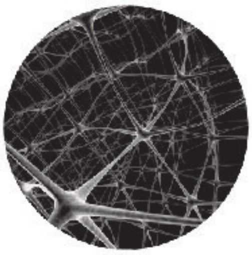
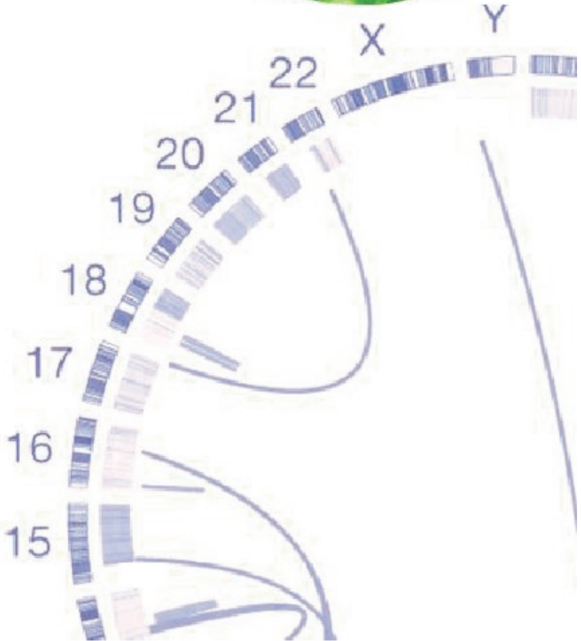
AYRCOB 2012

6th Asian Young Researchers Conference on Computational and Omics Biology





December 20 – 21, 2012

BGI, Shenzhen, China

6th Asian Young Researchers Conference on Computational and Omics Biology

AYRCOB 2012

Final Program and Abstract

BGI,
Shenzhen, China

December 20-21, 2012

Sponsors of the 6th AYRCOB

The University of Tokyo Global COE Program

We establish a center of excellence (COE) for educating graduate students in high-throughput biology, bioinformatics and high-performance computing, so that they will become leaders in the era of the Genome Big Bang.

BGI

BGI was founded in Beijing on Sept 9th, 1999 with the mission of supporting the development of science and technology, building strong research teams, and promoting the development of scientific partnership in genomics field.

Yonsei University

Being the oldest private university in Korea, Yonsei University was first established in 1885 by Christian missionaries. Our mission is to educate leaders who will contribute to humanity in the spirit of “truth and freedom.” The 300,000 Yonsei alumni who take this calling to heart can be found manifesting this proud spirit from leadership positions around the world.

National Chiao Tung University

National Chiao Tung University (NCTU) is a public university located in Hsinchu, Taiwan. Bioinformatics and Systems Biology Institute is the first Institute of Bioinformatics and is ranked among the best in Taiwan, as evidenced by the excellence of the faculty and students as well as the quality of interdisciplinary educational programs.

Committee

General Chairs

Hanhae Kim, Yonsei University, Korea
Yingrui Li, BGI-Shenzhen, China

Advisory Committee

Masanori Arita, The University of Tokyo, Japan
Insuk Lee, Yonsei University, Korea

Budget and Local Committee

Yingrui Li, BGI-Shenzhen, China (Chair)
Gang Chen, BGI-Shenzhen, China
Binghang Liu, BGI-Shenzhen, China
Yoshinori Fukasawa, The University of Tokyo, Japan
Haruka Ozaki, The University of Tokyo, Japan

Public Relation Committee

Binghang Liu, BGI-Shenzhen, China (Chair)
Nai-Wen Chang, National Taiwan University, Taiwan
Haruka Ozaki, The University of Tokyo, Japan
Yingrui Li, BGI-Shenzhen, China
Docyong Kim, KAIST, Korea

Proceedings Committee

Chih-Hung Chou, National Chiao Tung University, Taiwan (Chair)
Chao-Hsuan Ke, National Cheng Kung University, Taiwan
Chun-Hung Su, National Yang-Ming University, Taiwan
Hsin-Jung Li, Princeton University, USA
Anish Man Singh Shrestha, CBRC AIST, Japan
Nai-Wen Chang, National Taiwan University, Taiwan
Chyn Liaw, National Chiao Tung University, Taiwan

Program Committee

Junko Tsuji, The University of Tokyo, Japan (Chair and Coordinator)
Chia-Yi Wu, UCSD, USA
Gang CHEN, BGI-Shenzhen, China
Anish Man Singh Shrestha, CBRC AIST, Japan
Yoshinori Fukasawa, The University of Tokyo, Japan
Hanhae Kim, Yonsei University, Korea
Yukyung Jun, Ewha Womans University, Korea

Review Committee

Woochang Hwang, KAIST, Korea (Chair)
Chun-Hung Su, National Yang-Ming University, Taiwan
Chyn Liaw, National Chiao Tung University, Taiwan
Gang Chen, BGI-Shenzhen, China
Ruibang Luo, BGI-Shenzhen, China
Junho Kim, KAIST, Korea
Tak Lee, Yonsei University, Korea
Ara Cho, Yonsei University, Korea
Rohit Reja, Pennsylvania State University, India
Angela Jean, University of Queensland, Australia
Lawrence Wee, Institute for Infocomm Research, Singapore
Anish Man Singh Shrestha, CBRC AIST, Japan

Contents

Welcome Message 1

Conference Venue 2

Transportation Information 3

General Information 4

Conference Schedule 5

Keynote Speeches & Oral Presentation 10

Poster Abstract 37

Author Index 69

Welcome Message

Welcome to AYRCOB 2012

Dear friends,

On behalf of the international organizing committee, it is a great honor to welcome you to the 6th Asian Young Researchers' Conference on Computational and Omics Biology (AYRCOB). This year, we are pleased to host our conference at BGI, Shenzhen, China with admirable scientists and young researchers.

AYRCOB began as a joint conference between young researchers from Japan and Taiwan in 2007. Since the first AYRCOB, we have extended an invitation to young researchers from across Asia. Thanks to this effort, AYRCOB has expanded to 6 organizing countries (Australia, China, Japan, Korea, Taiwan and Singapore) that have come together to host our meaningful conference here at BGI, Shenzhen, China.

The 6th AYRCOB has been raised to international status thanks to a collaboration with ISCB-Asia/SCCG 2012. The 6th AYRCOB is also proud of having our conference at BGI, an organization that has been at the forefront of research during the current genomics era. In addition, many outstanding abstracts were submitted to our conference. As a result, selection for oral and poster presenters required tremendous effort. This year, we have selected 15 oral presenters from a pool of 65 submissions which spanned 11 countries.

On behalf of the organizing committee, we would like to thank our major sponsors, the Global Center of Excellence (GCOE) Program "Deciphering Biosphere from Genome Big Bang" at the Graduate School of Frontier Sciences, The University of Tokyo, Japan and BGI, Shenzhen, China for their generous support to our conference. We also would like to thank Yonsei University, Korea and National Chiao Tung University, Taiwan for their support.

We hope that our conference will be both beneficial and enjoyable. We also hope that our conference enables you to meet new friends and future collaborators.

Hanhae Kim & Yingrui Li,
General Chairs, 6th AYRCOB
(Junko Tsuji, a Coordinator)

Conference Venue

Venue

Venue at 2nd Building, BGI
- 812 conference hall
- 810, 807 for group discussions

Address:
BGI-Shenzhen, Beishan Industrial Zone,
Yantian District, Shenzhen,
518083, China

TEL: +86-755-25273620

Hotel

Pattaya Hotel

Address:
No.20 Huanmei Road, Dameisha, Yantian
District, Shenzhen 518083, China

Tel: +86-755-25252888

Transportation Information

Shuttle bus

AYRCOB will provide shuttle bus for both days of the conference between Pattaya Hotel and BGI-Shenzhen. The key time and place of the shuttle bus is as following:

Date	From	To	Time
19	Hong Kong airport	Pattaya Hotel	11:00AM; 7:00PM
19	Shenzhen airport	Pattaya Hotel	7:00PM
20	Pattaya Hotel	BGI-Shenzhen	8:00AM
21	Pattaya Hotel	BGI-Shenzhen	8:30AM
22	Pattaya Hotel	Luohu Port	8:30AM

*The shuttle bus would pick up at **exit B** in Hong Kong and Shenzhen airport.

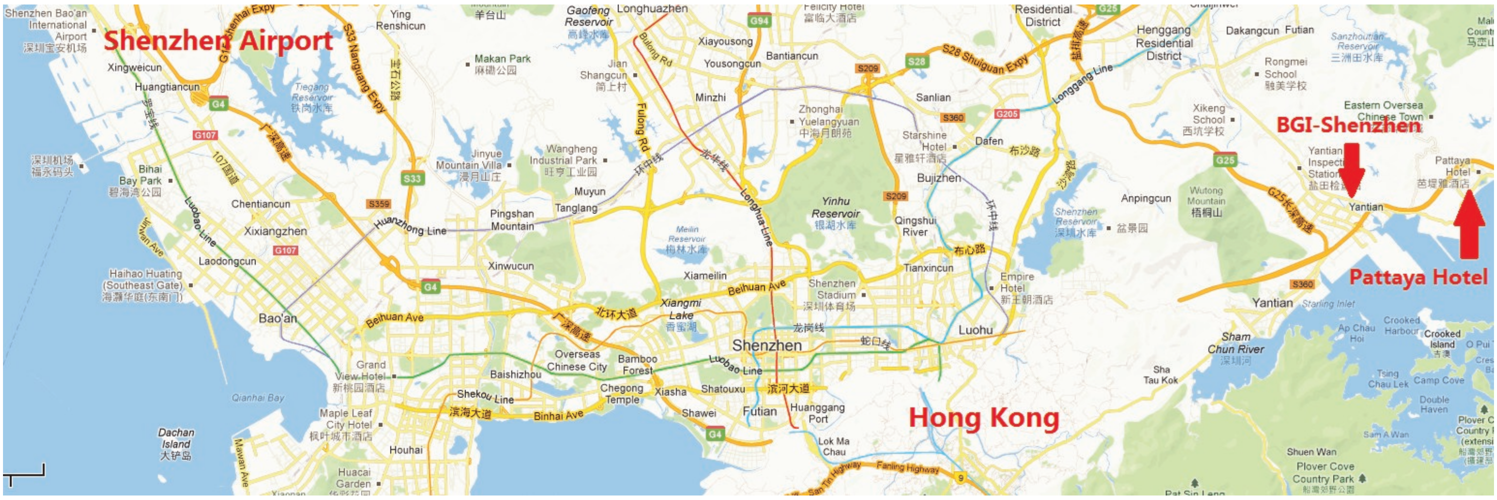
Public transportations

1. From Hong Kong International Airport:

Take Metro from the airport to the Luohu and Huanggang ports, or take a taxi to Sha Tau Kok ports, which is the nearest port to get to the venue. Then you can take a taxi to the venue. Show the relative Conference Notes to the taxi driver. There are five ports in the border of Hongkong and Shenzhen, and the nearest port to our conference venue is Sha Tau Kok port. Please take attention to the opening time of every port.

2. From Shenzhen Bao'an International Airport

Take the No.6 bus from Bao’an International Airport to the Dameisha bus stop (大梅沙公交站) which is very close to the meeting place.



General Information

Free lunch and banquet

- One both days of the conference, AYRCOB will provides lunch to all participants. This will be a good opportunity to meet with the keynote speakers.
- After the first day of the conference, AYRCOB will hold a banquet at Kingkey Palace Hotel, a five-star hotel which is close to Pattaya Hotel.

Adress: No.90 Yankui Road Yantian District, Yantian, 518083 Shenzhen.
Tel: +86-755-82688888.

Poster session

- On the first day (20th Dec), posters should be set up between 08:30-09:00 and removed on the second day (21th Dec). All posters will be set up throughout the entire conference (2 days).
- Oral presenters (IDs starting with O) are not required to make the 2-minute speech; poster presenters (IDs starting with P) are required to make a 2-minute speech each.
- 2-minute speeches for poster presenters will be carried out as follows: odd-numbered posters on the first day (20th Dec) and even-numbered on the second day (21th Dec).
- The poster sessions will run for a total of two days. During the two days, participants are free to ask questions.
- Please make sure to check your ID on p.70-71 of the booklet for the poster board number.

Schedule at a glance

December 20, 2012

8:30 - 9:00	Registration and Poster Set-up
9:00 - 9:20	Opening Remark
9:20 - 10:40	Session 1: High-throughput Genome Sequencing and Chromatin Structure
10:40 - 11:00	Coffee Break
11:00 - 12:00	Session 2: Single Molecule Methods and Systems Biology
12:00 - 13:30	Lunch
13:30 - 14:30	Session 3: Evolution and Genome Variation
14:30 - 14:50	Coffee Break
14:50 - 15:30	Session 4: Protein Biology
15:30 - 16:45	Poster Session (short talk: odd number)
16:45 - 17:45	Live with Today's Speaker
18:00 - 20:00	Banquet

December 21, 2012

9:00 - 10:40	Session 5: Machine Learning and Network Biology
10:40 - 11:00	Coffee Break
11:00 - 12:00	Session 6: Gene Regulation and Small RNAs
12:00 - 13:30	Lunch
13:30 - 14:30	Session 7: Biomedical Informatics
14:30 - 14:50	Coffee Break
14:50 - 16:00	Poster Session (short talk: even number)
16:00 - 17:00	Live with Today's Speaker
17:00 - 17:30	Closing Ceremony

Conference Schedule

December 20, 2012

08:30 AM - 09:00 AM	Registration and poster setup
09:00 AM - 09:20 AM	Opening Remark
09:20 AM - 10:40 AM	**Session 1 - High-throughput Genome Sequencing and Chromatin Structure (Session chair: Junko Tsuji)**
09:20 AM - 10:00 AM	Sequencing, Genomics, and Future of Man ***Huanming Yang, BGI, China (Keynote Speaker)***
10:00 AM - 10:20 AM	A Computational Approach to Estimate and Compare 3D Chromatin Conformation *Hayato Sakata, The University of Tokyo*
10:20 AM - 10:40 AM	Mapping and Aligning PacBio RNA-seq Data *Chao Zeng, Kyoto University*
10:40 AM - 11:00 AM	Coffee Break
11:00 AM - 12:00 PM	**Session 2 - Single Molecule Methods and Systems Biology (Session chair: Allison Wu)**
11:00 AM - 11:40 AM	Quantifying the *E. coli* Proteome and Transcriptome with Single-molecule Sensitivity in Single Cells ***Yuichi Taniguchi, RIKEN, Japan (Keynote Speaker)***
11:40 AM - 12:00 PM	Improving Alkane Tolerance of *Saccharomyces cerevisiae* by Introducing Novel ABC Transporters *Binbin Chen, Nanyang Technological University*
12:00 PM - 01:30 PM	Lunch

December 20, 2012

01:30 PM - 02:30 PM	**Session 3 - Evolution and Genome Variation** **(Session chair: Haruka Ozaki)**
01:30 PM - 01:50 PM	Sex-Specific Adaptation Drives Early Sex Chromosome Evolution in *Drosophila* *Qi Zhou, University of California, Berkeley*
01:50 PM - 02:10 PM	Comprehensive Analysis of Mitochondrial Pseudogenes in Mammalian Genomes *Junko Tsuji, The University of Tokyo*
02:10 PM - 02:30 PM	Discovering Potential Etiology for Schizophrenia with Somatic Deletions *Junho Kim, KAIST*
02:30 PM - 02:50 PM	Coffee Break
02:50 PM - 03:30 PM	**Session 4 - Protein Biology** **(Session chair: Junho Kim)**
02:50 PM - 03:10 PM	SplicePred: a Novel Resource for Identifying RNA Splicing-related Proteins *Kai-Yao Huang, Yuan Ze University*
03:10 PM - 03:30 PM	topPTM: a Database of Post-translational Modifications on Transmembrane Proteins *Min-Gang Su, Yuan Ze University*
03:30 PM - 04:45 PM	Poster Session (short talk: odd number)
04:45 PM - 05:45 PM	Live with Today’s Speaker
06:00 PM - 08:00 PM	Banquet

