ligases that may participate in the ubiquitination of p53.

Conclusion:
Aside from the ability to facilitate comprehensive investigations of protein ubiquitination networks, by integrating information regarding protein-protein interactions and substrate specificities, the proposed method could discover potential E3 ligases for ubiquitinated proteins. Our strategy presents an efficient means for the preliminary screen of ubiquitination networks and overcomes the challenge as a result of limited knowledge about E3 ligase-regulated ubiquitination.

A network based covariance test for detecting network multivariate eQTL in saccharomyces cerevisiae

Huili Yuan[1], Zhenye Li[1], Nelson L. S. Tang[4] and Minghua Deng[1,2,3*]
[1]LMAM, School of Mathematical Sciences, Peking University, Yiheyuan Road, 100871 Beijing, China.
[2]Center for Quantitative Biology, Peking University, Yiheyuan Road, 100871 Beijing, China.
[3]Center for Statistical Sciences, Peking University, Yiheyuan Road, 100871 Beijing, China.
[4]Department of Chemical Pathology, Prince of Wales Hospital, Faculty of Medicine, The Chinese University of Hong Kong, Shatin, Hong Kong, China.

Background:
Expression quantitative trait locus (eQTL) analysis has been widely used to understand how genetic variations affect gene expressions in the biological systems. Traditional eQTL is investigated in a pair-wise manner in which one SNP affects the expression of one gene. In this way, some associated markers found in GWAS have been related to disease mechanism by eQTL study. However, in real life, biological process is usually performed by a group of genes. Although some methods have been proposed to identify a group of SNPs that affect the mean of gene expressions in the network, the change of co-expression pattern has not been considered. So we propose a process and algorithm to identify the marker which affects the co-expression pattern of a pathway. Considering two genes may have different correlations under different isoforms which is hard to detect by the linear test, we also consider the nonlinear test.

Results:
When we applied our method to yeast eQTL dataset profiled under both the glucose and ethanol conditions, we identified a total of 166 modules, with each module consisting of a group of genes and one eQTL where the eQTL regulate the co-expression patterns of the group of genes. We found that many of these modules have biological significance.

Conclusions:
We propose a network based covariance test to identify the SNP which affects the structure of a pathway. We also consider the nonlinear test as considering two genes may have different correlations under different isoforms which is hard to detect by linear test.

Inference of domain-disease associations from domain-protein, protein-disease and disease-disease relationships

Wangshu Zhang[1], Marcelo P. Coba[2,3], Fengzhu Sun[1,4,*]
[1]Molecular and Computational Biology Program, University of Southern California, 1050 Childs Way, Los Angeles, USA; [2]Zilkha Neurogenetic Institute;3 Department of Psychiatry and Behavioral Sciences, Keck School of Medicine, University of Southern California, Los Angeles, CA; 4 Centre for Computational Systems Biology, School of Mathematical Sciences, Fudan University, Shanghai, China
*Corresponding author: fsun@usc.edu

Background
Protein domains can be viewed as portable units of biological function that defines the functional properties of proteins. Therefore, if a protein is associated with a disease, protein domains might also be associated and define disease endophenotypes. However, knowledge about such domain-disease relationships is rarely available. Thus, identification of domains associated with human diseases would greatly improve our understanding of the mechanism of human complex diseases and further improve the prevention, diagnosis and treatment of these diseases.

Methods
Based on phenotypic similarities among diseases, we first group diseases into overlapping modules. We then develop a framework to infer associations between domains and diseases through known relationships between diseases and modules, domains and proteins, as well as proteins and disease modules. Different methods including Association, Maximum likelihood estimation (MLE), Domain-disease pair exclusion analysis (DPEA), Bayesian, and Parsimonious explanation (PE) approaches are developed to predict domain-disease associations.

Results
We demonstrate the effectiveness of all the five approaches via a series of validation experiments, and show the robustness of the MLE, Bayesian and PE approaches to the involved parameters. We also study the effects of disease modularization in inferring novel domain-disease associations. Through validation, the AUC (Area Under the operating characteristic Curve) scores for Bayesian, MLE, DPEA, PE, and Association approaches are 0.86, 0.84, 0.83, 0.83 and 0.79, respectively, indicating the usefulness of these approaches for predicting domain-disease relationships. Finally, we choose the Bayesian approach to infer domains associated with two common diseases, Crohn's disease and type 2 diabetes.

Conclusions
The Bayesian approach has the best performance for the inference of domain-disease relationships. The predicted landscape between domains and diseases provides a more detailed view about the disease mechanisms

UbiSite: incorporating two-layered machine learning method with substrate motifs to predict ubiquitin-conjugation site on lysines

Chien-Hsun Huang[1,2], Min-Gang Su[1], Hui-Ju Kao[1], Jhih-Hua Jhong[1], Shun-Long Weng[3,4,5,*] and Tzong-Yi Lee[1,6,*]
[1]Department of Computer Science and Engineering, Yuan Ze University, Taoyuan 320, Taiwan;[2]Tao-Yuan Hospital, Ministry of Health & Welfare, Taoyuan 320, Taiwan;[3]Department of Obstetrics and Gynecology, Hsinchu Mackay Memorial Hospital, Hsin-Chu 300, Taiwan;[4]Mackay Medicine, Nursing and Management College, Taipei 112, Taiwan;
[5]Department of Medicine, Mackay Medical College, New Taipei City 252, Taiwan;[6]Innovation Center for Big Data and Digital Convergence, Yuan Ze University, Taoyuan 320, Taiwan
[§]Correspondence: francis@saturn.yzu.edu.tw 4467@mmh.org.tw

Background:
The conjugation of ubiquitin to a substrate protein (protein ubiquitylation), which involves a sequential process – E1 activation, E2 conjugation and E3 ligation, is crucial to the regulation of protein function and activity in eukaryotes. This ubiquitin-conjugation process typically binds the last amino acid of ubiquitin (glycine 76) to a lysine residue of a target protein. The high-throughput of mass spectrometry-based proteomics has stimulated a large-scale identification of ubiquitin-conjugated peptides. Hence, a new web resource, UbiSite, was developed to identify ubiquitin-conjugation site on lysines based on large-scale proteome dataset.