Conference Schedule

December 21, 2012

09:00 AM - 10:40 AM	**Session 5 - Machine Learning and Network Biology (Session chair: Woochang Hwang)**
09:00 AM - 09:40 AM	Computational Biology through Intra-relation, Inter-relation, and Integration of Diverse Genomic Data ***Hyunjung Shin, Ajou University, Korea (Keynote Speaker)***
09:40 AM - 10:00 AM	Network Assisted Arabidopsis Systems Genetics for Studying Plant Complex Traits *Tak Lee, Yonsei University*
10:00 AM - 10:20 AM	SNONet: Discovery of Protein S-Nitrosylation and Nitric Oxide Signaling Network *Cheng-Tsung Lu, Yuan Ze University*
10:20 AM - 10:40 AM	A Functional Gene Network for *Xanthomonas oryzae* pv. *oryzae* *Hanhae Kim, Yonsei University*
10:40 AM - 11:00 AM	Coffee Break
11:00 AM - 12:00 PM	**Session 6 - Gene Regulation and Small RNAs (Session chair: Chih-Hung Chou)**
11:00 AM - 11:40 AM	Studying Small Silencing RNAs Using High Throughput Sequencing ***Jui-Hung Hung, NCTU, Taiwan (Keynote Speaker)***
11:40 AM - 12:00 PM	Transcribed Pseudogene *ψPPM1K* Generates Endogenous siRNA to Suppress Oncogenic Cell Growth in Hepatocellular Carcinoma *Wen-Ling Chan, National Chiao Tung University*
12:00 PM - 01:30 PM	Lunch

December 21, 2012

01:30 PM - 02:30 PM	**Session 7 - Biomedical Informatics** **(Session chair: Gang Chen)**
01:30 PM - 01:50 PM	Rule-based Multi-scale Modeling for Analyzing Combination Drug Effects *Woochang Hwang, KAIST*
01:50 PM - 02:10 PM	In-silico Prediction of Drug Repositioning Candidates Using an Integrated Network Approach *Haeseung Lee, Ewha Womans University*
02:10 PM - 02:30 PM	Mechanistic Analysis of Prospective Natural Drugs for Checking Alzheimer's Plaque Pathology *Abhinav Grover, Jawaharlal Nehru University*
02:30 PM - 02:50 PM	Coffee Break
02:50 PM - 04:00 PM	Poster Session (short talk: even number)
04:00 PM - 05:00 PM	Live with Today's Speaker
05:00 PM - 05:30 PM	**Closing Ceremony**
05:00 PM - 05:15 PM	Closing Remarks
05:15 PM - 05:30 PM	Awards

Keynote Speeches & Oral Presentation

December 20, 2012

09:20 AM - 10:40 AM	**Session 1 - High-throughput Genome Sequencing and Chromatin Structure (Session chair: Junko Tsuji)**
09:20 AM - 10:00 AM	Sequencing, Genomics, and Future of Man ***Huanming Yang, BGI, China(Keynote Speaker)***
10:00 AM - 10:20 AM	A Computational Approach to Estimate and Compare 3D Chromatin Conformation *Hayato Sakata, The University of Tokyo*
10:20 AM - 10:40 AM	Mapping and Aligning PacBio RNA-seq Data *Chao Zeng, Kyoto University*

#K01 Dec. 20, 2012 09:20 AM - 10:00 AM

Huanming Yang

Chairman and Professor
BGI-Shenzhen, China

Sequencing, Genomics, and Future of Man

Technological breakthrough has proved to be one of the major driving forces for scientific development and social progress. Sequencing technology promises to provide the means to better understand life and to bring the new ways for a better future of man.

New technology means both a new opportunity, if you take it, and a new crisis, if you miss it, leading the later comers even more lagging behind. It is the case for a nation or an institution, as well as for an individual, especially for a young scientist.

Belief and confidence are two important factors for persistence. For example, belief and confidence in “life is of/in sequence” and “life is digital” have been the two pillars of genomics, based on which modern biology/biotech are laid.

Think farther and think wider. Sequencing is not everything, just like “nothing is everything”. Technologically, molecular techniques, stem cell and iPS technology, synthetic genomics, and animal cloning, together with genomic and other expected newly/future emerging technologies, as well as the idea of biobanking, will make the 21st century a BioCentury, and a better future of man.

We are responsible for a better future of man. We are doing science, we are also building culture, a culture of collaboration. The natural disasters repeatedly remind us that we are in the same family on the same boat. If you cannot do it, collaboration can make you able to do it. If you think you can do it, collaboration can make it bigger, better, and cheaper. Only “win-win” collaboration can help us build a better future of man.

#O01 Dec. 20, 2012 10:00 AM - 10:20 AM

A Computational Approach to Estimate and Compare 3D Chromatin Conformation

Hayato Sakata[1], Shinichi Morishita[1]

[1] The University of Tokyo, Computer Biology, Kashiwa, 2778562, Japan

Keywords: chromatin conformation, three dimension, Hi-C

Background and Motivations

Chromatin conformation is thought to have close relationship with gene regulation, development and evolution. Recently, high-throughput chromatin conformation capture methods such as Hi-C have been frequently used to monitor all-to-all DNA contact points in a nucleus. Many novel results have been reported, for example, relationship between oncogene activity and conformation changes, conservation of conformation among species, and correlation between conformation and epigenetics. The major drawback of traditional computational analysis, however, is its limitations in making a full use of DNA contact information, raising the concern that some crucial characteristics of chromatin structures might be overlooked. Although it is becoming popular to visualize DNA contact information in the two-dimensional plane associated with the degrees of individual contact points in heat map, no serious studies have been reported to estimate three- dimensional structure candidates that conform to real contact information collected experimentally. Moreover, to observe some significant differences in the 3D chromatin structures estimated in distinct tissue types, we need to have a method of comparing the 3D structures to uncover subtle structural changes with a high sensitivity. Developing these methods is highly nontrivial, and such methods may suffer from computational inefficiency because of the numerous size of the human genome. It is infeasible to utilize conventional methods for predicting the 3D structures of proteins.

Proposed Approaches

In this study, we approximate a 3D chromatin structure as a Hamiltonian path in a cubic lattice. We then propose an efficient way to produce an unbiased set of numerous independent Hamiltonian paths that are dissimilar to each other, so as to facilitate to select the best Hamiltonian path that is most consistent with contact points, in a reasonable amount of time. For detecting structural changes among more than one 3D chromatin structure, we also present an efficient algorithm that allows us to extract changing sub-structures that would characterize functional differences among multiple 3D structures and to identify similar sub-substructures that might be essential in maintaining common functions across difference tissue types.

Results and Conclusions

Among the features of chromatin conformation, we especially focus on the spatial distribution of epigenetic modification that might have a great impact on the gene regulation. Since epigenetic co-localization is speculated to have a fundamental role in organizing 3D structures, we check this hypothesis by testing the similarity between the model estimated from real Hi-C data and the model predicted from the epigenetic co-localization hypothesis.

#O02 Dec. 20, 2012 10:20 AM - 10:40 AM

Mapping and Aligning PacBio RNA-seq Data

Chao Zeng[1], Hiroaki Iwata[2], Natsuhiro Ichinose[1], Tetsushi Yada[1], Osamu Gotoh[3]

[1] Laboratory of Computational Biology, Kyoto University, Japan
[2] Medical Institute of Bioregulation, Kyushu University, Japan
[3] Computational Biology Research Center, National Institute of Advanced Industrial Science and Technology, Japan

Keywords: alignment, splice junction, mapping, NGS, RNA-seq, PacBio, long reads

Background and Motivations

Techniques of mapping RNA-seq reads onto genomic sequence have been rapidly adopted in studies of novel transcript detection, gene expression profiling, splice variant detection, etc. As the sequencing technology advances, Pacific Biosciences SMRT® platform can generate longer reads, which are expected to suffer from an increased rate of sequencing errors. Moreover, splice junctions will significantly reduce the fraction of continuously mapped reads as the read get longer. However, most currently popular mapping pipelines are designed for short reads with low rates of sequencing error and shows weak support for reads spanning exon-exon junctions. We extended Spaln to map and align PacBio RNA-seq reads against genomic sequence. Our tests on simulated and real data show that the extended Spaln is remarkably efficient for mapping long reads and suitable for detection of novel as well as known splice junctions.

Proposed Approaches

Spaln is a space efficient and fast method for mapping and aligning cDNA/EST sequences onto genomic sequence. In order to find the rough locations of the cDNA/EST sequence, Spaln employs a coarse grained strategy to look for a block in which the transcript may reside, where a block is a predefined segment of the genome with a fixed length. In this way, we can map the cDNA/EST sequences to the reference genome quickly before applying more precise spliced alignment with dynamic programming algorithm. Although the original version of Spaln is quite efficient and sensitive to map transcripts of the EST size or longer, its sensitivity considerably declines for short reads with sequencing errors. To improve the mapping rate, we used overlapping continuous seeds at the block mapping stage. In addition, we examined the performance of spaced seeds and also several combinations of continuous and spaced seeds.

Results and Conclusions

FLT3 sequencing data were downloaded from NCBI SRA050226. We applied Spaln to analyze 19,959 reads with length of 75nt ~ 2438nt, of which 19,893 (99.669%) reads were mapped onto whole human genome (GRCh37.3 build). In total, Spaln predicted 201,077 splicing events, including 196,225 (97.587%) known junctions that are reported in RefSeqGene. The results of our experiment clearly indicate that the extended Spaln is useful to map and align PacBio-like long reads with a substantial number of sequencing errors against reference genomic sequence. Reconstruction of the overall transcriptional scheme including alternative splicing and alternative transcriptional start and stop events from a depth of RNA-seq data is a direction of our future study.

December 20, 2012

11:00 AM - 12:00 PM	**Session 2 - Single Molecule Methods and Systems Biology (Session chair: Allison Wu)**
11:00 AM - 11:40 AM	Quantifying the *E. coli* Proteome and Transcriptome with Single-molecule Sensitivity in Single Cells ***Yuichi Taniguchi, RIKEN, Japan (Keynote Speaker)***
11:40 AM - 12:00 PM	Improving Alkane Tolerance of *Saccharomyces cerevisiae* by Introducing Novel ABC Transporters *Binbin Chen, Nanyang Technological University*

#K02 Dec. 20, 2012 11:00 AM - 11:40 AM

Taniguchi Yuichi

Unit Leader
Laboratory for Single Cell Gene Dynamics
RIKEN Quantitative Biology Center, Japan

Quantifying the *E. coli* Proteome and Transcriptome with Single-molecule Sensitivity in Single Cells

Protein and messenger RNA (mRNA) copy numbers vary from cell to cell in isogenic bacterial populations. However, these molecules often exist in low copy numbers and are difficult to detect in single cells. We carried out quantitative system-wide analyses of protein and mRNA expression in individual cells with single-molecule sensitivity using a newly constructed yellow fluorescent protein fusion library for *Escherichia coli*. We found that almost all protein number distributions can be described by the gamma distribution with two fitting parameters which, at low expression levels, have clear physical interpretations as the transcription rate and protein burst size. At high expression levels, the distributions are dominated by extrinsic noise. We found that a single cell's protein and mRNA copy numbers for any given gene are uncorrelated. In the symposium, our recent challenges on mRNA dynamics will be presented, in addition to the work on single-cell proteome and transcriptome analyses.

#O03 Dec. 20, 2012 11:40 AM - 12:00 PM

Improving Alkane Tolerance of *Saccharomyces cerevisiae* by Introducing Novel ABC Transporters

Binbin Chen[1], Matthew Wook Chang[2]

[1,2] Division of Chemical and Biomolecular Engineering, School of Chemical and Biomedical Engineering, Nanyang Technological University, 62 Nanyang Drive, 637459, Singapore.

Keywords: biofuel, bioinformatics, ABC transporters, S. *cerevisiae*, alkane, tolerance

Background and Motivations

The development of renewable biofuels, such as bio-ethanol, butanol, bio-diesel and jetfuels, helps to address energy security and climate change concerns. Considering the high energy content and compatibility with existing transportation infrastructure, medium chain alkanes draw much attention as the main component of gasoline and jet fuel. Recently, alkane biosynthetic pathway was identified and re-constructed in Escherichia coli and yeast. However, biofuel production is affected by the product toxicity as alkanes are proven to be toxic to microorganisms. Hence, biofuel toxicity is an important issue that needs to be addressed. To overcome biofuel toxicity, harnessing efflux pump is considered as a direct mechanism for reducing biofuel toxicity by expelling toxic compounds from cells. Hence, in this report, we focus on utilizing ATP-binding cassette (ABC) in yeast for reducing biofuel toxicity.

Proposed Approaches

To identify novel biofuel pumps, we used bioinformatics method to generate a list of ABC transporters from sequenced yeast genomes. Based on RT-PCR results, we selected two ABC transporters, ABC-T1 and ABC-T2, from our transporter candidate library for further analysis of alkane transport. To verify the transport capabilities of the transporters toward medium chain alkanes, we cloned out ABC-T1 and ABC-T2 and induced heterogeneous expression in *S. cerevisiae*, where we tested it against medium chain alkanes. The expression of ABC transporters on the plasma membrane was confirmed by western blot and fluorescence microscopy for EGFP-transporter fusion proteins. In addition, toxicity tests were done both on agar plate and in liquid culture. Intracellular alkane accumulation was analyzed with GC-FID after incubation with alkanes. Furthermore, through membrane transporters domain sequence analysis, we mutated glutamate and histidine, our results suggest that glutamate is essential for ABC-T1 and ABC-T2's activity, with ATP is most likely to be hydrolyzed by catalytic carboxylate mechanism.

Results and Conclusions

Thus, in this study, we used bioinformatics method to select potential transporters. RT-PCR was utilized to further screen out target transporters. Transport activity of ABC transporters were investigated by heterologous expression in *S. cerevisiae*. We demonstrated that both ABC-T1 and ABC-T2 are responsible for alkane export and their heterologous expression can significantly improve the tolerance of Baker's yeast against alkanes. Our findings have identified two novel transporters, ABC-T1 and ABC-T2, which can be used in engineering alkane tolerance in microorganisms for yield improvement and high efficient recovery of biofuel production.

December 20, 2012

01:30 PM - 02:30 PM	**Session 3 - Evolution and Genome Variation** **(Session chair: Haruka Ozaki)**
01:30 PM - 01:50 PM	Sex-Specific Adaptation Drives Early Sex Chromosome Evolution in *Drosophila* *Qi Zhou, University of California, Berkeley*
01:50 PM - 02:10 PM	Comprehensive Analysis of Mitochondrial Pseudogenes in Mammalian Genomes *Junko Tsuji, The University of Tokyo*
02:10 PM - 02:30 PM	Discovering Potential Etiology for Schizophrenia with Somatic Deletions *Junho Kim, KAIST*

#O04 Dec. 20, 2012 01:30 PM - 01:50 PM

Sex-Specific Adaptation Drives Early Sex Chromosome Evolution in *Drosophila*

Qi Zhou[1] and Doris Bachtrog[1]

[1]University of California, Berkeley, US

Keywords: sex chromosome, Drosophila miranda, genome evolution

Background and Motivations

X and Y chromosomes of almost all the species don't recombine with each other, and most Y chromosomes are full-degenerated to allow any investigation of their evolutionary history. In a fruit fly species Drosophila miranda, an autosome has fused to the Y chromosome about a million years ago and suppressed recombination with its homologous chromosome in males. Therefore, this Y-linked autosome becomes a 'neo-Y' and its pairing autosome becomes a 'neo-X', and they evolve exactly like ancient sex chromosome pairs and provide unique materials to study sex chromosome evolutionary history.

Proposed Approaches

We sequenced both male and female of Drosophila miranda by next-generation sequencing technology. To assemble the neo-X and neo-Y chromosome in separate, we took advantage of information of male-specific SNPs and a reduction of sequencing coverage at divergent regions between the neo-sex chromosomes. Original reads were divided and assembled into chromosomal sequence and we also produced RNA-seq data from nine different tissues/stages to study the gene expression patterns.

Results and Conclusions

We found within only 1 million years, the neo-Y chromosome has lost about 40% functional genes because of the inhibition of recombination. These genes show various signs of degeneration, including accumulation of nonsense mutations, down-regulation or silence of gene expression. However, we also found the neo-Y becomes 'masculinized' after it acquires male-specific inheritance. A lot of neo-Y genes have evolved biased expression specifically in male-reproductive tissues and those genes undergoing fast sequence evolution are enriched for those with male-beneficial but female-detrimental functions. In the mean time, as old X chromosomes show signs of demasculinization, i.e., a deficiency of male-biased genes and feminization (an enrichment of female-biased genes), male-related genes tend to evolve faster on the neo-X if haploid. These patterns, all together uncovered the important role of sexual antagonistic selection acting on the evolving sex chromosomes of D. miranda and indicated sex chromosomes are dominated by different evolutionary forces at different ages.