Results:
Given a total of 37,647 ubiquitin-conjugated proteins, including 128026 ubiquitylated peptides, obtained from various resources, this study carries out a large-scale investigation on ubiquitin-conjugation sites based on sequenced and structural characteristics. A TwoSampleLogo reveals that a significant depletion of histidine (H), arginine (R) and cysteine (C) residues around ubiquitylation sites may impact the conjugation of ubiquitins in closed three-dimensional environments. Based on the large-scale ubiquitylation dataset, a motif discovery tool, MDDLogo, has been adopted to characterize the potential substrate motifs for ubiquitin conjugation. Not only are single features such as amino acid composition (AAC), positional weighted matrix (PWM), position-specific scoring matrix (PSSM) and solvent-accessible surface area (SASA) considered, but also the effectiveness of incorporating MDDLogo-identified substrate motifs into a two-layered prediction model is taken into account. Evaluation by five-fold cross-validation showed that PSSM is the best feature in discriminating between ubiquitylation and non-ubiquitylation sites, based on support vector machine (SVM). Additionally, the two-layered SVM model integrating MDDLogo-identified substrate motifs could obtain a promising accuracy and the Matthews Correlation Coefficient (MCC) at 81.06% and 0.586, respectively. Furthermore, the independent testing showed that the two-layered SVM model could outperform other prediction tools, reaching at 85.10% sensitivity, 69.69% specificity, 73.69% accuracy and the 0.483 of MCC value.

Conclusion:
The independent testing result indicated the effectiveness of incorporating MDDLogo-identified motifs into the prediction of ubiquitylation sites. In order to provide meaningful assistance to researchers interested in large-scale ubiquitinome data, the two-layered SVM model has been implemented onto a web-based system (UbiSite), which is freely available at http://csb.cse.yzu.edu.tw/UbiSite/. Two cases given in the UbiSite provide a demonstration of effective identification of ubiquitylation sites with reference to substrate motifs.

Computational prediction of CRISPR cassettes in gut metagenome samples from Chinese type-2 diabetic patients and healthy controls

Tatiana C. Mangericao[1,2], Zhanhao Peng[1], Xuegong Zhang[1,3,*]
[1]MOE Key Lab of Bioinformatics/Bioinformatics Division, TNLIST/Center for Synthetic and Systems Biology, and Department of Automation, Tsinghua University, Beijing 100084, China
[2]Department of Bioengineering, Instituto Superior Técnico (IST), Lisbon, Portugal
[3]School of Life Sciences, Tsinghua University, Beijing 100084, China
*Corresponding author: zhangxg@tsinghua.edu.cn

Background:
CRISPR has been becoming a hot topic as a powerful technique for genome editing for human and other higher organisms. The original CRISPR-Cas (Clustered Regularly Interspaced Short Palindromic Repeats coupled with CRISPR-associated proteins) is an important adaptive defence system for prokaryotes that provides resistance against invading elements such as viruses and plasmids. A CRISPR cassette contains short nucleotide sequences called spacers. These unique regions retain a history of the interactions between prokaryotes and their invaders in individual strains and ecosystems. One important ecosystem in the human body is the human gut, a rich habitat populated by a great diversity of microorganisms. Gut microbiomes are important for human physiology and health. Metagenome sequencing has been widely applied for studying the gut microbiomes. Most efforts in metagenome study has been focused on profiling taxa compositions and gene catalogues and identifying their associations with human health. Less attention has been paid to the analysis of the ecosystems of microbiomes themselves especially their CRISPR composition.

Results:
We conducted a preliminary analysis of CRISPR sequences in a human gut metagenomic data set of Chinese individuals of type-2 diabetes patients and healthy controls. Applying an available CRISPR-identification algorithm, PILER-CR, we identified 3,169 CRISPR cassettes in the data, from which we constructed a set of 1,302 unique repeat sequences and 36,709 spacers. A more extensive analysis was made for the CRISPR repeats: these repeats were submitted to a more comprehensive clustering and classification using the web server tool CRISPRmap. All repeats were compared with known CRISPRs in the database CRISPRdb. A total of 784 repeats had matches in the database, and the remaining 518 repeats from our set are potentially novel ones.

Conclusions:
The computational analysis of CRISPR composition based contigs of metagenome sequencing data is feasible. It provides an efficient approach for finding potential novel CRISPR arrays and for analysing the ecosystem and history of human microbiomes.

Generalized logical model based on network topology to capture the dynamical trends of cellular signaling pathways

Fan Zhang[1], Haoting Chen[1,2], Li Na Zhao[3], Hui Liu[1,4], Teresa M. Przytycka[5] and Jie Zheng[1,6,7*]
[1]Biomedical Informatics Graduate Lab, School of Computer Engineering, Nanyang Technological University, 639798, Singapore, Singapore.
[2]School of Engineering and Applied Sciences, Columbia University, 10027, New York, NY, USA.
[3]Bioinformatics Institute, Agency for Science, Technology and Research, 138671, Singapore, Singapore.
[4]Lab of Information, Management, Changzhou University, 213164, Changzhou, Jiangsu, China.
[5]National Center for Biotechnology Information, NLM/NIH, Bethesda, MD, USA.
[6]Complexity Institute, Nanyang Technological University, Singapore, Singapore.
[7]Genome Institute of Singapore, Agency for Science, Technology and Research, 138672, Singapore, Singapore.
* Correspondence: zhengjie@ntu.edu.sg

Background:
Cellular responses to extracellular perturbations require signaling pathways to capture and transmit the signals. However, the underlying molecular mechanisms of signal transduction are not yet fully understood, thus detailed and comprehensive models may not be available for all the signaling pathways. In particular, insufficient knowledge of parameters, which is a long-standing hindrance for quantitative kinetic modeling necessitates the use of parameter-free methods for modeling and simulation to capture dynamic properties of signaling pathways.

Results:
We present a computational model that is able to simulate the graded responses to degradations, the sigmoidal biological relationships between signaling molecules and the effects of scheduled perturbations to the cells. The simulation results are validated using experimental data of protein phosphorylation, demonstrating that the proposed model is capable of capturing the main trend of protein activities during the process of signal transduction. Compared with existing simulators, our model has better performance on predicting the state transitions of signaling networks.

Conclusion:
The proposed simulation tool provides a valuable resource for modeling cellular signaling pathways using a knowledge-based method.

Meta-analysis of sex differences in gene expression in schizophrenia

Wenyi Qin[1], Cong Liu[1], Monsheel Sodhi[2,*], Hui Lu[1,3*]

[1]Department of Bioengineering, University of Illinois at Chicago, Chicago, IL, USA

[2]Department of Pharmacy Practice and Center for Pharmaceutical Biotechnology, University of Illinois at Chicago, Chicago, IL, USA

[3]SJTU-Yale Joint Center for Biostatistics, Shanghai Jiaotong University, Shanghai, China

*Co-corresponding authors: mssodhi@uic.edu huilu@uic.edu

Abstract:

Schizophrenia is a severe psychiatric disorder which influences around 1% of the worldwide population. Differences between male and female patients with schizophrenia have been noted. There is an earlier age of onset in males compared with females with this diagnosis, and in addition, there are differences in symptom profiles between the sexes. The underlying molecular mechanism of sex difference remains unclear. Here we present a comprehensive analysis to reveal the sex differences in gene expression in schizophrenia with stringent statistics criteria. We compiled a data set consisting of 89 male controls, 90 male schizophrenia patients, 35 female controls and 32 female schizophrenia patients from six independent studies of the prefrontal cortex (PFC) in postmortem brain. When we tested for a sex by diagnosis interaction on gene expression, 23 genes were up-regulated and 23 genes were down-regulated in the male group (q-value < 0.05), several genes are related to energy metabolism, while 4 genes are located on sex chromosome. No genes were statistically significant in the female group when multiple testing correction were conducted (q-value <0.05), most likely due to the small sample size. Our protocol and results from the male group provide a starting point for identifying the underlying different mechanism between male and female schizophrenia patients.