#O05 Dec. 20, 2012 01:50 PM - 02:10 PM

Comprehensive Analysis of Mitochondrial Pseudogenes in Mammalian Genomes

Junko Tsuji[1], Martin Frith[2], Kentaro Tomii[1,2], and Paul Horton[1,2]

[1] Department of Computational Biology, Graduate School of Frontier Sciences, University of Tokyo, Japan
[2] Computational Biology Research Center, AIST, Japan

Keywords: NUMTs, mitochondrial pseudogene, retrotransposons, DNA structure, open chromatin

Background and Motivations

Mitochondria possess their own small genomes (around 16,000bp in mammals). NUMTs (Nuclear MiTochondrial sequences) are (partial) copies of the mitochondrial DNA (mtDNA) inserted into the nuclear genome via the repair of DNA double-strand breaks. Several computational studies have investigated NUMTs, however those studies have not used appropriate methodology for sensitive detection of NUMTs and precise delineation of their boundaries.

Proposed Approaches

We developed a carefully considered protocol to redefine NUMT datasets of four mammalian species (human, rhesus, mouse, and rat), and by analyzing the datasets, found new characteristics of NUMT integration sites. The issues we considered include appropriate alignment parameters, correct handling of circular mtDNA, masking of low complexity sequences, post-insertion duplication of NUMTs, long indels and validation of E-value thresholds. Furthermore, to make it easy to excavate NUMT datasets of other species, we packaged this protocol as a software tool.

Results and Conclusions

As the characteristics of NUMT insertion sites, we found four general features related to chromatin and genomic contexts in human, rhesus, mouse, and rat.

(1) We resolved an outstanding controversy in the literature, by confirming that retrotransposons are highly enriched in NUMT flanks. NUMT insertion sites of all four mammalian species show the significant over-representation of retrotransposons (binomial test, $p < 10-4$).

(2) We discovered that NUMT insertion sites show a marked tendency to have high DNA curvature. In all organisms, NUMT flanks in the first 20bp showed significant high predicted DNA curvature.

(3) We found that each species show a significant enrichment of A+T oligomers in the first 10bp of NUMT flanks: TATATA ($p \approx 4.2\times10-4$) in human, ATTATT ($p \approx 7.9\times10-8$) in rhesus, AAACTT ($p \approx 1.1\times10-4$) in mouse, AATTTA ($p \approx 4.4\times10-4$) in rat. Interestingly, the oligonucleotide recognized by L1- endonucleases (TTTTAA) was also significantly enriched ($p \approx 1.4\times10-3$).

(4) We quantified the degree to which NUMT insertion sites correspond to open chromatin regions identified by the DNaseI-seq and FAIRE-seq experimental methods. By cross-checking with the open chromatin data, we found that NUMT insertions correlate with open chromatin regions in most measured cell types (binomial test, p⍰0.05).

More interestingly, we show that the correlation between NUMTs and open chromatin regions drops sharply when examining NUMTs older than the split between human and chimpanzee. This suggests that open chromatin regions where NUMTs insert (primarily intergenic regions) has shifted dramatically in the human line during relatively short evolutionary time scales. In addition to this, we observed that species-specific NUMT insertion sites are enriched in species-specific open chromatin regions. This highlights the possibility of the utility of NUMTs as open chromatin markers.

#O06 Dec. 20, 2012 02:10 PM - 02:30 PM

Discovering Potential Etiology for Schizophrenia with Somatic Deletions

Junho Kim[1], Sanghyeon Kim[2,*], and Doheon Lee[1,*]

[1] Department of Bio and Brain Engineering, KAIST, South Korea
[2] Stanley Brain Research Laboratory, Stanley Medical Research Institute, USA
* Corresponding authors

Keywords: brain mosaicism, Next-generation sequencing, somatic variation, schizophrenia

Background and Motivations

Schizophrenia is a severe and devastating brain disorder believed to be caused by the complex interaction of multiple genetic and environmental factors. Genetic association studies indicate that rare copy number variations (CNV) such as duplications and deletions are implicated in schizophrenia. However, more than 95% of schizophrenia cases cannot be explained by previously identified rare CNVs thus, the etiology of the majority of schizophrenia cases remains to be determined. Due to the unique characteristic of brain tissue that cannot be replicated after neural development, brain mosaicism according to somatic variations has been suspected for the supplementary cause of neuropsychiatric diseases. However, somatic variations in brain cells from an individual have not been well studied mainly due to technical limitations. In this study, we identify somatic deletions that occurred in a fraction of the cells in the brain with schizophrenia and assess their impact using whole genome sequencing (WGS) data.

Proposed Approaches

We developed a computational pipeline to detect somatic deletions that occur only in a fraction of brain cells. Currently no algorithms are available to detect such rare somatic deletions, we utilized several germline deletion calling algorithms including read-depth based analysis, paired-end mapping, and breakpoint mapping to build a convergent approach to identify novel somatic deletion candidates. Chromosomal DNAs from prefrontal cortex, cerebellum, and blood tissue of a female schizophrenic case were sequenced with high-depth WGS (70x), and generated data were applied to our computational pipeline. Sanger sequencing with nested PCR and laser capture microdissection were conducted to validate our somatic deletion candidates. Functional analysis of disrupted genes by somatic deletions was performed to predict affected molecular functions.

Results and Conclusions

We identified 98 somatic deletion candidates in DNA from prefrontal cortex and or cerebellum of an individual with schizophrenia. We validated the breakpoints of three somatic deletions which disrupted coding sequences in PRKRA, BOD1, and CBX3 genes in brain DNA. Somatic deletions disrupted genes that were over-represented in molecular pathways for embryonic organ development, nerve development and transcriptional regulation. Our results suggest that somatic deletions may affect brain development and gene expression regulation in a region specific manner and may be a mechanism underlying the variation that occurs in the normal brain as well as underlying the pathophysiology of many neurological and psychiatric disorders that cannot be explained by current rare CNVs.

December 20, 2012

02:50 PM - 03:30 PM	**Session 4 - Protein Biology** **(Session chair: Junho Kim)**
02:50 PM - 03:10 PM	SplicePred: a Novel Resource for Identifying RNA Splicing-related Proteins *Kai-Yao Huang, Yuan Ze University*
03:10 PM - 03:30 PM	topPTM: a Database of Post-translational Modifications on Transmembrane Proteins *Min-Gang Su, Yuan Ze University*

#O07 Dec. 20, 2012 02:50 PM - 03:10 PM

SplicePred: a Novel Resource for Identifying RNA Splicing-related Proteins

Kai-Yao Huang[1] and Tzong-Yi Lee[1,*]

[1] Department of Computer Science and Engineering, Yuan Ze University, Taoyuan 320, Taiwan

Keywords: RNA splicing, splicing factor, spliceosome, alternative splicing

Background and Motivations

Machinery of pre-mRNA splicing is achieved through the interaction of RNA sequence elements and a variety of RNA splicing-related proteins (e.g. spliceosome and splicing factors). However, the regulation of RNA splicing is not yet fully elucidated, partly because splicing-related proteins (SRPs) have not yet been exhaustively identified and the experimental identification is time-consuming and lab-intensive. Although various computational methods have been proposed for the identification of RNA-binding proteins, there exists no online resource aiming at the identification of SRPs so far. Therefore, we are motivated to design a novel method, named SplicePred, for the identification of RNA splicing-related proteins using experimentally verified spliceosomes and splicing factors.

Proposed Approaches

The experimentally verified splicing-related proteins in humans were collected from published literatures. After the removal of redundant protein entries, it resulted in a total of 283 human splicing factors which are regarded as positive data for feature investigation and model training. Additionally, human proteins which are not included in the positive data were extracted from the UniProtKB and were regarded as the candidate set of non-splicing-related proteins. In order to filter out potential noise data for non-splicing proteins, the remaining proteins consisting of keywords “RNA splicing”, “spliceosome”, or “splicing factors” are removed. The investigation of amino acid composition reveals that there are significant differences between SRPs and non-splicing proteins. Additionally, the information of functional domain is investigated as a feature for classifying splicing factors from non-splicing proteins. A hybrid approach is investigated in this work by combining different sets of feature vectors with the goal of improving splicing factor prediction performance. The public SVM library, LibSVM, is utilized to generate the predictive model with positive and negative training sets, which are encoded with reference to various training features.

Results and Conclusions

The five-fold cross-validation evaluation indicates that the SVM model trained with amino acid composition and functional domains could provide a promising accuracy (82.04%). The result of independent testing demonstrates that SplicePred could effectively differentiate between splicing-related proteins and RNA-binding proteins in mammals and plants. Moreover, the SplicePred could identify the types of SRPs including small nuclear ribonucleoproteins, splicing factors, splicing regulation proteins, and novel spliceosomes. SplicePred is the first online resource for the identification of RNA splicing-related proteins and is now freely accessible via http://csb.cse.yzu.edu.tw/SplicePred/.

#O08 Dec. 20, 2012 03:10 PM - 03:30 PM

topPTM: a Database of Post-translational Modifications on Transmembrane Proteins

Min-Gang Su[1] and Tzong-Yi Lee[1,*]

[1]Department of Computer Science and Engineering, Yuan Ze University, Chung-Li 320, Taiwan

Keywords: post-translational modification (PTM), transmembrane protein, structural topology

Background and Motivations

Transmembrane proteins play crucial roles in various cellular processes. The biological effects of protein modifications on transmembrane proteins include phosphorylation for signal transduction and ion transport, acetylation for structure stability, attachment of fatty acids for membrane anchoring and association, as well as the glycosylation for substrates targeting, cell-cell interactions, and viruses infection. With the importance of PTMs functioning on transmembrane proteins, we are motivated to develop a database, topPTM, that integrates experimentally verified post-translational modifications (PTMs) from available databases and research articles, and annotates the PTM sites on transmembrane proteins with structural topology.

Proposed Approaches

The experimentally verified PTMs are mainly collected from public resources including dbPTM, Phospho.ELM, PhosphoSite, OGlycBase, and UbiProt. Additionally, due to the emerging evidences in nitric oxide (NO)-related pathway, the experimentally verified protein S-nitrosylation sites are manually extracted from approximately 200 S-nitrosylation-related research articles using a text mining approach. For transmembrane proteins, the information of membrane topologies is collected from TMPad, TOPDb, PDBTM, and OPM. After the removal of redundant protein entries, a total of 2234 TM proteins containing experimentally curated annotations of membrane topology remained. In order to fully investigate the PTMs on transmembrane proteins, a candidate set of TM proteins is extracted from UniProtKB by choosing protein entries which contain the keyword “TRANSMEM” in feature (“FT”) line, the localization of “membrane”, and the information of transmembrane topology. The candidate TM proteins are further filtered using HMMTOP and MEMSAT to determine their membrane topologies. The structural topology of transmembrane proteins is represented by graphical visualization, as well as the PTMs. Moreover, the tertiary structure of PTM sites on transmembrane proteins is visualized by Jmol program.

Results and Conclusions

The experimentally verified PTMs are mainly collected from public resources including dbPTM, Phospho.ELM, PhosphoSite, OGlycBase, and UbiProt. Additionally, due to the emerging evidences in nitric oxide (NO)-related pathway, the experimentally verified protein S-nitrosylation sites are manually extracted from approximately 200 S-nitrosylation-related research articles using a text mining approach. For transmembrane proteins, the information of membrane topologies is collected from TMPad, TOPDb, PDBTM, and OPM. After the removal of redundant protein entries, a total of 2234 TM proteins containing experimentally curated annotations of membrane topology remained. In order to fully investigate the PTMs on transmembrane proteins, a candidate set of TM proteins is extracted from UniProtKB by choosing protein entries which contain the keyword “TRANSMEM” in feature (“FT”) line, the localization of “membrane”, and the information of transmembrane topology. The candidate TM proteins are further filtered using HMMTOP and MEMSAT to determine their membrane topologies. The structural topology of transmembrane proteins is represented by graphical visualization, as well as the PTMs. Moreover, the tertiary structure of PTM sites on transmembrane proteins is visualized by Jmol program.

December 21, 2012

09:00 AM - 10:40 AM	**Session 5 - Machine Learning and Network Biology (Session chair: Woochang Hwang)**
09:00 AM - 09:40 AM	Computational Biology through Intra-relation, Inter-relation, and Integration of Diverse Genomic Data ***Hyunjung Shin, Ajou University, Korea (Keynote Speaker)***
09:40 AM - 10:00 AM	Network Assisted Arabidopsis Systems Genetics for Studying Plant Complex Traits *Tak Lee, Yonsei University*
10:00 AM - 10:20 AM	SNONet: Discovery of Protein S-Nitrosylation and Nitric Oxide Signaling Network *Cheng-Tsung Lu, Yuan Ze University*
10:20 AM - 10:40 AM	A Functional Gene Network for *Xanthomonas oryzae* pv. *oryzae* *Hanhae Kim, Yonsei University*

#K03 Dec. 21, 2012 09:00 AM - 09:40 AM

Hyunjung (Helen) Shin

Professor
Department of Industrial Engineering
Ajou University, Korea

Computational Biology through Intra-relation, Inter-relation, and Integration of Diverse Genomic Data

In computational biology, a novel knowledge has been obtained mostly by identifying `intra-relation,' the relation between entities on a specific biological level, and many such researches have been successful. Nowadays, a number of heterogeneous types of data have become more available (i.e., TCGA, The Cancer Genome Atlas) generated from all molecular levels of `omic' dimensions from genome to phenome. Given multi-levels of data, information from a level to another may lead to some hints that we can uncover an unknown biological knowledge. Therefore, it is expected that the next attempt will be more focused on how to utilize the information from `inter-relation,' the relation between different levels: from the genome level to epigenome, transcriptome, proteome, and further stretched to the phenome level. In the meantime, `integration' of different levels of data can aid in extracting knowledge by drawing an integrative conclusion from many pieces of information collected from diverse types of intra-relation or inter-relation. In this talk, the prototypes of the research schemes for intra-relation, inter-relation, and integration will be discussed. The three schemes will be exemplified based on the pilot experimental results on the prediction problem of cancer clinical outcomes using the TCGA data.

#O09 Dec. 21, 2012 09:40 AM - 10:00 AM

Network Assisted *Arabidopsis* Systems Genetics For Studying Plant Complex Traits

Tak Lee[1], Jung Eun Shim[1] and Insuk Lee[1]

[1]Department of Biotechnology, College of Life Science and Biotechnology, Yonsei University, 262 Seongsanno, Seodaemun-Gu, Seoul, 120-749, Korea

Keywords: Genome wide association study, functional gene network, Arabidopsis

Background and Motivations

As next generation sequencing (NGS) technology develops rapidly, Genome Wide Association Study (GWAS) is being highlighted for searching genes that are associated with certain traits such as disease genes in humans and stress resistant genes in plants. GWAS uses high throughput technology and can analyze massive amount of data. By sequencing genomes of organisms and statistically associating sequence variants to certain traits, GWAS is expected to show high performance on the discovery of novel genes. However, even though GWAS has high cost and requires intensive work, it cannot discover as many genes as we expected and shows low performance. Here, we present a novel way of analyzing associations between genetic variants and phenotypes of a plant model organism, *Arabidopsis thaliana*, by using a Network guided approach.

Proposed Approaches

Using the *Arabidopsis* functional gene network (AraNet), we develop a unique algorithm that would effectively predict the significant variant-phenotype associations of *Arabidopsis* GWAS. AraNet is constructed by integrating various omics data and predicts functional relationships for 73% of total *Arabidopsis* genome. So such algorithm that combines GWAS data and integrated omics data of AraNet, would give more power in predicting genes that have low significance in GWAS but still important in certain phenotypes.

In detail, GWAS data with 199accessions, 178384SNPs and 107phenotypes have been analyzed by using a R package EMMA(efficient mixed model association). EMMA corrects population structures which can cause confounding in population studies and it has previously been used in maize and rice studies. To develop the algorithm that combines the Network omics data to GWAS data, a naïve Bayesian approach was used.

Results and Conclusions

With the aid of functional gene network AraNet, we expect improved prediction of significant genetic variants, overcome statistical defects of GWAS and unveil more genetic characteristics of Arabidopsis thaliana by discovering genes truly related to phenotypes.