Characterizing redescriptions using persistent homology to isolate genetic pathways contributing to pathogenesis

Daniel E Platt[1*], Saugata Basu[2], Pierre A Zalloua[3,4] and Laxmi Parida[1]
[1]Computational Biology Center, IBM T. J. Watson Research Center, 1101 Kitchawan Rd., 10598, Yorktown Hgts, NY, United States.
[2]Department of Mathematics, Purdue University, 150 N. University St., 47907, West Lafayette, IN, United States.
[3]Graduate Studies and Research, Lebanese American University, P.O. Box 13-5053, Chouran Beirut: 1102 2801, Lebanon.
[4]Department of Environmental Health, Harvard University, 401 Park Drive, Boston, MA, United States.
*Correspondence: watplatt@us.ibm.com

Background:
Complex diseases may have multiple pathwyas leading to disease. E.g. coronary artery disease evolves from arterial damage to their epithelial layers, but has multiple causal pathways. More challenging, those pathways are highly correlated within metabolic syndrome. The challenge is to identify specific clusters of phenotype characteristics (composite phenotypes) that may reflect these different etiologies. Further, GWAS seeking to identify SNPs satisfying multiple composite phenotype descriptions allows for lower false positive rates at lower α thresholds, allowing for the possibility of reducing false negatives. This may provide a window into the missing heritability problem.

Methods:
We identify significant phenotype patterns, and identify fuzzy redescriptions among those patterns using Jaccard distances. Further, we construct Vietoris-Rips complexes from the Jaccard distances and compute the persistent homology associated with those. The patterns comprising these topological features are identified as composite phenotpyes, whose genetic associations are explored with logistic regression applied to pathways and to GWAS.

Results:
We identified several phenotypes that tended to be dominated by metabolic syndrome descriptions, and which were distinct among the combinations of metabolic syndrome conditions. Among SNPs marking the RAAS complex, various SNPs associated specifically with different groups of composite phenotypes, as well as distinguishing between the composite phenotypes and simple phenotypes. Each of these showed different genetic associations, namely rs6693954, rs762551, rs1378942, and rs1133323. GWAS identified SNPs that associated with composite phenotypes included rs12365545, rs6847235, and rs701319. Eighteen GWAS identified SNPs appeared in combinations supported in composite combinations with greater power than for any individual phenotype.

Conclusions:
We do find systematic associations among metabolic syndrome variates that show distinctive genetic association profiles. Further, the systematic characterization involves composite phenotype descriptions that allow for combined power of individual phenotype GWAS tests, yielding more significance for lower individual thresholds, permitting the exploration of SNPs that would otherwise show as false negatives.

The Modularity and Dynamicity of miRNA-mRNA Interactions in High-Grade Serous Ovarian Carcinomas and the Prognostic Implication

Wensheng Zhang[1], Andrea Edwards[1], Wei Fan[2], Erik K. Flemington[3], Kun Zhang[1§]
[1]Department of Computer Science, Xavier University of Louisiana, 1 Drexel Drive, New Orleans LA 70125
[2]Big Data Lab, Baidu Research, 1195 Bordeaux Dr., Sunnyvale, CA 94089
[3]Tulane Health Sciences Center, Tulane Cancer Center, Tulane University, 1700 Tulane Ave, New Orleans, LA 70112
[§]Corresponding author: kzhang@xula.edu

Abstract

Ovarian carcinoma is the fifth-leading cause of cancer death among women in the United States. Major reasons for this persistent mortality include the poor understanding of the underlying biology and a lack of reliable biomarkers. Previous studies have shown that aberrantly expressed MicroRNAs (miRNAs) are involved in carcinogenesis and tumor progression by post-transcriptionally regulating gene expression. However, the interference of miRNAs in tumorigenesis is quite complicated and far from being fully understood. In this work, by an integrative analysis of mRNA expression, miRNA expression and clinical data published by The Cancer Genome Atlas (TCGA), we studied the modularity and dynamicity of miRNA-mRNA interactions and the prognostic implications in high-grade serous ovarian carcinomas. With the top transcriptional correlations (Bonferroni-adjusted p-value < 0.01) as inputs, we identified five miRNA-mRNA module pairs (MPs), each of which included one positive-connection (correlation) module and one negative-connection (correlation) module. The number of miRNAs or mRNAs in each module varied from 3 to 7 or from 2 to 873. Among the four major negative-connection modules, three fit well with the widely accepted miRNA-mediated post-transcriptional regulation theory. These modules were enriched with the genes relevant to cell cycle and immune response. Moreover, we proposed two novel algorithms to reveal the group or sample specific dynamic regulations between these two RNA classes. The obtained miRNA-mRNA dynamic network contains 3350 interactions captured across different cancer progression stages or tumor grades. We found that those dynamic interactions tended to concentrate on a few miRNAs (e.g. miRNA-936), and were more likely present on the miRNA-mRNA pairs outside the discovered modules. In addition, we also pinpointed a robust prognostic signature consisting of 56 modular protein-coding genes, whose co-expression patterns were predictive for the survival time of ovarian cancer patients in multiple independent cohorts.

Protein Inference: A Protein Quantification Perspective

Zengyou He[a,b,*], Ting Huang[c], Xiaoqing Liu[a], Peijun Zhu[a], Ben Teng[a], Shengchun Deng[d]
[a]School of Software, Dalian University of Technology, Dalian, China
[b]Key Laboratory for Ubiquitous Network and Service Software of Liaoning, Dalian, China.
[c]College of Computer and Information Science, Northeastern University, USA.
[d]School of Computer Science and Engineering, Harbin Institute of Technology, China.