#O10 Dec. 21, 2012 10:00 AM - 10:20 AM

SNONet: Discovery of Protein S-Nitrosylation and Nitric Oxide Signaling Network

Cheng-Tsung Lu[1] and Tzong-Yi Lee[1,2]

[1] Department of Computer Science and Engineering, Yuan Ze University, Chung-Li 320, Taiwan
[2] Graduate Program in Biomedical Informatics, Yuan Ze University, Chung-Li 320, Taiwan

Keywords: S-nitrosylation; post-translational modification; nitric oxide; signaling network

Background and Motivations

S-Nitrosylation (SNO), a selective and reversible protein post-translational modification (PTM) that involves the covalent attachment of nitric oxide (NO) to the sulfur atom of cysteine, critically regulates protein activity, localization, and stability. Due to its importance in regulating various protein functions and cell signaling, a mass spectrometry-based proteomics method is used to increase the acquisition of of experimentally determined SNO sites. However, given the intricacy of the cellular regulation, using mass spectrometry alone cannot detect the S-nitrosylation-driven signaling network. Furthermore, there is currently no study dedicated to map S-nitrosylation pathway. Thus, we are motivated to develop a new approach that utilizes all available S-nitrosylated proteins to reconstruct the NO signaling networks.

Proposed Approaches

The experimentally verified S-nitrosylation data were manually extracted from research articles via a literature survey. First, all fields in the PubMed database are searched based on the keywords “S-nitrosylation” or “S-nitrosylated” followed by downloading the full text of these research articles. A text-mining system is then developed to survey the full-text literature that potentially describes the site-specific identification of S-nitrosylated sites. Approximately 400 original and review articles associated with protein S-nitrosylation are retrieved from PubMed (July 2012). Next, the full-length articles are manually reviewed for the precise extraction of the S-nitrosylated peptides and the modified cysteines. To determine the locations of S-nitrosylated cysteines within a full-length protein sequence, the experimentally verified S-nitrosylated peptides are then mapped to UniProtKB protein entries based on sequence identity. A total of 3374 S-nitrosylated cysteines are manually determined in 1757 S-nitrosylated proteins. In an attempt to investigate networks among those S-nitrosylated proteins, five protein-protein interaction databases, including DIP, MINT, IntAct, HPRD, and STRING, are integrated. Additionally, the information of metabolic pathways is extracted from KEGG.

Results and Conclusions

This work presents a novel method for investigating the nitric oxide signaling network based experimentally verified S-nitrosylated proteins. Results show that linking protein-protein interaction and metabolic pathway to S-nitrosylation networks would provide a new perspective for cellular signaling networks.

#O11 Dec. 21, 2012 10:20 AM - 10:40 AM

A Functional Gene Network for *Xanthomonas oryzae pv. oryzae*

Hanhae Kim[1], Ofir Bahar[2], Pamela Ronald[2*], Insuk Lee[1*]

[1] Department of Biotechnology, Yonsei University, Korea
[2] Dept. Plant Pathology and the Genome Center, University of California, Davis, USA

Keywords: Functional Network, *Xanthomoans*, plant pathogen

Background and Motivations

Xanthomonas oryzae pv. oryzae (*Xoo*) is a gram-negative bacterial pathogen that causes bacterial blight of rice, a devastating disease in rice worldwide resulting in significant yield losses annually. The rice XA21 pattern recognition receptor confers robust resistance against ***Xoo*** infection. This study aims to construct a predictive functional gene network for *Xoo*, called XooNET and use this tool to investigate the genetic basis underlying *Xoo* infection.

Proposed Approaches

To infer evolutionary co-inheritance among *Xoo* genes, phylogenetic profiles and gene neighborhood were used. Functional linkages were generated using publicly available *Xoo* and *Escherichia coli* (*E. coli*) proteomics data and by measuring Pearson correlation of mRNA co-expression data sets of *Xoo* and *E. coli*. The generated functional linkages of *E. coli* were then transferred to XooNET by an assosialog approach. To produce evidence networks, we benchmarked the generated functional linkages with Gene Ontology based gold standards using bootstrapping. We then subjected the evidence networks through weighted sum scheme to construct XooNET.

Results and Conclusions

A prototype of XooNET covers approximately 70 % the *Xoo* proteome and consists of ca. 57,000 linkages. To improve the coverage and the number of linkages, and because of the current lack of *Xoo* data, we will further analyze *E. coli* omics data sets to generate associalogs. XooNET will be presented during the 6th AYRCOB conference.

December 21, 2012

11:00 AM - 12:00 PM	**Session 6 - Gene Regulation and Small RNAs (Session chair: Chih-Hung Chou)**
11:00 AM - 11:40 AM	Studying Small Silencing RNAs Using High Throughput Sequencing ***Jui-Hung Hung, NCTU, Taiwan (Keynote Speaker)***
11:40 AM - 12:00 PM	Transcribed Pseudogene *ψPPM1K* Generates Endogenous siRNA to Suppress Oncogenic Cell Growth in Hepatocellular Carcinoma *Wen-Ling Chan, National Chiao Tung University*

#K04 Dec. 21, 2012 11:00 AM - 11:40 AM

Jui-Hung Hung

Assistant Professor
Institute of Bioinformatics
National Chiao Tung University, Taiwan

Studying Small Silencing RNAs Using High-throughput Sequencing

High-throughput sequencing technology facilitates the surge of sequencing throughput and the reduction of the costs in a great degree. With the help of such technology and advanced bioinformatics, we now have an unprecedented chance to look at the biogenesis of small silencing RNAs in more details. In this talk, I will cover the current research projects I conducted pertaining to the trimming and tailing mechanism of endo-siRNA and miRNA and how this mechanism plays a role in miRNA and Argonaute sorting in animal cells.

#O12 Dec. 21, 2012 11:40 AM - 12:00 AM

Transcribed Pseudogene *ψPPM1K* Generates Endogenous siRNA to Suppress Oncogenic Cell Growth in Hepatocellular Carcinoma

Wen-Ling Chan[1,2,3], Hsien-Da Huang[1,2*], Jan-Gowth Chang[3*]

[1]Institute of Bioinformatics and Systems Biology, [2]Department of Biological Science and Technology, National Chiao Tung University, Hsin-Chu, Taiwan; [3]Center of RNA Biology and Clinical Application, China Medical University Hospital; *Correspondence author

Keywords: transcribed pseudogene, endogenous siRNA (esiRNA), tumor suppressor

Background and Motivations

RNA interference (RNAi) is a natural cellular process that defends cells against viruses and transposons, and also regulates gene expression in a sequence-specific manner. The transcribed pseudogenes (TPGs) express endogenous siRNAs (esiRNAs) in mice and flies but this remains unclear in humans.

Proposed Approaches

To investigate human TPGs producing esiRNAs that regulate protein-coding genes, we mapped the TPGs to small RNAs (sRNAs) that were supported by public deep sequencing data from various sRNA libraries, and constructed the TPG-derived esiRNA-target interactions. Following, we performed a serial of bio-experiments to verify our predictions.

Results and Conclusions

A genome-wide survey revealed a partial retrotranscript pseudogene *ψPPM1K* containing inverted repeats capable of folding into hairpin structures that can be processed into two esiRNAs; these esiRNAs potentially target many cellular genes, including *NEK8*. Examining 41 paired surgical specimens, we find significantly less expression of the two predicted of *ψPPM1K-specific* esiRNAs, and the cognate-*PPM1K* in the hepatocellular carcinoma (HCC) than in paired non-tumor tissues, whereas the target-*NEK8* expression is increased in the tumor. Additionally, *NEK8* and *PPM1K* were down-regulated in stably transfected *ψPPM1K*-overexpressing cells, but not in cells transfected with an esiRNA1-deletion mutant of *ψPPM1K*. Furthermore, expression of *NEK8* in *ψPPM1K*-transfected cells demonstrated that *NEK8* can counteract the growth inhibitory effects of *ψPPM1K* in HCC cells. These findings indicate that a transcribed pseudogene can exert tumor suppressor activity independent of its parental gene by generation of esiRNAs that regulate human cell growth.

December 21, 2012

01:30 PM - 02:30 PM	**Session 7 - Biomedical Informatics** **(Session chair: Gang Chen)**
01:30 PM - 01:50 PM	Rule-based Multi-scale Modeling for Analyzing Combination Drug Effects *Woochang Hwang, KAIST*
01:50 PM - 02:10 PM	In-silico Prediction of Drug Repositioning Candidates Using an Integrated Network Approach *Haeseung Lee, Ewha Womans University*
02:10 PM - 02:30 PM	Mechanistic Analysis of Prospective Natural Drugs for Checking Alzheimer’s Plaque Pathology *Abhinav Grover, Jawaharlal Nehru University*

#O13 Dec. 21, 2012 01:30 PM - 01:50 PM

Rule-based Multi-scale Modeling for Analyzing Combination Drug Effects

Woochang Hwang[1], Youngdeuk Hwang[2], Sunjae Lee[3], Doheon Lee[4]

[1,2,3,4] Department of Bio and Brain Engineering KAIST, Republic of Korea

Keywords: multi-scale modeling, Whole-body simulation, Rule-based simulation, Combination drug

Background and Motivations

Essential reasons including robustness, redundancy, and crosstalk of biological systems, have been reported to explain the limited efficacy and unexpected side-effects of drugs. Many pharmaceutical laboratories have begun to develop multicompound drugs to remedy this situation, and some of them have shown successful clinical results. Simultaneous application of multiple compounds could increase efficacy as well as reduce side-effects through pharmacodynamics and pharmacokinetic interactions. However, such approach requires overwhelming cost of preclinical experiments and tests as the number of possible combinations of compound dosages increases exponentially. Computer model-based experiments have been emerging as one of the most promising solutions to cope with such complexity. Though there have been many efforts to model specific molecular pathways using qualitative and quantitative formalisms, they suffer from unexpected results caused by distant interactions beyond their localized models.

Proposed Approaches

Here we propose a rule-based whole-body modeling platform. We have tested this platform with Type 2 diabetes (T2D) model, which involves the malfunction of numerous organs such as pancreas, circulation system, liver, and muscle. We made our own rule formalism for capturing multi-level model. The rule can express action subject such as gene, protein, hormone, metabolite etc. It also can express where the subject is in such as cells, organs, tissues etc. We extracted rules by manual curaion from literature and different types of existing models, such as ordinary differential equation model and Petri-net model. We modeled multi-scale system that each level has different response time. We simulated the T2D model by rules. Initial condition of this model was T2D state. Drug injection is the start action of the simulation. We simulated multi drugs effects on the T2D model.

Results and Conclusions

We have extracted T2D-related 117 rules by manual curation from literature and different types of existing models show drug effect pathways of T2D drugs and how combination of drugs could work on the whole-body scale. Metformin, which is one of T2D drugs, works with other drugs as combination drug in 16 cases in our result. We expect that it would provide the insight for identifying effective combination of drugs and its mechanism for the drug development.

#O14 Dec. 21, 2012 01:50 PM - 02:10 PM

In-silico Prediction of Drug Repositioning Candidates Using an Integrated Network Approach

Haeseung Lee[1], Hanna Ryu[1], Sanghyuk Lee[1,2§], Wankyu Kim[1§]

[1] Ewha Research Center for Systems Biology, Division of Life and Pharmaceutical Sciences, Ewha Womans University, Seoul Korea

[2] Korean Bioinformation Center (KOBIC), 52 Eoeun-dong, Yuseong-gu, Daejeon, 305-806, Korea

Keywords: drug repositioning, gene set analysis, microarray

Background and Motivations

With increasing cost of developing novel drugs, drug repositioning (DR) is being actively sought, which is using known drugs for novel indications. Elucidation of new drug-disease links using comparison between biological signatures is efficient way for DR.

Proposed Approaches

Here, we propose an integrative computational approach to identify DR candidates and apply our method to three types of cancer (glioblastoma multiforme, lung and ovarian cancer). Our method is based on the idea that novel drug-disease links can be established by using the similarity of their signatures in terms of genes, biological pathways and chemical structures. An extensive dataset is collected for >40,000 chemicals in total, such as drug targets, chemical signature genes from CMAP, chemical structures as well as gene expression profiles for the target disease. DR candidate chemicals are associated to the target disease by comparing chemical target/signature genes, pathway activity profiles and chemical structures with those for the disease or its 'GOLD STANDARD' drugs. Eight distinct types of drug-disease associations are established as a result of more than 1 billion comparisons in total, which were further integrated by a logistic regression method. Our method consistently shows high accuracy in predicting DR candidates for all the three types of cancers (AUC: 0.88~0.91).

Results and Conclusions

Our integrated approach is shown to be effective in finding chemical-disease associations for targeted drug repositioning.

#O15 Dec. 21, 2012 02:10 PM - 02:30 PM

Mechanistic Analysis of Prospective Natural Drugs for Checking Alzheimer's Plaque Pathology

Jaspreet Kaur Dhanjal[1], Sudhanshu Sharma[1], Heena Dhiman[1], Abhinav Grover[2]

[1] Delhi Technological University, India
[2] School of Life Sciences, Jawaharlal Nehru University, India

Keywords: Alzheimer's, natural drugs, β-Secretase

Background and Motivations

Alzheimer's is a neurodegenerative disorder that results in loss of memory and decline in cognitive abilities. Accumulation of extra cellular β amyloidal plaques is one of the major pathology associated with this disease. β-Secretase or BACE performs the rate limiting step of amyloidal pathway. Inhibition of this enzyme offers a viable prospect to check the growth of these plaques. Numerous efforts have been made in the recent past for generation of BACE inhibitors, however many of them failed during the preclinical and clinical trials owing to drug related/induced toxicity. In the present work, we have used computational methods to screen a large dataset of natural compounds to search for small molecules having BACE inhibitory activity with low toxicity to normal cells and studied their detailed mechanistic behaviors. Preliminary experimental bioactivity assays of these compounds have shown favorable results.

Proposed Approaches

The structure of human BACE was prepared using Schrödinger. A data set consisting of 1,69,109 natural compounds from 10 different suppliers of ZINC database was prepared using LigPrep. Clustering of probe sites in BACE based on their spatial proximity and total interaction energies confirmed the binding cavity. Prepared data-set of natural compounds was then virtually screened against BACE using Glide's HTVS docking. The compounds above threshold of 6 HTVS score were selected and subjected to high-precision Glide's XP protocol for refinement. The top scoring compounds above a cutoff of 11 XP docking score were analyzed for structural and thermal stabilities using MD simulations on Desmond with OPLS all-atom force field 2005.

Results and Conclusions

The individual probe site related closely to the favored high-affinity binding site of BACE, thus validating the ligand binding pocket of the enzyme. Out of several hundred compounds screened using HTVS, six compounds showed significant binding affinity with high-precision Glide XP docking. The two top scoring natural compounds ATAET and DHED were studied for their detailed interactions with BACE. High ligand-efficiency and glide-Emodel scores obtained for these compounds suggested significant binding affinity for BACE. MD simulations carried out to mimic bodily environment and to study the dynamical behavior of binding revealed steady and low-value RMSD trajectories of ligand-bound enzyme complexes, indicating stabilization of these ligands in the BACE functional cavity. Our analysis showed that the first proteolytic cleavage step APP - the rate limiting step of BACE, will be inhibited by binding of ATAET while DHED will restrict flexibility of the flap and modulate the functionality of the enzyme leading to non-formation of the intermediate complex, thus preventing Alzheimer's. Preliminary bioactivity assays of these natural compounds have confirmed computational findings, thus providing experimental evidence to the study.