Abstract

In mass spectrometry-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 from the raw data. Then the relative or absolute protein abundance is estimated with quantification methods, such as spectral counting. Until now, most researchers have been dealing with these two processes separately. In fact, the protein inference problem can be regarded as a special protein quantification problem in the sense that truly present proteins are those proteins whose abundance values are not zero. Some recent published papers have conceptually discussed this possibility. However, there is still a lack of rigorous experimental studies to test this hypothesis.
In this paper, we investigate the feasibility of using protein quantification methods to solve the protein inference problem. Protein inference methods aim to determine whether each candidate protein is present in the sample or not. Protein quantification methods estimate the abundance value of each inferred protein. Naturally, the abundance value of an absent protein should be zero. Thus, we argue that the protein inference problem can be viewed as a special protein quantification problem in which one protein is considered to be present if its abundance is not zero. Based on this idea, our paper tries to use three simple protein quantification methods to solve the protein inference problem effectively. The experimental results on six data sets show that these three methods are competitive with previous protein inference algorithms. This demonstrates that it is plausible to model the protein inference problem as a special protein quantification task, which opens the door of devising more effective protein inference algorithms from a quantification perspective. The source codes of our methods are available at: http://code.google.com/p/protein- inference/.

Identification of microRNA precursor based on gapped n-tuple structure status composition kernel

Bin Liu[1,2*], Longyun Fang[1]
[1]School of Computer Science and Technology, Harbin Institute of Technology Shenzhen Graduate School, Shenzhen, Guangdong, China 518055
[2]Key Laboratory of Network Oriented Intelligent Computation, Harbin Institute of Technology Shenzhen Graduate School, Shenzhen, Guangdong, China 518055
*Corresponding authors:bliu@insun.hit.edu.cn

Abstract

MicroRNAs (miRNAs) are small non-coding RNA molecules, functioning in transcriptional and post-transcriptional regulation of gene expression. The discrimination of the real premiRNAs from the false ones (such as hairpin sequences with similar stem-loops) is necessary for the understanding of miRNAs' role in the control of cell life and death. Since both their small size and sequence specificity, it cannot be based on sequence information alone, but requires structure information about the miRNA precursor to get satisfactory performance. Kmers are convenient and widely used features for modeling the properties of miRNAs and other biological sequences. However, Kmers suffer from the inherent limitation that if the parameter K is increased to incorporate long range effects, some certain Kmer will appear rarely or even not appear, as a consequence, most Kmers absent and a few present once. Thus, the statistical learning approaches using Kmers as features become susceptible to noisy data once K becomes large. To address this problem, we introduce alternative feature sets using gapped n-tuple structure status composition, a new classifier, imiRNA-GSSC, and a general method for robust estimation of Kmer frequencies. To make the method applicable to large-scale genome wide applications, we adopted an efficient tree data structure for computing the kernel matrix. We show that compared to the original imiRNA-kmer and alternative approaches, our imiRNA-GSSC identifies miRNA precursors with significantly improved accuracy. We then show that, imiRNA-GSSC trained with human data can correctly predict 82.35% of the 4022 pre-miRNAs from 11 different species ranging from animals, plants and viruses. imiRNA-GSSC would be a useful high throughput tool for large-scale analysis of microRNA precursors

Multi-Instance Multi-Label Distance Metric Learning for Genome-Wide Protein Function Prediction

Yonghui Xu[a], Huaqing Min[b], Hengjie Song[b], Qingyao Wu[b]
[a]School of Computer Science and Engineering, South China University of Technology, Guangzhou, China, 510006.
[b]School of Software Engineering, South China University of Technology, Guangzhou, China, 510006.
Email addresses:
xu.yonghui@hotmail.com (Yonghui Xu),
hqmin@scut.edu.cn (Huaqing Min),
song009@e.ntu.edu.sg (Hengjie Song),
qyw@scut.edu.cn (Qingyao Wu)

Abstract

Multi-instance multi-label (MIML) learning has been proven to be effective for the genome-wide protein function prediction problems where each training ex- ample is associated with not only multiple instances but also multiple class labels. To find an appropriate MIML learning method for genome-wide protein function prediction, many studies in the literature attempted to optimize objective functions in which dissimilarity between instances is measured using the Euclidean distance. But in many real applications, Euclidean distance may be unable to capture the intrinsic similarity/dissimilarity in feature space and label space. Unlike other previous approaches, in this paper, we propose to learn a multi-instance multi-label distance metric learning framework (MIMLDML) for genome-wide protein function prediction. Specifically, we learn a Mahalanobis distance to preserve and utilize the intrinsic geometric information of both fea- ture space and label space for MIML learning. In addition, we try to deal with the sparsely labeled data by giving weight to the labeled data. Extensive ex- periments on seven real-world organisms covering the biological three-domain system (i.e., archaea, bacteria, and eukaryote[1]) show that the MIMLDML algorithm is superior to most state-of-the-art MIML learning algorithms.

Protein-Protein Interface Residues Share Similar Hexagon Neighborhood Conformations

Fei Guo[a], Yijie Ding[a], Shuai Cheng Li[b], Lusheng Wang[b,*]
[a]School of Computer Science and Technology, Tianjin University, 92 Weijin Road, Nankai District, Tianjin, P.R.China
[b]Department of Computer Science, City University of Hong Kong, 83 Tat Chee Avenue, Kowloon, Hong Kong
∗Corresponding author:cswangl@cityu.edu.hk (Lusheng Wang)

Abstract

Studies on protein-protein interaction are important in proteome research. How to build more effective models based on sequence information, struc- ture information and physicochemical characteristics, is the key technology in protein-protein interaction identification. In this paper, we aim to ana- lyze the neighborhood information on interface, through a hexagon structure. Using 6,438 complexes, we build a non-redundant database of hexagon neigh- borhood on interface. We propose a novel method for identifying protein- protein interface, via hexagon structure similarity. Interacting residues are extracted with top values of hexagon structure similarity, and interacting patches can be clustered as predicted interface residues. Experiments show that our method achieves better results than some state-of-the-art methods for identifying protein-protein interface. Comparing to existing methods, our approach improves F – measure value by at least 0.03. On Benchmark v4.0, our method has overall precision and recall values of 55% and 56%. On CAPRI targets, our method has overall precision and recall values of 52% and 55%. We produce an efficient and accurate method for protein-protein interaction identification in the molecular biology research.

SUMONA: A Supervised Method for Optimizing Network Alignment

[1]Erhun Giray Tuncay and [2]Tolga Can
[1]Ministry of Science, Industry and Tech- nology, Ankara, Turkey and the Department of Computer Engineering, Ege University, Izmir, Turkey
[2]Department of Computer Engineering, Middle East Technical University, Ankara, Turkey
Correspondence: erhungiray.tuncay@sanayi.gov.tr

Abstract

This study focuses on improving the multi objective memetic algorithm for PPI alignment, Optimizing Network Aligner - OptNetAlign, via integration with other existing network alignment methods such as Spinal, NETAL and HubAlign. The output of this algorithm is an elite set of aligned networks all of which are optimal with respect to multiple user defined criteria. However, OptNetAlign is an unsupervised genetic algorithm that initiates its search with completely random solutions and it requires substantial running times to generate an elite set of solutions that have high scores with respect to the given criteria. In order to improve running time, the search space of the algorithm can be narrowed down by focusing on the most desired criteria and trying to optimize other relevant criteria on a more limited set of solutions. The method presented in this study improves OptNetAlign in a supervised fashion by utilizing the alignment results of different network alignment algorithms with varying parameters that depend upon user preferences. Therefore, the user can prioritize certain objectives upon others and achieve better running time performance while optimizing the secondary objectives.