Poster Abstract

Poster

P01	Risa Kawaguchi	Computational Prediction of miRNA Regulations in Whole Brain of Mouse Using in situ Hybridization Data
P02	Sungmin Kim	Pyrosequencing Analysis of Biodiversity Patterns in Marine Invertebrates Captured by Light-Traps
P03	Moustafa Tarek Gabr	Molecular Docking & Analysis of Peptide Deformylase (PDF) with Hydrazides: Molecular Modeling Study of New Anti-Leptospiral Drugs
P04	Yi-Chun Lai	Identification of Functional Role of Tyrosine Residues in *Clostridium tetani* H+-PPase
P05	Takaho Tsuchiya	Fully Automated Technique for Cellular Signaling Analysis with High Temporal Resolution
P06	Kun Gao	Distinguishing Orthologs and between-species Paralogs in Human-chimp Whole Genome Alignment
P07	Haruka Ozaki	Enumerating DNA-binding Motifs from ChIP-Seq Data
P08	Chih-Hung Chou	Identifying microRNA-target Interactions Using CLIP and PAR-CLIP Sequencing Data
P09	Chih-Wei Chen	Systematically Screen Factors which Affect Motors in *C. elegans*
P10	Satyendra Pratap Singh	Comparative Metagenomic: Microbial Community Analysis of Vermi-compost
P11	Thanet Praneenararat	NaviClusterCS: a Cytoscape Plug-in for Interactively Navigating Large Biological Networks in a Multi-scale Manner
P12	Chyn Liaw	Computational Identification of Important Features for Predicting Antibody Amyloidogenesis
P13	Docyong Kim	Network-based Extraction of Pathological Attributes by Combining Data Classification and Literature Mining
P14	Han-Qin Zheng	Metabolic Pathway Analysis and Gene Discovery for Biofuels Production in *Neodesmus danubialis* (UTEX 2219 -4) by Transcriptome Sequencing Data
P15	Ryota Mori	Essential Ambiguity of RNA Secondary Structure Estimation and a Method to Overcome It

Poster

P16	Allison C.Y. Wu	A Quantitative Study on Gene Expression in *C. elegans* Intestinal Specification Network at Low Temperature Reveals Possible Mechanisms behind Low Penetrance of Mutations and Novel Gene Functions
P17	Tsukasa Fukunaga	An Investigation of Structural Profiles around Target Sites of RNA Binding Proteins
P18	Anish Man Singh Shrestha	An Approximate Bayesian Approach to Mapping Paired-end DNA Reads to a Reference Genome
P19	Junko Tsuji	HeatLogo: a New Sequence Logo Displaying Statistical Significance of Sequence Symbols and Physicochemical Properties
P20	Junko Tsuji	Bisulfighter: Pipeline for Accurate Methylated Cytosine Calling
P21	Yoshinori Fukasawa	MoiraiSP: a Novel Mitochondrial Localization Signal Predictor
P22	Chao-Hsuan Ke	Effective Bio-entities Recognition and Cross-mapping Service to Assist Biocuration Task
P23	Wenlong Jia	SOAPfuse: Software for Identifying Gene Fusions from Paired-end RNA-seq Data
P24	Yu Wen Chen	Long Term Medical Spending in Systemic Lupus Erythematosus and Lupus Nephritis
P25	Sung Moo Byeon	Phylogenetic Analysis of Bifunctional DNase-RNase Domain Containing Proteins
P26	Hsin-Ling Yeh	Applications of Heat Map with Hierarchical Clustering in Mortality Risk - "Hot-Spots" Risk Assessment in Industrial Polluted Area
P27	Bo Wei	GRecScan: An RNA virus Sequence Auto-typing and Inter-subtype Recombination Detection Tool Based on MFE
P28	Tatyana Goldberg	LocTree2 Predicts Localization for all Domains of Life
P29	Ye Yanbo	A New Approach to Identify Possible Core Genes in Baculovirus

#P01 Dec. 20, 2012 03:30 PM - 04:45 PM

Computational Prediction of miRNA Regulations in Whole Brain of Mouse Using in situ Hybridization data

Risa Kawaguchi[1], Hisanori Kiryu[1]

[1] University of Tokyo, Graduate School of Frontier Sciences, Department of Computational Biology, Japan

Keywords: miRNA, regulatory sequence, computational prediction

Background and Motivations

Recent research for microRNA (miRNA) expression has revealed its importance in vertebrate animals especially for gene silencing at the synapse. Many researchers have measured miRNA expressions to investigate such miRNA activity by using microRNA array, RT-PCR, RNA-seq, and so on. It is, however, still impossible to investigate an extent of miRNA regulation directly and easily. Therefore, we tried to predict them by examining relationships between mRNA expressions and an existence of some 7-mer (seed of miRNA) sequences in its 3'UTR using the data from Allen Brain Atlas database.

Proposed Approaches

At first, we calculated probabilities of miRNA repressions using hyper geometric distribution (HGD) for strength of mRNA expression and seed existence using shuffled mRNA 3'UTR sequences as a control. We also applied a FDR control for p-values obtained from HGD. However we found that the result was consequently influenced by seed sequence biases.

To avoid them, it is necessary to account for a frequency of seed appearances in shuffled 3'UTRs. So we calculated each seed frequency compared with shuffled sequences and "wordscore", which means Z-score of mean-adjusted cumulated score of logarithms of seed frequency in order of mRNA expressions. We used Kolmogorov-Smirnov (K-S) test for the Z-score distribution of each seed.

Results and Conclusions

We were able to predict many regulations of miRNA specifically to brain organs by HGD and some of which agreed with previous research, miR-200 family and let-7 family, for example. After applying the false discovery rate (FDR) control method, however, many local regulations were veiled and the influence of seed sequence specific bias was also observed too much. It supposed to be because of differences of seed frequency after 3'UTR shuffling.

Next, we calculated the wordscores and their ranks at each position. The histograms of wordscore rank showed some seeds had high orders of wordscore at many positions and inferred to have relationships with mRNA expressions. The result of K-S test for wordscore also showed the seed group having a lot of high D-value includes miRNA previously reported to predominantly express in mouse brain. In conclusion, our method is useful to predict miRNA or some signatures regulating broadly and also has a potential to predict local regulations combining the method to detect localizing of high-Dvalues.

#P02 Dec. 21, 2012 02:50 PM - 04:00 PM

Pyrosequencing Analysis of Biodiversity Patterns in Marine Invertebrates Captured by Light-Traps

Sungmin Kim[1], Won Kim[1]

[1] School of Biological Sciences, Seoul National University, Seoul 151-747, South Korea

Keywords: Marine invertebrate, biodiversity, pyrosequencing, light-trap

Background and Motivations

Marine invertebrates play an important role for biological monitoring and assessment of changes to biodiversity. The assessment of biodiversity covers all research areas ranging from species identification to ecological monitoring. Several studies have used Sanger sequencing as an effective DNA barcoding technique for the identification of species. However, this is not feasible when dealing with bulk environmental samples. Pyrosequencing technology enables rapid genome sequencing from hundreds of species ranging from bacteria to higher eukaryotes. This technique is less expensive and less time-consuming, and it can generate a large number of DNA amplicons rapidly for discriminatory purposes, when compared with Sanger sequencing.

Proposed Approaches

A biodiversity study of coastal areas was conducted using light trapping which is a useful alternative to benthic invertebrate sampling for monitoring water condition and for estimating the abundance of larval assemblages. Genomic DNA from the bulk environment samples was extracted. The 5′ region of the 18S rDNA gene was amplified by using a universal primer with 454 adaptors and sample-specific key tags. Rapid denoising of pyrosequencing amplicon reads was performed by using sequential filtering flow. In order to obtain a rough estimate of interspecific divergence, 18S rDNA gene fragments were aligned with reliable reference sequences from SILVA database.

Results and Conclusions

Truly barcoded pyrosequencing reads were selected, and their average length was 380 bp. To infer biologically meaningful taxonomic units, we plotted the number of clusters estimated by Usearch uclust with 1%–10% dissimilarity. On the basis of 97% similarity that is determined from the 18S rDNA reference dataset consisting of 4,495 sequences, the number of OTUs was 22, 19, 24, and 19 in the 4 regions, respectively. The present study provides novel insights into the composition, richness, and abundance of marine invertebrates captured by light-traps, regardless of their maturity stages. Such advances might help researchers to better understand and predict the relationship between biodiversity and ecosystem function.

#P03 Dec. 20, 2012 03:30 PM - 04:45 PM

Molecular Docking & Analysis of Peptide Deformylase (PDF) with Hydrazides: Molecular Modeling Study of New Anti-Leptospiral Drugs

Mustafa. T. Gabr[1], S. Madathil[2], Q. Hussain[3], K. R. Madhavi[4], R. Kumar[5], D. Das[6], Asif Naqvi[6]

[1] Mansura University, Elmansura, Egypt.
[2] Centre for Plant Molecular Biology & Biotechnology,Coimbatore India
[3] University of Tampa, Florida, USA.
[4] K L University, Vijayawada, A.P, India.
[5] Singhania University, Rajasthan, India.
[6] Bio Discovery- Solutions for future, Chennai, India

Keywords: Peptide Deformylase, PDF, Hydrazide, Leptospirosis, Lamarckian Genetic Algorithm, Docking.

Background and Motivations

Leptospirosis is a rare and severe bacterial infection that occurs when people are exposed to certain environments. Leptospirosis is caused by exposure to several types of the Leptospira bacteria, which can be found in fresh water that has been contaminated by animal urine. Peptide Deformylase (PDF) is a metalloproteinase and performs a vital role in maturation of protein in bacteria by removing the formyl group from the N-terminal methionine residue of ribosome synthesized polypeptides. Peptide Deformylase (PDF) is essential for normal growth of bacteria for higher organisms and explored as an attractive target for developing novel antibiotics.

Proposed Approaches

In this study we report the binding mode of Peptide Deformylase (PDF) with derivatives of Hydrazides on the basis of structural similarity, substructure, isomers & conformers. Molecular docking approach using Lamarckian Genetic Algorithm was carried out to find out the binding mode of PDF on the basis of calculated ligand-protein pairwise interaction energies. Study was carried out on 3000 molecules which were virtually screened from different databases on the basis of the structural similarity of Tosylhydrazide. The grid maps representing the protein were calculated using auto grid and grid size was set to 60*60*60 points with grid spacing of 0.375 Ǻ. Docking was carried out with standard docking protocol on the basis of a population size of 150 randomly placed individuals; a maximum number of 2.5 *107 energy evaluations, a mutation rate of 0.02, a crossover rate of 0.80 and an elitism value of 1. Fifteen independent docking runs were carried out for each ligand and results were clustered according to the 1.0 Ǻ rmsd criteria.

Results and Conclusions

The docking result of the study of 3000 molecules demonstrated that the binding energies were in the range of -8.64 kcal/mol to -2.56 kcal/mol, with the minimum binding energy of –8.64 kcal/mol. 6 molecules showing hydrogen bonds with the active site residue TYR 71. Further in-vitro and in-vivo study is required on these molecules as the binding mode provided hints for the future design of new derivatives with higher potency and specificity.

#P04 Dec. 21, 2012 02:50 PM - 04:00 PM

Identification of Functional Role of Tyrosine Residues in *Clostridium tetani* H+-PPase

Yi-Chun Lai[1] and Rong-Long Pan[1]

[1] College of Life Science, Institute of Bioinformatics and Structural Biology National Tsing Hua University, Hsin Chu, Taiwan

Keywords: Proton-translocating pyrophosphatase, H+-PPase, CtH+-PPase

Background and Motivations

Proton-translocating pyrophosphatase (H+-PPase, EC 3.6.1.1) is a crucial enzyme which sustains pH homeostasis of organisms. This enzyme generates and maintains the proton gradient across the vacuolar membrane by hydrolyzing the PPi as energy, thus enabling to transport other important ions and metabolites through the biomembrane. Though the research of H+-PPase in plants has been conducted, H+-PPase of Clostridium tetani (CtH+-PPase) was selected as the model for further studies in this thesis. Previous studies indicated that tyrosine residues play an important role in proton translocation and its hydroxyl group in the functional group is able to accept and release protons.

Proposed Approaches

We thus replaced nineteen tyrosine residues in CtH+-PPase individually by alanine with the site-directed mutagenesis technique and analyzed their hydrolysis, proton-translocation and coupling ratio. The enzymatic activities of mutants on Y175, Y226, Y392, Y414 and Y471 were significantly decreased. These five tyrosine residues are highly-conserved in several species. Three dimensional structure suggests they also surround the proton channel of CtH+-PPase. Therefore, we speculated these five positions were involved in the catalytic activity. We then substituted these five tyrosine residues with other amino acids.

Results and Conclusions

The enzymatic activities of Y414S and Y414T were restored to approximately 75% of wild-type so that in this position the hydroxyl group is important to CtH+-PPase. From ion effects study, Y414 was also found to be associated with Na+-binding. In conclusion, the functional role of tyrosine residues in CtH+-PPase was substantially elucidated in our study.

#P05 Dec. 20, 2012 03:30 PM - 04:45 PM

Fully Automated Technique for Cellular Signaling Analysis with High Temporal Resolution

Takaho Tsuchiya[1], Takamasa Kudo[1], Yasunori Komori[1] and Shinya Kuroda[1,2]

[1] Department of Biophysics and Biochemistry, Graduate School of Science, University of Tokyo, Tokyo, Japan

[2] CREST, Japan Science and Technology Agency, Tokyo, Japan

Keywords: cellular signaling, ERK, high temporal resolution

Background and Motivations

Mathematical modeling is an effective way to comprehend the cellular signaling dynamics. However, reliable modeling for cellular response requires quantitative time series data on the basis of experimental observation. A stimulation method with high temporal resolution is required since many signaling molecules show their activation peak within a few minutes after stimulation. A previous stimulation method could not technically produce the desired temporal resolution because it is difficult to stabilize temperature and CO_2 concentration during stimulation. To address these problems, it is crucial to construct the automated stimulation method in a CO_2 incubator with high temporal resolution to obtain high quality data.

Proposed Approaches

A whole experiment consisted of three stages, stimulation experiment, immune-fluorescence assay, and image-processing. About the latter two stages, we have previously developed automated immune -fluorescence assay technique, QIC, which integrates fully automated liquid handling system and high-precision image-processing [1]. For the stimulation experiment, we here report the development of the automated stimulation machine. This machine carries out stimulation experiment in a CO_2 incubator which enables us to obtain less than one minute interval time series data in the stable environment throughout the experiment.

Results and Conclusions

This fully automated experimental method enabled us to measure ERK activation every minute up to 60min in response to nerve growth factor (NGF) in PC12 cells. Obtained time course has high quality. It is previously reported that ERK oscillates in a single cell level in response to many kinds of stimulation in human mammary epithelial cells [2]. Though, our result suggests there is no oscillatory ERK activation in mean population level in response to NGF in PC12 cells. Under this fully automated system, we will further address this issue.

Reference:

[1]Ozaki Y-i, Uda S, Saito TH, et al. (2010) A Quantitative Image Cytometry Technique for Time Series or Population Analyses of Signaling Networks. PLoS ONE 5(4): e9955.

[2] Harish Shankaran, et al (2009) Rapid and sustained nuclear–cytoplasmic ERK oscillations induced by epidermal growth factor. Molecular Systems Biology 5:332.

#P06 Dec. 21, 2012 02:50 PM - 04:00 PM

Distinguishing Orthologs and Between-species Paralogs in Human-chimp Whole Genome Alignment

kun Gao[1] and Junathan Miller[1]

[1] OKinawa Institute of Science and Technology, Japan

Keywords: homology, ortholog, paralog, whole-genome alignment, gene conservation

Background and Motivations

The process of comparative sequence analysis begins with identification of homology within or across species. Homologous regions generally share common ancestry, and can be further distinguished into two classes: orthologs, which are gene pairs diverged via evolutionary speciation, and paralogs, which are gene pairs diverged subsequent to gene duplication. By definition, only genes from different species can be orthologs, whereas paralogs can be defined both within and across species; however, when the species are sufficiently divergent, paralogs among them are generally rare and difficult to distinguish from orthologs. Therefore, so-called "between-species paralogs" are seldom annotated. On the other hand, for closely-related species such as human and chimp, where more than 80% of their genomes are identical, the situation is quite different and between-species paralogs are abundant.

As we have described previously, duplicated sequence lengths in a variety of species, including human and chimp, exhibit power-law length distributions with their associated heavy tails. The lengths of sequences paralogous between human and chimp are comparable to the lengths of sequences orthologous between them, so that it remains a challenge to distinguish orthologs from paralogs between closely-related species like human and chimp.

Proposed Approaches

A standard approach to distinguishing orthologs from paralogs between species is to say that orthologs most are likely to be the homologs between the two species that diverged least. In other words, sequence conservation between orthologs is stronger than sequence conservation between paralogs. We therefore propose the following methods to distinguish between orthologs and paralogs, all based on the BLASTZ (or LASTZ) raw whole-genome alignment.

The first method is the UCSC net alignment, which aims to filter the "best" alignments from the LASTZ raw alignment. Under the assumption that these "best" alignments consist primarily of orthologs, the residual should consist primarily of paralogs.

The second method is to set up an artificial threshold for the LASTZ raw alignment score. High-scoring homologs are classified as orthologs, low-scoring ones as paralogs.

The third method is to geometrically split the dot plot of the LASTZ raw alignment into "proximal-diagonal" and "off-diagonal" regions. Homologs within the "proximal-diagonal region" are classified as orthologs, and "off-diagonal" as paralogs.

The fourth and final method is to inspect the "nesting" of homologous sequences: whether a homologous region is also contained in a longer region of homology in the same genome. Within a simple point mutation model, we anticipate that for the most part, paralogs are "nested" into orthologs. Identifying all the nested overlaps among homologs may enable us to distinguish between paralogs and orthologs.

In all our calculations, we study whole chromosomes of human and chimp, not merely genes.

Results and Conclusions

For human and chimp, the classifications of paralogs obtained by these four different methods agree with one another quite well. The length distributions of the paralogs are algebraic, which we can explain via certain models of segmental duplication. On the other hand, the orthologs exhibit an exponential length distribution, consistent with random uncorrelated base substitution.