Gene expression variability in mammalian embryonic stem cells using single cell RNA-seq data

Anna Mantsoki[1], Guillaume Devailly[1], Anagha Joshi[1§]
[1]The Roslin institute, University of Edinburgh, Easter bush campus, Midlothian, EH25 9RG.
[§]Corresponding author: Anagha.joshi@roslin.ed.ac.uk

Background
Gene expression heterogeneity contributes to development as well as disease progression. Due to technological limitations, most studies to date have focused on differences in mean expression across experimental conditions, rather than differences in gene expression variance. The advent of single cell RNA sequencing has now made it feasible to study gene expression heterogeneity and to characterise genes based on their coefficient of variation.

Methods
We collected single cell gene expression profiles for 32 human and 39 mouse embryonic stem cells and studied correlation between diverse characteristics such as network connectivity and coefficient of variation (CV) across single cells. We further systematically characterised properties unique to High CV genes.

Results
Highly expressed genes tended to have a low CV and were enriched for cell cycle genes. In contrast, High CV genes were co-expressed with other High CV genes, were enriched for bivalent (H3K4me3 and H3K27me3) marked promoters and showed enrichment for response to DNA damage and DNA repair.

Conclusions
Taken together, this analysis demonstrates the divergent characteristics of genes based on their CV. High CV genes tend to form co-expression clusters and they explain bivalency at least in part.

MOCCS: clarifying DNA-binding motif ambiguity using ChIP-Seq data

Haruka Ozaki[a,*], Wataru Iwasaki[a,b,c,*]

[a]Department of Computational Biology, Graduate School of Frontier Sciences, The University of Tokyo, Kashiwanoha 5-1-5, Kashiwa, 277-8568, Chiba, Japan

[b]Department of Biological Sciences, Graduate School of Science, The University of Tokyo, Hongo 7-3-1, Bunkyo-ku, 113-0032, Tokyo, Japan

[c]Atmosphere and Ocean Research Institute, The University of Tokyo, Kashiwanoha 5-1-5, Kashiwa, 277-8564, Chiba, Japan

[*]Correspondence: harukao.cb@gmail.com (H.O.), iwasaki@bs.s.u-tokyo.ac.jp (W.I.)

Background:

As a key mechanism of gene regulation, transcription factors (TFs) bind to DNA by recognizing specific short sequence patterns that are called DNA-binding motifs. A single TF can accept ambiguity within its DNA-binding motifs, which comprise both canonical (typical) and non-canonical motifs. Clarification of such DNA-binding motif ambiguity is crucial for revealing gene regulatory networks and evaluating mutations in *cis*-regulatory elements. Al- though chromatin immunoprecipitation sequencing (ChIP-seq) now provides abundant data on the genomic sequences to which a given TF binds, existing motif discovery methods are unable to directly answer whether a given TF can bind to a specific DNA-binding motif.

Results:

Here, we report a method for clarifying the DNA-binding motif ambiguity, MOCCS. Given ChIP-Seq data of any TF, MOCCS comprehensively analyzes and describes every *k*-mer to which that TF binds. Analysis of simulated datasets revealed that MOCCS is applicable to various ChIP-Seq datasets, requiring only a few minutes per dataset. Application to the ENCODE ChIP-Seq datasets proved that MOCCS directly evaluates whether a given TF binds to each DNA-binding motif, even if known position weight matrix models do not provide sufficient information on DNA-binding motif ambiguity. Furthermore, users are not required to provide numerous parameters or background genomic sequence models that are typically unavailable. MOCCS is implemented in Perl and R and is freely available via https://github.com/yuifu/moccs.

Conclusions:

By complementing existing motif-discovery software, MOCCS will contribute to the basic under- standing of how the genome controls diverse cellular processes via DNA-protein interactions.

Characterizing mutation-expression network relationships in multiple cancers

Shila Ghazanfar[a,b,*], Yee Hwa Yang[a]
[a]School of Mathematics and Statistics at The University of Sydney, F07, The University of Sydney, NSW, 2006, Australia
[b]Data61, CSIRO, Locked Bag 17, North Ryde NSW 1670, Australia
*Corresponding author: s.ghazanfar@maths.usyd.edu.au (Shila Ghazanfar)

Background:
Data made available through large cancer consortia like The Cancer Genome Atlas make for a rich source of information to be studied across and between cancers. In recent years, network approaches have been applied to such data in uncovering the complex interrelationships between mutational and expression profiles, but lack direct testing for expression changes via mutation. In this pan-cancer study we analyze mutation and gene expression information in an integrative manner by considering the networks generated by testing for differences in expression in direct association with specific mutations. We re- late our findings among the 19 cancers examined to identify commonalities and differences as well as their characteristics.

Results:
Using somatic mutation and gene expression information across 19 cancers, we generated mutation-expression networks per cancer. On evaluation we found that our generated networks were significantly enriched for known cancer-related genes, such as skin cutaneous melanoma ($P<0.01$ using Network of Cancer Genes 4.0). Our framework identified that while different cancers contained commonly mutated genes, there was little concordance between as- sociated gene expression changes among cancers. Comparison between cancers showed a greater overlap of network nodes for cancers with higher overall non- silent mutation load, compared to those with a lower overall non-silent mutation load.

Conclusions:
This study offers a framework that explores network informa- tion through co-analysis of somatic mutations and gene expression profiles. Our pan-cancer application of this approach suggests that while mutations are fre- quently common among cancer types, the impact they have on the surrounding networks via gene expression changes varies. Despite this finding, there are some cancers for which mutation-associated network behaviour appears to be similar: suggesting a potential framework for uncovering related cancers for which simi- lar therapeutic strategies may be applicable. Our framework for understanding relationships among cancers has been integrated into an interactive R Shiny application, PAn Cancer Mutation Expression Networks (PACMEN), containing dynamic and static network visualization of the mutation-expression networks. PACMEN also features tools for further examination of network topology char- acteristics among cancers.

Autumn Algorithm – Computation of Hybridization Networks for Realistic Phylogenetic Trees

Daniel H. Huson[1] and Simone Linz[2]
[1]Center for Bioinformatics (ZBIT), University of Tu¨bingen, Germany
[2]Department of Computer Science at the Univer- sity of Auckland, New Zealand.
Correspondence: daniel.huson@uni-tuebingen.de

Abstract

A minimum hybridization network is a rooted phylogenetic network that displays (or contains) two given rooted phylogenetic trees using a minimum number of reticulations. Much mathematical work has been done on the calculation of such networks, usually making simplifying assumptions on the input trees, such as requiring them to be bifurcating, correctly rooted or that they both contain the same taxa. In biological studies, these assumptions usually do not hold and "realistic" trees have multifurcations (obtained by contracting uncertain edges), are difficult to root and rarely contain exactly the same taxa (due to absent genes or missing data). In this paper, we present a new algorithm for computing minimum hybridization networks for a given pair of "realistic" rooted phylogenetic trees that are not necessarily bifurcating and do not necessarily have exactly the same taxa. We also describe how the algorithm might be used to improve the rooting of the input trees. We introduce the concept of "autumn trees", which provides a nice framework for the formulation of algorithms based on the mathematics of "maximum acyclic agreement forests". While the main computational problem addressed in this paper is hard, the run-time depends mainly on how different the given input trees are. In biological studies, where the input trees are reasonably similar, our parallel implementation performs well in practice. We have implemented the algorithm in our open source program Dendroscope 3, which provides an ideal platform for biologists to explore the use of rooted phylogenetic networks to represent their data. We demonstrate the utility of the algorithm by applying it to several previously studied data sets.

DTL-RnB: Algorithms and Tools for Summarizing the Space of DTL Reconciliations

W. Ma* ,D. Smirnov†, J. Forma*,A.Schweickart*,C. Slocum‡,S. Srinivasan* and R. Libeskind-Hadas*
*Harvey Mudd College, Claremont, California, USA
†Pomona College, Claremont, California, USA
‡California Polytechnic University, Pomona, California, USA

Abstract

Phylogenetic tree reconciliation is an important technique for reconstructing the evolutionary histories of species and genes and other dependent entities. Reconciliation is typically performed in a maximum parsimony framework and the number of optimal reconciliations can grow exponentially with the size of the trees, making it difficult to understand the solution space. This paper demonstrates how a small number of reconciliations can be found that collectively contain the most highly supported events in the solution space. While we show that the formal problem is NP-complete, we give a 1-1/e approximation algorithm, experimental results that indicate its effectiveness, and the new DTL-RnB software tool that uses our algorithms to summarize the space of optimal reconciliations (www.cs.hmc.edu/dtlrnb).

Algorithms for Pedigree Comparison

[1]Zhi-Zhong Chen, [2]Qilong Feng, [3]Chao Shen, [1]Jianxin Wang, [3]Lusheng Wang
[1]Division of Information System Design, Tokyo Denki University, Hatoyama, Saitama, Japan, 350-0394.
[2]School of Information Science and Engineering, Central South University, Changsha, P.R. China, 410083
[3]Department of Computer Science, City University of Hong Kong, Tat Chee Avenue, Kowloon, Hong Kong
E-mail address:
zzchen@mail.dendai.ac.jp
csufeng@csu.edu.cn
chaoshen2-c@my.cityu.edu.hk
jxwang@csu.edu.cn
cswangl@cityu.edu.hk

Abstract

Reconstruction of ancestral relationships among genera, species and populations is a core task in evolutionary biology. At the population level, pedigrees have been commonly used. Reconstruction of pedigree is required in practice due to legal or medical reasons. Pedigrees are very important to geneticists for inferring haplotype segments, recombination, and allele sharing status with which disease loci can be identified. Evaluating reconstruction methods requires comparing the inferred pedigree and the known pedigrees. Moreover, comparison of pedigrees is required in studying relationships among crops such as maize, wheat and barley, etc. In this paper, we discuss three models for comparison of pedigrees, the maximum pedigree isomorphism problem, the maximum paternal-path-preserved mapping problem and the minimum edge-cutting mapping problem. For the maximum pedigree isomorphism problem, we prove that the problem is NP-hard and give a fixed-parameter algorithm for the problem. For the maximum paternal-path-preserved mapping problem, we give a dynamic-programming algorithm to find the mapping that preserves the maximum number of paternal paths between the two input pedigrees. For the minimum edge-cutting mapping problem, we prove that the problem is NP-hard and give a fixed-parameter algorithm with running time $O(n(1+\sqrt{2}))^k$ where n is the number of vertices in the two input pedigrees and k is the number of edges to be cut. This algorithm is useful in practice when comparing two similar pedigrees.

Predicting the Absorption Potential ofChemical Compounds through a DeepLearning Approach

[1]Moonshik Shin, [1]Donjin Jang, [2]Hojung Nam, [1]Kwang Hyung Lee, and [1]Doheon Lee
[1]Department of Bio and Brain Engineering, Korea Advanced Institue of Science and Technology (KAIST), Dajeon, Korea.
[2]School of Information and Communications, Gwangju Institute of Science and Technology (GIST), Gwangju, Korea.
E-mail address:
{msshin.kr, djjang, khlee, dhlee}@kaist.ac.kr hjnam@gist.ac.kr

Abstract

The human colorectal carcinoma cell line (Caco-2) is a commonly used in-vitro test that predicts the absorption potential of orally administered drugs. In-silico prediction methods, based on the Caco-2 assay data, may increase the effectiveness of the high-throughput screening of new drug candidates. However, previously developed in-silico models that predict the Caco-2 cellular permeability of chemical compounds use handcrafted features that may be dataset-specific and induce over-fitting problems. Deep Neural Network (DNN) generates high-level features based on non-linear transformations for raw features, which provides high discriminant power and, therefore, creates a good generalized model. We present a DNN- based binary Caco-2 permeability classifier. Our model was constructed based on 663 chemical compounds with in-vitro Caco-2 apparent permeability data. 209 molecular descriptors are used for generating the high-level features during DNN model generation. Dropout regularization is applied to solve the over-fitting problem and the non-linear activation. The Rectified Linear Unit (ReLU) is adopted to reduce the vanishing gradient problem. Based on the 125 independent test sets, the average classification accuracy of the DNN-based classifier was 73.92%, while LDA and GBT-based classifiers provide the classification accuracy as 65.76% and 69.44%, respectively. The results demonstrate that the high-level features generated by the DNN are more robust than handcrafted features for predicting the cellular permeability of structurally diverse chemical compounds in Caco-2 cell lines.

hc-OTU: A Fast and Accurate Method forClustering Operational Taxonomic Unit based onHomopolymer Compaction

[1,4]Seunghyun Park, [1]Hyun-soo Choi, [1]Byunghan Lee, [2,5]Jongsik Chun, [3]Joong-Ho Won and [1]Sungroh Yoon
[1]Department of Electrical and Computer Engineering, Seoul National University, Seoul 151-744, Republic of Korea
[2]Program in Bioinformatics, Seoul National University, Seoul 151-747, Republic of Korea
[3]Department of Statistics, Seoul National University, Seoul 151-747, Republic of Korea
[4]School of Electrical Engineering, Korea University, Seoul 136- 713, Republic of Korea
[5]School of Biological Sciences, Seoul National University, Seoul 151-742, Republic of Korea
Correspondence author: sryoon@snu.ac.kr