#P07 Dec. 20, 2012 03:30 PM - 04:45 PM

Enumerating DNA-binding Motifs from ChIP-Seq Data

Haruka Ozaki[1], Wataru Iwasaki[1,2], Toshihisa Takagi[1]

[1] The university of Tokyo, Japan
[2] Atmosphere and Ocean Research Institute, Japan

Keywords: motif finding, ChIP-seq, next generation sequencing

Background and Motivations

Deciphering DNA binding motifs of transcription factors (TFs) is important to predict their target genes and understand their biological roles. However, our knowledge on those binding motifs is often biased because existing studies typically focused on limited TF-binding events・The recent emergence of ChIP-seq (chromatin immunoprecipitation combined with high-throughput DNA sequencing) has enabled us to detect TF-binding events on a genome-wide scale, making it possible to quantitatively investigate which DNA binding motifs are actually used across the genome.

Many motif-finding methods, including the widely used tool MEME, use position weight matrices or sequence logos to represent found DNA binding motifs. Both representations provide us an abstract view of the motifs by describing the frequency of each base at each position. However, methods using these representations are unsuitable for capturing all significant motifs from tens of thousands of input sequences, as is a usual case in analyzing ChIP-seq data.

Proposed Approaches

Here, we developed an enumerative method that can comprehensively evaluate every k-mer as a DNA binding motif. Previous ChIP-seq studies have revealed that experimentally confirmed DNA binding motifs show clear peaks of frequencies around the TF-binding sites (TFBSs). These observations provided us a good proxy for motif prediction. Our proposed method first converts the raw frequency distributions of every k-mer around TFBSs into cumulative relative frequency curves. Then, it calculates the area under the curve (AUC) score for each k-mer, which we call 'motif-AUC', to quantify whether the distribution have peaks around TFBSs. The motif-AUC scores become bigger when the peaks become clearer, and are used to evaluate if the motif is a true TF-binding motif.

Results and Conclusions

We evaluated the proposed method by applying it to published ChIP-seq data on human cultured cells. In this presentation, we discuss these results and the applicability of the method.

#P08 Dec. 21, 2012 02:50 PM - 04:00 PM

Identifying microRNA-target Interactions Using CLIP and PAR-CLIP Sequencing Data

Chih-Hung Chou[1], Min-Te Chou[1], Jui-Hung Hung[1,2*] and Hsien-Da Huang[1,2*]

[1] Institute of Bioinformatics and Systems Biology, National Chiao Tung University, Hsin-Chu 300, Taiwan

[2] Department of Biological Science and Technology, National Chiao Tung University, Hsin-Chu 300, Taiwan

Keywords: microRNA, CLIP-Seq, PAR-CLIP, miRNA-target interaction

Background and Motivations

MicroRNAs (miRNAs) play a critical role in down-regulating gene expression. By coupling with Argonaute family proteins, miRNAs bind to target sites on mRNAs and employ translational repression. A large amount of miRNA-target interactions (MTIs) have been identified by the crosslinking and immunoprecipitation (CLIP) and the photoactivatable-ribonucleoside-enhanced CLIP (PAR-CLIP) along with the next-generation sequencing (NGS). PAR-CLIP shows high efficiency of RNA co-immunoprecipitation, but it also lead to T to C conversion in miRNA-RNA-protein crosslinking regions. This artificial error obviously reduces the mappability of reads. However, a specific tool to analyze CLIP and PAR-CLIP data that takes T to C conversion into account is still in need.

Proposed Approaches

We herein propose the first CLIP and PAR-CLIP sequencing analysis platform specifically for miRNA target analysis, namely miRTarCLIP. From scratch, it automatically removes adaptor sequences from raw reads, filters low quality reads, reverts C to T, aligns reads to 3'UTRs, scans for read clusters, identifies high confidence miRNA target sites, and provides annotations from external databases. With multi-threading techniques and our novel C to T reversion procedure, miRTarCLIP greatly reduces the running time comparing to conventional approaches. In addition, miRTarCLIP serves with a web-based interface to provide better user experiences in browsing and searching targets of interested miRNAs.

Results and Conclusions

miRTarCLIP not only shows comparable results to that of other existing tools in a much faster speed, but also reveals interesting features among these putative target sites. Specifically, we used miRTarCLIP to disclose that T to C conversion within position 1-7 and that within position 8-14 of miRNA target sites are significantly different (p value = 0.02), and even more significant when focusing on sites targeted by top 102 highly expressed miRNAs only (p value = 0.01). These results comply with previous findings and further suggest that combining miRNA expression and PAR-CLIP data can improve accuracy of the miRNA target prediction. To sum up, we devised a systematic approach for mining miRNA-target sites from CLIP-Seq and PAR- CLIP sequencing data, and integrated the workflow with a graphical web-based browser, which provides a user friendly interface and detailed annotations of MTIs. Our integrated tool can be accessed online freely at http://miRTarCLIP.mbc.nctu.edu.tw.

#P09 Dec. 20, 2012 03:30 PM - 04:45 PM

A Systematic Screen for Factors Regulating Motor Neurons in *C. elegans*

Chih-Wei Chen[1], Oliver Wagner[1]

[1] National Tsing Hua University, Institute of Molecular and Cellular Biology & Department of Life Science, Hsinchu, Taiwan

Keywords: Kinesin, dynein/dynactin, motor, *C. elegans*

Background and Motivations

Many neurodegenerative diseases display accumulation of proteins (cargo) in the nervous system and one hypothesis is that cargo accumulation (for example APP or tau) is based on defective axonal transport and powered by molecular motors. Kinesins and their opposing dynein/dynactin motor complex are the major motor proteins for cargo transport along axonal microtubules in neuronal cells. The frequently observed directional changes of motors and synaptic vesicles on axon in living worms might be explained by cooperative interactions between kinesins and dynein or by tug-of-war model. The bidirectional movement is from unbalance opposite forces producing by kinesin and dynein/dynactin motors. Our hypothesis is that kymograph patterns will change while related factors/proteins affect motors. Therefore, we use real experimental data from neurons of wild type worms to create various tug-of-war patterns and compare patterns from p150/DNC-1 or other mutants.

Proposed Approaches

Our currently results showing p150/DNC-1, adaptor of Dynein, enables physical interact with Unc -104, and not only Yeast two-hybrid but also BiFC (bimolecular fluorescence complementation) assay prove this interaction. We create a mathematical standard model by fitting real kymograph data from motor neuron of wild type *C. elegans*, and then compare with kymograph patterns from p150/DNC-1 mutant strain or404.

Results and Conclusions

We analyze over 400 events from wild type and mutants. Our model displays various tag-of-war patterns by unbalance between kinesin and dynein/dynactin. This model displays the p150/DNC-1 (dynactin) mutant affect Unc-104 (Kinesin). Our goal is using standard model for systematic screen to export factors which may affect motor protein and change kymograph patterns.

#P10 Dec. 21, 2012 02:50 PM - 04:00 PM

Comparative Metagenomic: Microbial Community Analysis of Vermi-compost

Satyendra Pratap Singh[1], Rupali Gupta[2], Manish Kumar[3], Ram Nageena Singh, Prem Lal Kashyap , Sudheer Kumar, Alok Kumar Srivastava and Arun Kumar Sharma

[1] National Bureau of Agriculturally Important Microorganisms, Kusmaur, Uttar Pradesh-275101, India

Keywords: Microbial Community Analysis, Vermi-compost, Organic Matter Decomposition, DGGE.

Background and Motivations

Vermi-Compost is an attractive habitat for microorganisms due to an abundance of nutrients and Relative environmental stability. They decompose organic matter into a stable amendment for improving soil quality and fertility. The Microorganisms that occupy this habitat assist in the uptake to carbon source from vermi compost and decompose organic matter into a stable amendment for improving overall soil quality and fertility. Recent technical approaches in environmental microbiology have enabled the tracing and assessment of these microorganisms using rapid and simple molecular techniques with-out culture dependent bias. Because most of the microorganisms in nature are inaccessible as they are uncultivable in the laboratory and traditional culture-based methods are generally considered to be insufficient to describe microbial communities.

Proposed Approaches

The aim of this study was to apply a group specific PCR system followed by denaturing gradient gel electrophoresis (DGGE) analysis to evaluate the microbial community and the effect of decomposing process on the diversity of microbial populations in vermi-compost ecosystems. Direct DNA extraction from different stages of vermi compost samples was performed. Specific primers were used to amplify 16S rRNA genes and then a semi-nested PCR reaction was applied to obtain smaller fragments for comparing the PCR products by DGGE.

Results and Conclusions

Whether in bulk, different stages of vermi compost samples, the DGGE profiles revealed little change in microbial community. The presence of a few additional bands were observed only in 2nd and 3rd stage samples indicated that a microbial shift occurred with the addition of nutrients and decomposition of organic matter. The combination of semi-nested PCR and DGGE was found to be a rapid and sensitive technique to study the microbial diversity and may be suitable for further studies concerning the role of this bacterial group in large-scale composting for improvement in fertility of soil and cropping system.

#P11 Dec. 20, 2012 03:30 PM - 04:45 PM

NaviClusterCS: a Cytoscape Plug-in for Interactively Navigating Large Biological Networks in a Multi-Scale Manner

Thanet Praneenararat[1,*], Toshihisa Takagi[1,2,3], Wataru Iwasaki[1,4,*]

[1] Department of Computational Biology, The University of Tokyo, Kashiwa, Chiba, 277-8568, Japan

[2] National Bioscience Database Center, Japan Science and Technology Agency, Chiyoda, Tokyo, 102-0081, Japan

[3] Center for Information Biology, National Institute of Genetics, Mishima, Shizuoka, 411-8540, Japan

[4] Current address: Atmosphere and Ocean Research Institute, the University of Tokyo, Kashiwa, Chiba, 277-8564, Japan

[*] These authors contributed equally to this work

Keywords: Network, Interactive Navigation, Cytoscape, Multi-Scale, Big Data

Background and Motivations

The overwhelming amount of biological network data in the big data era is making its visualization cluttered with jumbling nodes and edges, which is well known as “hair-balls”. The visual complication is significantly hindering data analysis and communication of researchers. Effective navigation approaches that can always abstract network data properly and present them insightfully are hence required, to help researchers interpret and analyze the data and acquire knowledge efficiently. Cytoscape is a *de facto* standard platform for network visualization and analysis. Apart from its core sophisticated features, it easily allows for extension of the functionalities by loading extra plug-ins.

Proposed Approaches

We developed NaviClusterCS, which enables researchers to interactively navigate large biological networks of ~100,000 nodes in a multi-scale manner, similar to web mapping services, in the Cytoscape environment. NaviClusterCS rapidly and automatically identifies biologically meaningful clusters in large networks, e.g., proteins sharing similar biological functions in protein-protein interaction networks. Then, it displays only preferable numbers of those clusters at any magnification to avoid cluttered visualization, while its *zooming* and *re-centering* functions still enable researchers to interactively analyze the networks in detail.

Results and Conclusions

Its application to a real *Arabidopsis* co-expression network dataset shows a practical use of the tool for suggesting hidden knowledge in large biological networks, which is not trivial to be obtained using existing tools. NaviClusterCS provides interactive and multi-scale network navigation to a wide range of biologists in the big data era, via the *de facto* standard platform for network visualization. It can be freely downloaded at http://navicluster.cb.k.u-tokyo.ac.jp/cs/ and installed as a plug-in of Cytoscape.

#P12 Dec. 21, 2012 02:50 PM - 04:00 PM

Computational Identification of Important Features for Predicting Antibody Amyloidogenesis

Chyn Liaw [1], Chun-Wei Tung [2], Shinn-Ying Ho[1,3*]

[1] Institute of Bioinformatics and Systems Biology, National Chiao Tung University, Hsinchu 300, Taiwan

[2] School of Pharmacy, College of Pharmacy, Kaohsiung Medical University, Kaohsiung 807, Taiwan

[3] Department of Biological Science and Technology, National Chiao Tung University, Hsinchu 300, Taiwan

Keywords: Antibody Amyloidogenesis; Ubiquitylation Site; Random Forests

Background and Motivations

Antibody amyloidogenesis, the aggregation of soluble proteins into amyloid fibrils, is one of the major problems caused by the humanization process of antibody. Humanized antibodies having a longer half-life can enhance and restore therapeutic effects. However, the humanization process might decrease thermal stability of antibodies that could affect their affinities to targets and lead to amyloid fibril formation.

Proposed Approaches

This study proposes a prediction method based on Random Forests classifier capable of predicting antibody amyloidogenesis across different germline and identify informative features.

Results and Conclusions

Our method performs well with the accuracy of 84.49% using 10-fold cross-validation in across-germline prediction that is impossible for the existing method. Furthermore, the analysis of informative features and prediction of antibody amyloidogenesis can provide better understanding of antibody amyloidogenesis. We found that ubiquitylation might play important roles in determining the antibody amyloidogenesis. An antibody with higher number of putative ubiquitylated lysines tends to be degraded by proteasome that is less likely to be ubiquitylated. In contrast, an antibody with lower number of putative ubiquitylated lysines is less possible for degradation by proteasome that might accumulate to be amyloidogenic.

#P13 Dec. 20, 2012 03:30 PM - 04:45 PM

Network-based Extraction of Pathological Attributes by Combining Data Classification and Literature Mining

Docyong Kim[1*], Kyunghyun Park[1*], Doheon Lee[1]

[1] Department of Bio and Brain engineering, KAIST, Korea
[*]These authors contributed equally to this work

Keywords: Literature mining, Network-based approach, Disease-associated pathological attribute

Background and Motivations

Recently, many researchers have looked for obvious causalities and treatments of diseases. However, diseases are generally influenced by multiple genes and environmental factors. There is a chance to find disease-associated pathological attributes from various biological databases.

Proposed Approaches

In our study, we used the National Health and Nutrition Examination Survey (NHANES) clinical database and the PubMed biomedical literature database to identify the disease- associated pathological attributes by combining data classification and literature mining. This network-based combining approach not only reduces false positive rates compared with literature mining, but also extracts pathological novel attributes. Therefore, in a case study we analyzed, the network inferred the pathological candidate attributes associated with a liver disease.

Results and Conclusions

By the suggested approach, 11 pathological attributes were extracted from a total of 60 pathological attributes from the NHANES clinical database. This result confirmed that four of the extracted pathological attributes were not extracted by other data classification approaches. The identified disease-associated pathological attributes could be used to discover new diagnosis strategies or drug repositioning candidates.

#P14 Dec. 21, 2012 02:50 PM - 04:00 PM

Metabolic Pathway Analysis and Gene Discovery for Biofuels Production in *Neodesmus danubialis* (UTEX 2219-4) by Transcriptome Sequencing Data

Han-Qin Zheng[1], Yi-Fan Chiang-Hsieh[2], Wen-Chi Chang[1,2,*] and Ching-Nen Nathan Chen[3,*]

[1] Institute of Bioinformatics and Biosignal Transduction, National Cheng Kung University, Tainan 701, Taiwan

[2] Institute of Tropical Plant Sciences, National Cheng Kung University, Tainan 701, Taiwan

[3] Institute of Marine Biology, National Sun Yat-sen University, Kaohsiung 804, Taiwan

Keywords: alga, photosynthesis, biofuel, transcriptome, stress

Background and Motivations

Nowadays, energy source is limited and gradually exhausted on earth. Therefore, to solve the shortage problem of energy source, discovery and investigation of renewable energy source become an important and pressing issue. Biomass, a source of renewable energy, included the biofuels which derived from crops. However, crops for biofuels deeply against the forest and poor people's food. Therefore, some scientists try to find the plant which has high fix energy efficiency and little cultivated land request, such as algae. Recently, a new alga strain, UTEX 2219-4, closely related to *Neodesmus danubialis*, was identified to generate large oil body under several stresses such as nitrogen starvation, sorbitol stress and salt stress.

Proposed Approaches

In order to understand the mechanism of large oil body production, next generation 454 sequencing technology was applied to study UTEX 2219-4 transcriptome under four different conditions (normal, nitrogen starvation, sorbitol stress and salt stress). Total 1,255,030 reads produced 22,071 contigs, 20,490 isotigs and 9,339 singletons. Assembled unique sequences were annotated by BLAST similarity search with NCBI non-redundant (NR) and Swiss-Prot Uniref50 protein database. Furthermore, Gene ontology (GO) and Kyoto Encyclopedia of Genes and Genomes(KEGG) orthology (KO) pathway of each annotated sequence are identified.