Abstract

In the recent studies of microbial ecology, the distribution of a species into operational taxonomic units (OTUs) is derived by clustering amplicon sequences of 16s rRNA genes, based on environmental samples. Many existing tools for OTU clustering trade off between accuracy and computational efficiency. We propose a novel OTU clustering algorithm, hc-OTU, which achieves high accuracy and fast runtime by exploiting homopolymer compaction and k-mer profiling, to significantly reduce computing time for pairwise distances of amplicon sequences. We compare the proposed method with other widely used methods, including UCLUST, CD-HIT, MOTHUR, ESPRIT, ESPRIT-TREE, and CLUSTOM comprehensively, using nine different experimental datasets and many evaluation metrics, such as normalized mutual information, adjusted rand index, measure of concordance and F-score. Our evaluation reveals that the proposed method achieves comparable accuracy to MOTHUR and ESPRIT-TREE, two widely considered “best” clustering methods, with orders of magnitude speed-up.

Codon Context Optimization inSynthetic Gene Design

[1]Dimitris Papamichail, [2]Hongmei Liu, [3]Vitor Machado, [3]Nathan Gould, [4]J. Robert Coleman,[5] Georgios Papamichail
[1]Department of Computer Science, The College of New Jersey, 2000 Pennington Road, Ewing, NJ 08628-0718, USA.
[2]Division of Biostatistics, Department of Public Health Science, Miller School of Medicine, University of Miami, Miami, FL 33136, USA
[3]Department of Computer Science, The College of New Jersey, 2000 Pennington Road, Ewing, NJ 08628-0718, US
[4]Department of Biology, Farmingdale State Col- lege (SUNY), Farmingdale, NY 11735-1021, USA.
[5]Department of Informatics, New York College, Amalias 38 Ave., Athens, 10558, Greece
E- mail: papamicd@tcnj.edu

Abstract

Advances in de novo synthesis of DNA and computational gene design methods make possible the customization of genes by direct manipulation of features such as codon bias and mRNA secondary structure. Codon context is another feature significantly affecting mRNA translational efficiency, but existing methods and tools for evaluating and designing novel optimized protein coding sequences utilize untested heuristics and do not provide quantifiable guarantees on design quality. In this study we examine statistical properties of codon context measures in an effort to better understand the phenomenon. We analyze the computational complexity of codon context optimization and design exact and efficient heuristic gene recoding algorithms under reasonable constraint models. We also present a web-based tool for evaluating codon context bias in the appropriate context.

3D genome reconstruction with ShRec3D+and Hi-C data

Jiangeng Li, Wei Zhang, Xiaodan Li
Electronic and Control Engineering, and the Beijing Key Laboratory of Computational Intelligence and Intelli- gent System, Beijing University of Technology, Beijing, China.
E-mail address: lijg@bjut.edu.cn zhanwe1bjut@163.com lixiaodan2011@163.com

Abstract

Hi-C technology, a chromosome conformation capture (3C) based method, has been developed to capture genome-wide interactions at a given resolution. The next challenge is to reconstruct 3D structure of genome from the 3C- derived data computationally. Several existing methods have been proposed to obtain a consensus structure or ensemble structures.These method can be categorized as probabilistic models or restraint-based models.In this paper, we propose a method, named ShRec3D+, to infer a consensus 3D structure of a genome from Hi-C data. The method is a two-step algorithm which is based on ChromSDE and ShRec3D methods. First, correct the conversion factor by golden section search for converting interaction frequency data to a distance weighted graph. Second, apply shortest-path algorithm and multi- dimensional scaling (MDS) algorithm to compute the 3D coordinates of a set of genomic loci from the distance graph. We validate ShRec3D+ accuracy on both simulation data and publicly Hi-C data. Our test results indicate that our method successfully correct the parameter with a given resolution, is more accurate than ShRec3D, and is more efficient and robust than ChromSDE.

Algorithmic Mapping and Characterization of the Drug-Induced Phenotypic-Response Space of Parasites Causing Schistosomiasis

[1]Rahul Singh, [2]Rachel Beasley, [3]Thavy Long, and [4]Conor R. Caffrey
[1]Department of Computer Science, San Francisco State University, San Francisco, CA 94132 USA and the Center for Discovery and Innovation in Parasitic Diseases Univer- sity of California San Diego, San Diego, California, USA
[2]Department of Computer Science, San FranciscoState University, San Francisco, CA 94132
[3]Center for Discovery and Innovation in Parasitic Diseases,University of California San Diego, San Diego, California, USA
[4]Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California San Diego, San Diego, California
E-mail: rahul@sfsu.edu

Abstract

Neglected tropical diseases, especially those caused by helminths, constitute some of the most common infections of the world's poorest people. Amongst these, schistosomiasis (bilharzia or 'snail fever'), caused by blood flukes of the genus Schistosoma, ranks second only to malaria in terms of human impact: two hundred million people are infected and close to 800 million are at risk of infection. Drug screening against helminths poses unique challenges: the parasite cannot be cloned and is difficult to target using gene knockouts or RNAi. Consequently, both lead identification and validation involve phenotypic screening, where parasites are exposed to compounds whose effects are determined through the analysis of the ensuing phenotypic responses. The efficacy of leads thus identified derives from one or more or even unknown molecular mechanisms of action. The two most immediate and significant challenges that confront the state-of-the-art in this area are: the development of automated and quantitative phenotypic screening techniques and the mapping and quantitative characterization of the totality of phenotypic responses of the parasite. In this paper we investigate and propose solutions for the latter problem in terms of the following: (1) mathematical formulation and algorithms that allow rigorous representation of the phenotypic response space of the parasite, (2) application of graph-theoretic and network analysis techniques for quantitative modeling and characterization of the phenotypic space, and (3) application of the aforementioned methodology to analyze the phenotypic space of S. mansoni – one of the etiological agents of schistosomiasis, induced by compounds that target its polo-like kinase 1 (PLK 1) gene – a recently validated drug target. In our approach, first, bio-image analysis algorithms are used to quantify the phenotypic responses of different drugs. Next, these responses are linearly mapped into a low- dimensional space using Principle Component Analysis (PCA). The phenotype space is modeled using neighborhood graphs which are used to represent the similarity amongst the phenotypes. These graphs are characterized and explored using network analysis algorithms. We present a number of results related to both the nature of the phenotypic space of the S. mansoni parasite as well as algorithmic issues encountered in constructing and analyzing the phenotypic-response space. In particular, the phenotype distribution of the parasite was found to have a distinct shape and topology. We have also quantitatively characterized the phenotypic space by varying critical model parameters. Finally, these maps of the phenotype space allows visualization and reasoning about complex relationships between putative drugs and their system-wide effects and can serve as a highly efficient paradigm for assimilating and unifying information from phenotypic screens both during lead identification and lead optimization.