Results and Conclusions

Expectably, genes involved in fatty acid biosynthesis were up-regulated under various stress conditions. In addition, some genes involved in photosynthesis were up-regulated and they could function to enhance efficiencies of the photosystems. Based on the comparative of gene expression level, the energy flux was mainly from carbon fixation and went through the Glycolysis / Gluconeogenesis to fatty acid biosynthesis, but not from autophagy and endocytosis. These findings help people to understand the mechanism of biofuels production in alga.

#P15 Dec. 20, 2012 03:30 PM - 04:45 PM

Essential Ambiguity of RNA Secondary Structure Estimation and a Method to Overcome It

Ryota Mori[1], Michiaki Hamada[1,2], Kiyoshi Asai[1,2]

[1] Graduate School of Frontier Sciences, the University of Tokyo, Japan
[2] Computational Biology Research Center, National Institute of Advanced Industrial Science and Technology, Japan

Keywords: RNA secondary structure, energy distribution, sub-optimal structure, riboswitch

Background and Motivations

RNA conformation is widely regarded as an issue of importance in its biological function. For several decades, many kinds of tools for predicting RNA secondary structure are proposed by both dry and wet strategy. From the point of view of base-by-base predicting accuracy, recent methods have accomplished quite high performance for RNAs of known structure. All of these methods, however, have common latent problems such as unclear reliability or omission of information on thermal fluctuation and significant sub-optimal structures. These problems are caused by essential ambiguity of current estimation style, which is called “point estimation”.

Proposed Approaches

There are no problems from the beginning if we could calculate the existing probability for all candidate structures and then interpret the distribution but it is not quite realistic because of computational complexity and its high dimensional space. To approach these problems, we require an alternative method for extracting information from exact probability distribution. The basic idea is calculating existing probability for each distance from a certain reference structure by applying a dynamic programming technique. We constructed this algorithm by adopting McCaskill model for calculation of partition function and hamming distance as difference of structures. We can compute this certain-structure-centered probability distribution by $O(n^3)$ time and $O(n^2 d_{max})$ memory (n: RNA length, d_{max}: maximum hamming distance) with maximum parallelization. Additionally, we expand the algorithm into two dimensions for comparing some structural clusters.

Results and Conclusions

We indicate how much ambiguity exists in current estimation methods and then show several outputs of our algorithm. For example, the distribution of a certain riboswitch around its minimum free energy structure implies two distinct structural clusters. This riboswitch is known to change its conformation dynamically in the presence of SAM. Actually, SAM+ and SAM- structure correspond to primary and secondary peaks respectively. These two peaks seem to be separated by quite high potential barrier but there might exist a channel to associate these peaks. Our proposing method yields us profound information on reliability and stability. Applying this algorithm, we can extend a range of analysis such as expression level, thermal stability, and so on.

#P16 Dec. 21, 2012 02:50 PM - 04:00 PM

A Quantitative Study on Gene Expression in *C. elegans* Intestinal Specification Network at Low Temperature Reveals Possible Mechanisms behind Low Penetrance of Mutations

Allison C.Y. Wu[1], Scott A. Rifkin[2]

[1]Bioinformatics and Systems Biology Ph.D Program, UCSD, 9500 Gilman Drive, La Jolla, CA 92093
[2]Division of Biology, UCSD, 9500 Gilman Drive, La Jolla CA92093

Keywords: Quantitative biology, worm development, single-molecule technique

Background and Motivations

The intestinal specification pathway in *C. elegans* is a small genetic regulatory network with redundant genes[2], and previous studies have shown that these can buffer genetic and environmental variation. Mutations in *skn-1* result in greater variation in expression throughout the network including elimination of *med-1/2* and *end-3* expression. This gene expression variability leads to incomplete penetrance in *skn-1* mutants. Because of the lack of genetic buffers from *med-1/2* and *end-3* pathways in these *skn-1* mutant embryos, *end-1* expression is required to reach a threshold level within the 65-120 nuclei stage, to activate *elt-2*[3]. While the regulatory relationships of all the above mentioned genes are well-studied, there is one other gene, *elt-7*, that has only been poorly studied and of which the functional role is still unclear. We intend to put this gene into our study and investigate its function in buffering genetic and environmental variation.

In a previous study by Bowerman, et al.[4], a larger fraction of *skn-1* mutant embryos was found to develop intestinal cells and pharyngeal cells under lower temperature, 15 °C than their counterparts at 25 °C. What causes this lower penetrance at low temperature?

Proposed Approaches

We propose to use single-molecule fluorescence *in situ* hybridization (smFISH) to look into two possible hypotheses for this phenomenon: first, longer cell-cycles might give more time for *end-1* to accumulate to the threshold level. In this case, no change of dose-response curve will be involved. Second, the threshold for *elt-2* induction might shift to a lower level and the decision window for *end-1* might also change. By using single-molecule fluorescence *in situ* hybridization (smFISH), we are able to measure the mRNA transcript number at single-molecule level in each individual without any transgenic methods.

Results and Conclusions

With this technique, we investigate the gene expression of *med-1/2*, *end-1*, *end-3*, *elt-2*, and *elt-7* at two temperatures, 25 °C and 15 °C in wildtype *C. elegans*. Our data revealed a possible change of threshold on *end-1* to *elt-2* at low temperature and a possible regulatory relationship among *elt-7* and *end-1*, *end-3* in this *skn-1* intestinal specification network.

#P17 Dec. 20, 2012 03:30 PM - 04:45 PM

An Investigation of Structural Profiles around Target Sites of RNA Binding Proteins

Tsukasa Fukunaga[1], Hisanori Kiryu[1]

[1] The University of Tokyo, Japan

Keywords: RNA binding proteins, RNA secondary structure, CLIP-seq

Background and Motivations

RNA binding proteins (RBPs) play an integral role in post-transcriptional regulation by binding to target transcripts. Therefore, an understanding of RBPs and their target sites lead us to understanding gene regulatory networks. Recent improvement of experimental technologies including CLIP-seq enables us to identify target sites of RBPs comprehensively. Thus, we can discover the sequential motif of the CLIP-ed RBP. However, RBPs recognize not only specific sequential motifs but also RNA secondary structure in their target sites. Despite the importance of the RNA secondary structure in target sites of RBPs, there have been few studies investigating RNA secondary structures in target sites. When an RNA sequence forms a secondary structure, each nucleotide in the sequence takes any of six different loops, bulge loop, hairpin loop, internal loop, multi loop, outer loop and stem. Here, we define that the structural profile of nucleotide i in an RNA sequence are probabilities that i takes each six loop type. Although we need structural profiles in order to detect binding to specific loop type, none of the existing programs can exactly compute the structural profiles.

Proposed Approaches

Therefore, we developed algorithms for computing the structural profiles exactly using a dynamic programming method, and implemented in software called 'CapR'. We computed the structural profiles based on the Rfold model. The Rfold model is an unambiguous grammar that generates all the secondary structures excluding pseudoknots without redundancy. This model is also able to be directly applied to the energy model and we apply Turner Energy model. To apply the algorithm to long sequences, we restricted the maximal span of the base pairs to a fixed value W. The computational complexities of our algorithm are O(NW^2).

Results and Conclusions

First, we analyzed the basic properties of structural profiles. We found that maximal span and GC content have an impact on structural profiles. Second, we evaluated accuracy of structural profiles calculated by CapR. We showed that our method accurately infers loop types. Third, we investigated structural profiles of target sites of RBPs determined by RIP-Chip and CLIP-seq. We verified that yeast protein Vts1p preferentially bind hairpin loops in vivo. Also, structural profiles are more useful than accessibilities in order to distinguish the false binding site from the true binding site. Last, we investigate the structurally most important positions around binding sites of RBPs. While neighboring binding site is important for the QKI protein, 5’-end of their binding site is important for the Nova protein. These results may indicate the kinetic aspects of the binding mechanism.

#P18 Dec. 21, 2012 02:50 PM - 04:00 PM

An Approximate Bayesian Approach to Mapping Paired-End DNA Reads to a Reference Genome

Anish Man Singh Shrestha[1] and Martin C. Frith[1]

[1] Computational Biology Research Center, AIST, Japan

Keywords: high-throughput sequencing, paired-end DNA reads, Last

Background and Motivations

Many high-throughput sequencers provide a paired-end option, in which each of the two opposite strands of a DNA fragment is read from the edge to the interior in the 5' to 3' direction, generating a pair of reads. Paired-end reads can be obtained by a simple modification to the standard single-end workflow, yet they provide several benefits over single-end reads. They contain extra positional information that aids in accurate mapping of reads to a reference, for instance by disambiguating alignments when one of the ends aligns to a repetitive region. They are also extremely useful in downstream analyses of structural variations such as detection of indels or rearrangements. In this work, we focus on the former: the task of mapping a set of paired-end reads to a reference genome, which is often the first and fundamental step in inferring biological phenomena from high-throughput sequencing data.

Proposed Approaches

We present a new probabilistic framework to predict the alignment of paired-end reads to a reference genome. Using simulated as well as real data, we compare the performance of our method against six other read-mapping tools that provide a paired-end option.

Results and Conclusions

We show that our method provides a good combination of sensitivity, error rate, and computation time, especially in more challenging and practical cases such as when the reference genome is incomplete or unavailable for the sample, or when there are large variations between the reference genome and the source of the reads. An open-source implementation of our method is available as part of Last, a multi-purpose alignment program freely available at http://last.cbrc.jp.

#P19 Dec. 20, 2012 03:30 PM - 04:45 PM

HeatLogo: a New Sequence Logo Displaying Statistical Significance of Sequence Symbols and Physicochemical Properties

Junko Tsuji[1], Szu-Chin Fu[2], and Paul Horton[1,2]

[1] Department of Computational Biology, Graduate School of Frontier Sciences, University of Tokyo, Japan

[2] Computational Biology Research Center, AIST, Japan

Keywords: sequence logo, heat map, protein disorder region, protein secondary structure

Background and Motivations

Sequence logos are simple graphical representation of conserved sequence patterns in multiple alignments. Although there are several logo generation tools which collaterally evaluate statistical significance of each symbol's occurrence rate, no tool is available for visualizing sequence logos in a color-coded heat map representation according to statistical significance of over/under-represented symbols at specific positions.

Proposed Approaches

In this study, we materialized the solution of this issue as HeatLogo by extending the WebLogo3 package, and designed the tool for dealing with all possible sequence types including codons. HeatLogo displays sequence conservations and the statistical significance of over/under-represented symbols at once. Furthermore, especially for protein sequences, HeatLogo provides an interesting function; to analyze disorder and secondary structure information of protein multiple sequence alignment, HeatLogo calculates such physicochemical properties with third-party predictors and visualizes the calculated results as multi-track sequence logos color-coded according to p-value of each column.

Results and Conclusions

We tested the utility of HeatLogo using leucine-rich nuclear export signal (NES) sequences and codon sequences of the signal peptide in maltose-binding protein (malE) as examples, and we confirmed that HeatLogo visually captures important features of the two example datasets. HeatLogo is the first tool for observing the conservation pattern of DNA, RNA, codons, proteins, protein physicochemical properties, and their statistical significance of each position at a time.

#P20 Dec. 21, 2012 02:50 PM - 04:00 PM

Bisulfighter: Ppeline for Accurate Methylated Cytosine Calling

Junko Tsuji[1], Paul Horton[1,2], Toutai Mitsuyama[2]

[1] Department of Computational Biology, Graduate School of Frontier Sciences, University of Tokyo, Japan

[2] Computational Biology Research Center, AIST, Japan

Keywords: bisulfite sequencing, DNA methylation, differentially methylated region

Background and Motivations

Methylated cytosines (mCs) affect many biological processes such as gene expression, silencing or genomic imprinting. A combination method of bisulfite-treated DNA and high throughput sequencing, known as bisulfite-seq, is widely applied to capture a snapshot of epigenomic state of cells.

To find mC regions, bisulfite-converted reads are mapped to a reference genome, and mC levels are estimated with the mapped reads. Since the mC calling procedure highly depends on the alignment correctness of mapped reads, those computational tasks still have had room for improvement. In response to this, with the local alignment software, LAST, we developed a new mC estimation pipeline, Bisulfighter. This pipeline calls mC levels more accurately than other published tools.

Proposed Approaches

Bisulfighter first converts all cytosines of bisulfite-converted reads to thymines and maps those reads to the reference genome. After restoring converted cytosines of mapped reads, mC ratios are computed by considering quality scores of mapped reads and alignment probabilities which measure the reliability of each aligned column.

Results and Conclusions

Using simulated datasets of chromosome X in human, Bisulfighter succeeded to exhibit the best performance, exceeding other mC callers such as Bismark, Brat, Bsmap, etc. We will show the improved performance of our pipeline in all mC contexts (CpG, CHG, CHH; H = A, T, C) calls.

#P21 Dec. 20, 2012 03:30 PM - 04:45 PM

MoiraiSP: a Novel Mitochondrial Localization Signal Predictor

Yoshinori Fukasawa[1,2], Kenichiro Imai[2,3], Szu-Chin Fu[3], Junko Tsuji[1], Noriyuki Sakiyama[3] and Paul Horton[1,3]

[1] Dept. of Comp. Biol., the University of Tokyo, Japan

[2] JSPS Research Fellow, Japan

[3] CBRC, Institute of Advanced Industrial Science and technology, Japan

Keywords: Sequence analysis, mitochondria, signal prediction and cleavage

Background and Motivations

1000-1500 different proteins are estimated to localize in mitochondria, however numerous mitochondrial proteins remain undiscovered. Prediction of mitochondrial targeting signal is an efficient approach when identifying undiscovered mitochondrial proteins. A cleavable N-terminal presequence is the best characterized mitochondrial targeting signal: it is said that about half of known mitochondrial proteins possess the presequence. Mitochondrial proteins with presequence are imported into mitochondria via the translocase and then the presequence is cleaved off by mitochondrial processing protease (MPP) in the matrix. However, the detail mechanisms remain unclear. Moreover, the data of experimentally identified presequences was limited. Thus, current predictors cannot produce sufficient performances in presequence and cleavage site prediction. Fortunately, large scale proteomic analyses of presequence were recently reported in yeast and plant. This proteomic data gave us an opportunity to improve the signal prediction.

Proposed Approaches

In this work, we therefore developed a predictor for presequences and their cleavage sites trained on recent proteomic data as well as annotated sequences extracted from Swiss-Prot. We trained an SVM classifier for this task and applied several features to predict presequence such as amino acid composition, physico-chemical properties and import receptor recognition motif. We furthermore performed novel motif search and generated profiles for cleavage sites of MPP. These novel characteristic features were integrated into our predictor as features of presequence.

Results and Conclusions

Our predictor attains better performances than the present predictors: 0.70 of Predotar, 0.65 of TargetP and 0.77 of our predictor in Matthew's correlation coefficient. In addition, we achieved a significant performance improvement in cleavage site prediction. Prediction of cleavage site shows better performance with comparing TargetP: our predictor predicts 71% of canonical cleavage sites and TargetP does about 54% within predicted as presequence containing proteins. The results indicate that, having the advantage of a large training dataset for cleavage site, our predictor makes more accurate predictions than previous methods. Thus, our method is valuable for finding candidates of undiscovered mitochondria proteins and their signal regions.

#P22 Dec. 21, 2012 02:50 PM - 04:00 PM

Effective Bio-entities Recognition and Cross -mapping Service to Assist Biocuration Task

Chao-Hsuan Ke, Jung-Hsien Chiang

Department of Computer Science and Information Engineering, National Cheng Kung University, Tainan, Taiwan

Keywords: bio-entities, biocuration, text mining, annotation

Background and Motivations

To collect and organize biological information for researchers has become an essential process of biomedical research; therefore many literature biocuration and annotation efforts have been carried out. Typically, curators read scientific articles and transform it into easy-access records. However, this procedure is time-consuming and data cannot regularly be updated. Now majority of curators consider developing text annotation tools probably assist in carrying out the main biocuration tasks. The first step in biocuration is bio-entities recognition and normalization, different format of bio-entities are presented in the literature greatly affects how fast biocurators can identify and curate them. In other side, the bio-entities unambiguously identified and its identifier (ID) establishes links between various biological databases is another crucial issue. Therefore an effective method to integrate bio-entities with scientific publication and directly mapping existing biological databases is necessary; it can obviate the need for curators to search in multiple locations for information relating to a specific item of interest.