Posters

Prathik Naidu. *A Novel Approach to Gene Expression Analysis of Ethnicities.*

Yoshimasa Aoto, Tsuyoshi Hachiya, Kazuhiro Okumura, Sumitaka Hase, Kengo Sato, Yuichi Wakabayashi and Yasubumi Sakakibara. *Digital clustering: An accurate clustering algorithm based on statistical test results.*

Manabu Sugii, Hirotada Mori and Hiroshi Matsuno. *Extension of Genetic Toggle Switch from 2 to 3-variable with Mathematical Formalization.*

Srikiran Chandrasekaran and Karthik Raman. *An Analysis of Stochastic Algorithms for Parameter Estimation in Biological Systems.*

Marghoob Mohiyuddin, John C. Mu, Pegah T. Afshar, Xi Chen, Jian Li, Narges Baniasadi, Mark B. Gerstein, Wing H. Wong and Hugo Y. K. Lam. *A comprehensive genomics resource for assessing variant-calling accuracy (**Sponsor Poster**).*

Min Ye, Gabriela Racz, Qijia Jiang, Xiuwei Zhang and Bernard Moret. *NEMo: An Evolutionary Model with Modularity for PPI Networks.*

Hussein Hejase and Kevin Liu. *A scalability study of computational methods for inferring phylogenetic networks using multi-locus sequence data.*

Yu-Wen Huang, Chih-Min Chang and Jenn-Kang Hwang. *The Effect of Nucleic Acids on Protein Evolution.*

Yu-Feng Lin, Chih-Wen Cheng, Chung-Shiuan Shih, Jenn-Kang Hwang, Chin-Sheng Yu and Chih-Hao Lu. *MIB: Prediction server of Metal Ion Binding sites using fragment transformation method.*

Bayo Lau, John C. Mu, Li Tai Fang, Marghoob Mohiyuddin, Narges Bani Asadi and Hugo Y. K. Lam. *A realistic simulation for benchmarking germline and somatic mutation detection with long read sequencing.*

Caroline Larlee, Alex Brandts and David Sankoff. *Compromise or optimize? The breakpoint anti-median.*

Masayuki Ishitsuka, Tatsuya Akutsu and Jose Nacher. *Determining the optimal critical control set of proteins in large protein networks.*

Kai Wang. *Long-read sequencing and assembly of an Asian genome revealed novel functional genomic elements.*

Chih-Hao Lu, Chin-Sheng Yu, Yu-Feng Lin and Jin-Yi Chen. P*redicting Flavin- and Nicotinamide Adenine Dinucleotide–Binding Sites in Proteins Using the Fragment Transformation Method.*

Zhi Lu. *Improved prediction of RNA secondary structure by integrating the free energy model with restraints derived from experimental probing data.*

Tsun-Tsao Huang, Jenn-Kang Hwang and Chih-Chieh Chen. *Structure Prediction of homodimeric and heterodimeric Protein Complexes.*

Tieming Ji and Jie Chen. *Modeling NGS read count data for CNV study.*

Qian Yang, Xiangzhi Li, Jun Li and Xuegong Zhang. *Ultra Fast Splicing QTL Analysis.*

Sora Yoon, Seonkyu Kim, Sang-Mun Chi, Seon-Young Kim and Dougu Nam. *Improved gene-set enrichment analysis of RNA-seq dat*a.

Miriam Elman and Dongseok Choi. *Discriminative ability of classification methods with high dimensional data.*

Posters

So Young Ryu, Jone Jacobs, Yong-Ming Yu, David Camp, Richard Smith, Ronald Davis, Ronald Tompkins and Wenzhong Xiao. *Measuring in vivo protein synthesis using stable isotope tracer labeling coupled with high-throughput mass spectrometry.*

Xinmin Zhang, Derrick Cheng and Yuerong Zhu. *Cloud-Based Big Data Integration Accelerates Genomics Research.*

Amrit Rau. *A time-staggered combinatorial antibiotic treatment strategy mitigates multi-drug resistance cost in silico.*

Min Yi, Nancy Flournoy and Eloi Kpamegan. *A Novel Method of Bioassay Validation with Heterogeneous Response Variance.*

Hsin-Nan Lin and Wen-Lian Hsu. Kart *An ultra-fast algorithm for NGS read mapping with high error tolerance.*

Byungmin Kim, Saud Alguwaizani, Byungkyu Park and Kyungsook Han. *A method for predicting interactions between virus and human proteins.*

Yuting Ye and Jingyi Jessica Li. *A non-negative matrix factorization based preselection procedure for more accurate isoform discovery from RNA-seq data.*

Preethy Sasidharan Nair, Minna Ahvenainen, Anju Philips, Harri Lähdesmäki and Irma Järvelä. *MicroRNA expression profiling to study the effect of active music listening.*

Tinyi Chu and Charles Danko. *Identification of Active Transcriptional Regulatory Elements using Precision Run-on Sequencing (PRO-seq) at an Unprecedented Resolution.*

Masahito Ohue and Yutaka Akiyama. *A rescoring method of protein-peptide docking prediction with amino acid profiles of rigid-body search results.*

Yuchao Xia, Minghua Deng and Ruibin Xi. *SVmine improves structural variation detection by integrative mining of predictions from multiple algorithms.*

Shih-Chung Yen, Chih-Wen Chen and Jenn-Kang Hwang. *Comprehensive analysis of knotted proteins in the relationship of structural packing density, dynamics and sequence conservation profile*.

Eunyoung Kim, Sangwoo Kim, Suhyun Ha and Hojung Nam. *Developing compound toxicity classification models using physicochemical properties and structure information.*

Carl Angelo Medriano, Jung-Eun Lim, Sun Ha Jee and Youngja Park. *Urine analysis relating poor liver condition and environmental hormones (Bisphenol A and Phthalates): A metabolomic approach.*

Parameswaran Ramachandran, Gareth Palidwor and Theodore Perkins. *BIDCHIPS: Bias Decomposition and Removal from ChIP-seq Data.*

Nikita Alexeev and Max Alekseyev. *Estimation of the True Evolutionary Distance under the Fragile Breakage Model.*

Eric Ho. *Codon Usage Under the Lens of Machine Learning.*

Vijayachitra Modhukur and Jaak Vilo. *Integrated analysis and visualization of DNA methylation , gene expression and copynumber variation in human cancer.*

Yared Kidane. *Computational Prediction of Ionizing Radiation Responsive Genes.*

Adam Johnson and James Birchler. *Chromosome Dosage Effects on the Transcriptome.*