Proposed Approaches

In this study, we propose a hybrid method to identify bio-entities from text and then establish identifier mapping links between biological databases. The method contains two phase. In the first phase, we used different machine learning-based recognizer and lexicon respectively to identify five bio-entities that are gene, protein, chemical, drug and disease and contain specific fields (official name, symbol name, synonyms, database cross-reference links, species name, entry ID). In the second phase, matching identified bio-entities to biological identifiers and linking to their database of origin.

Results and Conclusions

Herein, we randomly select 100 full text articles which all contained gene function from BioCreAtive III (IAT task) corpus as experiential material. We want to know whether superior bio-entities identification can assist curators. First, all testing articles were annotated by two human assessors experienced in the field of bioinformatics. Afterwards, three Ph.D students who study in bioinformatics read non-labeled testing articles then carry out the curation task (i.e. to label gene function). In addition, another three Ph.D students used an open-access gene annotation software named MyMiner, and carry out the same curation task. The result shows first three curators labeled gene function can save double time than non-identified data, and save 10% time than other curators who used MyMiner. Hence, we believe upholds information integrity, facilitates the proper linkage of data to other resources and support effective mining of data from scientific papers, and would accelerate literature curation.

#P23 Dec. 20, 2012 03:30 PM - 04:45 PM

SOAPfuse: Software for Identifying Gene Fusions from Paired-end RNA-seq data

Wenlong Jia[1,2], Kunlong Qiu[1] and Guangwu Guo[1,2]

[1] BGI-Shenzhen, Shenzhen, China
[2] BGI-Americas, Cambridge, USA

Keywords: bioinformatics, gene fusion, RNA-Seq

Background and Motivations

Gene fusions play important roles in carcinogenesis and can serve as valuable diagnostic and therapeutic targets in cancer. Recently, many computational methods have been developed to identify fusion candidates by analyzing RNA-seq data. Although the methods were capable of detecting genuine gene fusions, many challenges and limitations remain.

Proposed Approaches

We have developed open-source software, SOAPfuse, which detects gene fusions with single base resolution from paired-end RNA-seq data. SOAPfuse applies several pioneering algorithms to overcome deficiencies of other wildly used fusion-detecting software. In our methods, the most important is the local exhaustion algorithm for high efficient construction of putative junction library. Moreover, combining an iteratively trimming-realigning procedure and several strong filtrations, SOAPfuse achieves good detectability and high accuracy.

Results and Conclusions

We compared SOAPfuse with other tools, including chimerascan, deFuse, FusionHunter and TopHat-Fusion, based on both actual and simulated datasets. Actual data are from two published studies, which have several validated fusions. SOAPfuse consumed the least computing resources (cpu-time and memory) to redetect the most validated fusions. We simulated 150 fusions, and generated RNA-Seq data at nine depth levels (5-200X). SOAPfuse shows the lowest FN and FP rates at all depths. We also applied SOAPfuse to RNA-seq data from two bladder cancer cell lines, and identified 16 fusions, in which 15 (93%) are validated successfully. We reanalyzed this dataset using deFuse, which detected only 9 out of these 15 validated fusions. We also found 6 pairs of recurrent fusions, some of which show strong signals inferring their structure variation sources. Interestingly, these recurrent fusions are not reported by previous studies on bladder cancer. To our knowledge, SOAPfuse is the only one that uses local exhaustion algorithm. It runs fast and saves computation resources, quite suitable for analyzing vast samples in parallel. SOAPfuse not only identifies genuine fusions effectively but also provides practicable means to explore the potential SVs that transcribe to fusions. Our work gives a novel and efficient tool for fusion detection, and we think it is quite proper for researches on credible fusions. SOAPfuse is freely available from http://soap.genomics.org.cn/soapfuse.html.

#P24 Dec. 21, 2012 02:50 PM - 04:00 PM

Long Term Medical Spending in Systemic Lupus Erythematosus and Lupus Nephritis

Yu-Wen Chen[1], Yen-Jen Sung[1], Der-Ming Liou[1]

[1] National Yang-Ming University, Taiwan

Keywords: Systemic Lupus Erythematosus, Nephritis, National Health Insurance Research Database

Background and Motivations

Systemic lupus erythematosus is a chronic disorder which needs long-term treatment, most common in Asian population. More renal involvement in lupus lead to higher healthcare cost. However, publications on the economic burden of lupus are scarce in Taiwan. Our propose is to estimate health care costs among persons with systemic lupus erythematosus (SLE) with nephritis or not in Taiwan.

Proposed Approaches

Data are derived from 80,000 National Health Insurance Research Database beneficiaries with ICD-9-CM 710.0 and nephritis diagnosis codes between 2006 and 2010. The NHIRD includes demographic and enrollment data, such like inpatient, hospitalization data, inpatient drug exposure data, and outpatient diagnosis data, and outpatient pharmacy claims.

Results and Conclusions

Data were obtained from 70 SLE patients who met our standard. 10% (n =7) of the SLE patients had nephritis during the follow-up. SLE is more common in women than in men (89.38% of SLE patients and 86.14% of SLE patients with nephritis). The mean age is 46.8 years in SLE patients and 44.5 years in the group of nephritis. Mean medical costs were predominantly higher for SLE patients with nephritis (US$12,107) compare with SLE patients (US$5,355) in 2006 (P<0.001). Mean medical payment for SLE patients decrease in 2007 and 2009 but increase in 2008 and 2010, with a wavy motion performance. Final year cost was US$9,538 per patient payment. The mean annual costs for SLE patients were US$5,095 to US$3,306.

The annual mean medical payment for the SLE patients with nephritis was US$12,107 in the first year. Decreased by approximately 50% at the first two years, increased back at an average rate of 50% that arrived US$9,538 in 2010, which is more double the cost of the SLE patients group. Inpatient care is the mainly composed of medical payment, accounted for the largest part of costs in each year for SLE populations, comprising 75% till 80% of annual costs. Outpatient services accounted for 20% till 25% of costs. Similar situation showed in the group of SLE patient with nephritis. Costs of inpatient payment had the highest percentage, the range from 30% to 90%. By providing this meaningful information, it will provide a basis for the medical care cost and help decision makers for setting priorities for resource allocation and research activities.

#P25 Dec. 20, 2012 03:30 PM - 04:45 PM

Phylogenetic Analysis of Bifunctional DNase-RNase Domain Containing Proteins

Sung Moo Byeon[1], Jeong Sheop Shin[1]

[1] Korea University, Republic of Korea

Keywords: DNase-RNase, Bifunctional Nuclease, PF02577, phylogenetic analysis

Background and Motivations

In our early study, we found noble bifunctional nucleases from Arabidopsis thaliana(AtBBD1) and Oriza minuta(OmBBD) which contained a domain of unknown function(DUF151) with highly conserved sequence throughout species. AtBBD1 and OmBBD showed both DNase and RNase activity and have been categorized as DNase-RNase family by pfam databank(pfam.sanger.ac.uk). Proteins containing this domain were specific to Archaea, Bacteria and plants. To verify whether key amino acid residues for DNase-RNase activity were conserved throughout evolution, we performed a phylogenetic analysis of DNase-RNase domain containing proteins. Crystal structure of the protein containing homologous domain(TM0160) was also reviewed for possible active sites of bifunctional enzyme activity.

Proposed Approaches

Proteins containing DNase-RNase domain were clustered into four major groups according to their domain components. Group A consisting of only DNase-RNase domain, group B of DNase-RNase domain plus UVR domain, group C consisting of DNase-RNase plus UVR and HCR domain and group D consisting of variable regions within the HCR, DNase-RNase and UVR domains. Sequences were compared to find conserved residues within domain and phylogenetic tree of each group was constructed using Uniprot protein database (www.uniprot.org). An X-ray crystal analysis data of protein consisted only of DNase-RNase domain(TM0160) by Spraggon team (Spraggon et al., 2004) was also analyzed for possible catalytic site reference.

Results and Conclusions

Phylogenetic analysis revealed a core region that was conserved throughout sepecies which could be a good candidate for the putative enzymatic active site. Region 95-109 was very well conserved and Arginine 99 showed no exception throughout whole species. This result is in concordance with the suggestion of structural study by Spraggon et al. that aspartic acid 102 is one of the residues of putative active site.

#P26 Dec. 21, 2012 02:50 PM - 04:00 PM

Applications of Heat Map with Hierarchical Clustering in Mortality Risk – "Hot-Spots" Risk Assessment in Industrial Polluted Area

Hsin-Ling Yeh[1], Po-Huang Chiang[2], Der-Ming Liou[1], Hui-Chen Tsou[2]

[1] National Yang-Ming University, Taiwan
[2] National Health Research Institutes, Taiwan

Keywords: spatial analysis, cluster heap map, industrial pollution

Background and Motivations

Changhua and Yunlin are the most important cities of agriculture in Taiwan. In 1976, the Industrial Development Bureau of Ministry of Economic Affairs built the coastal industry for industrial development which invited many different plants which created many factories such as petrochemical plants, oil refinery plants, chemical plants, etc... Although the economy was induced, it created serious pollution of the environment and affected health of residents.

The purpose of this research is to survey the relationship between the highest death rate local and industrial pollution by analyzing the top ten leading causes of death in Changhua and Yunlin, in the years 2001-2011. Previous studies of a region cluster state of mortality use Geographic Information Systems (GIS) and spatial analysis. However, these methods only focus on one disease, while we need to combine all the data to overlay and compare outcomes. Thus we are unable to use this method because the final image from using this type of analysis on our data would complicate the presentation of the results. Hence, we used cluster heap map to display the relationship between spatial and temporal information of mortality data visually and rapidly.

Proposed Approaches

This study used GIS spatial autocorrelation analysis and cluster heat map in the statistical software R to survey the relationship between the highest death rate local and industrial pollution by analyzing the top ten leading causes of death in Changhua and Yunlin, from 2001-2011. The cluster heat map is an integrated heat map with hierarchical clustering that displayed the results of a cluster analysis by permuting the rows and the columns of a matrix to place similar values near each other. We used this method to survey distribution of mortality by year.

Results and Conclusions

We found the cluster states of male and female individual in Changhua 2006, 2001 and Yunlin 2005, 2008 spatial analysis were significant (Moran's I, $p<0.05$). Thus we graphically visualized these clusters and found that in these clusters the industrial air pollution had similar distribution to the death rates.

#P27 Dec. 20, 2012 03:30 PM - 04:45 PM

GRecScan: An RNA Virus Sequence Auto-typing and Inter-subtype Recombination Detection Tool Based on MFE

Bo Wei[1], Simon Rayner[1]

[1] Wuhan Institute of Virology, Chinese Academy of Sciences

Keywords: recombination, auto-typing, RNA virus, secondary structure

Background and Motivations

The error-prone replications of RNA viruses genomes results in rapid accumulation of mutations in the genome and generates genetic diversity, leading to the creation of genotypes, serotypes or subtypes with distinctive phenotypes. Diversity and evolution are further exacerbated by genetic recombination and reassortment events that can occur both within and between subtypes or even between different viruses or with host genes. Therefore, typing and recombination detection of RNA virus strains are significant in insight into evolution of RNA virus.

Proposed Approaches

In this research a heuristic clustering algorithm through phylogenetic tree is used to automatically objective typing RNA virus nucleotide sequences. Then possible recombination sequences between these subtypes are detected based on the similarity of folding Minimal Free Energy (MFE) of RNA sequence secondary structure and Mann-Whitney nonparametric U test is applied to test whether the variation of a fragment in the test sequence is significantly larger than the random variation within subtype, that is, if the variation of the fragment is greater than the intra-subtype variation, it is caused by recombination.

Results and Conclusions

A dataset including Dengue virus I recombination sequence D00503 and 35 reference sequences of all five genotypes of Dengue virus were tested through the algorithm. It successfully divided the dataset into five subtypes and the recombination strain D00503 was detected with P-value <0.05 and the recombination breakpoints were consistent with those previously reported. Another empirical dataset with HBV recombination isolate CHN-QH18 also passed the detection. A direct-viewing and friendly Java graphic interactive interface has been implemented for the algorithm, which can run on any platform with acceptable speed.

#P28 Dec. 21, 2012 02:50 PM - 04:00 PM

LocTree2 Predicts Localization for All Domains of Life

Tatyana Goldberg[1], Tobias Hamp[1], Burkhard Rost[1, 2]

[1]TUM, Bioinformatik-I12, Informatik, Boltzmannstrasse 3, Garching 85748, Germany
[2]New York Consortium on Membrane Protein Structure (NYCOMPS) and Department of Biochemistry and Molecular Biophysics, Columbia University, New York, NY 10032, USA

Keywords: Sub-cellular localization, sequence-based prediction, machine learning, support vector machines, sequence kernel, hierarchical ontology, evolutionary profile, trans-membrane proteins

Background and Motivations

The knowledge of the sub-cellular localization of a protein can help in elucidating its function, as a protein's localization can provide hints about its functional role in a cell. Despite advances in high-throughput imaging, localization maps remain importantly incomplete. Several methods have been developed that accurately predict localization, yet many challenges remain to be tackled.

Proposed Approaches

Here, we introduce LocTree2 to predict localization in all three domains of life, spanning water-soluble globular and membrane proteins. It predicts three localization classes for Archaea, six classes for Bacteria, and eighteen classes for Eukaryota. LocTree2 uses a hierarchical system of Support Vector Machines (SVMs) implemented to imitate the cascading mechanism of cellular sorting. The classification is made using sequence information only.

Results and Conclusions

The method reaches high levels of sustained performance for both Eukaryota (Q18=65%; Q18 is eighteen-state accuracy for classifying proteins to eighteen localization classes) and Bacteria (Q6=84%; Q6 is six-state accuracy). Our method also accurately distinguishes between membrane and non-membrane proteins. LocTree2 works well even for protein fragments; these may result from erroneous assemblies or wrong gene predictions that are common in genome projects. In our hands, LocTree2 compared favorably to other state-of-the-art methods when tested on new data.

Availability

Online through PredictProtein (predictprotein.org); as standalone version or a web-server at http://www.rostlab.org/services/loctree2.

#P29 Dec. 20, 2012 03:30 PM - 04:45 PM

A new Approach to Identify Possible Core Genes in Baculovirius

Ye Yanbo[1,2], Simon Rayner[1*]

[1] Wuhan Institute of Virology, China
[2] Graduate School of Chinese Academy of Sciences, China

Keywords: baculovirus, core gene, clustering

Background and Motivations

Baculovirus is a large group of dsDNA viruses, which can be classified into 4 different groups (αβγδ) based on phylogenetic analysis. Among those, γ and δ are more ancient groups with less species isolated. A recent study suggests there are 37 core genes in all baculovirus. However, additional core genes may exist. Two factors make it difficult to recognize the gene orthology and identify baculovirus core genes: 1. large evolutionary distances between each group, especially from αβ to γδ; 2. the diverse gene conservation of baculovirus genes makes it impossible to set a global cutoff. General sequence similarity search programs and clustering method cannot be applied directly to identify core genes.

Proposed Approaches

In this study, a new approach which takes these two factors into consideration was used to detect new possible core genes. Two relaxed psiblasts, αβ against αβ and γδ against all, with different substitution matrices (BLOSUM60 and BLOSUM45), were used to find all possible homology connections. Then a weighted graph combine both blast results was generated to represent their protein similarity relationships. The edge weights were calculated from e-values and evolutionary distances, and pairs with far distances were strengthened. After this, the Markov cluster algorithm (MCL) was used to cluster all genes, following by a manual pruning based on other important information (such as relative position of genes). 56 baculovirus genomes were downloaded from NCBI and analyzed in this study. All works were done in an interactive Java program BlastGraph developed by our group.

Results and Conclusions

33 out of 37 previous core genes were identified using this method. Ac53 and Ac101 (P40), which were identified by a recent study, were confirmed to be core genes by this approch. Besides, one more gene (calyx/pep/p10) might be considered as another core gene as it was detected to present in all baculoviruses from our result.

Author Index

Author Index

Author	ID	Page NO.
B		
Byeon, Sung Moo	P25	64
C		
Chan, Wen-Ling	O12	32
Chen, Binbin	O03	17
Chen, Chih-Wei	P09	48
Chen, Yu-Wen	P24	63
Chou, Chih-Hung	P08	47
F		
Fukasawa, Yoshinori	P21	60
Fukunaga, Tsukasa	P17	56
G		
Gabr, Moustafa Tarek	P03	42
Gao, Kun	P06	45
Goldberg, Tatyana	P28	67
Grover, Abhinav	O15	36
H		
Huang, Kai-Yao	O07	23
Hung, Jui-Hung	K04	31
Hwang, Woochang	O13	34
J		
Jia, Wenlong	P23	62
K		
Kawaguchi, Risa	P01	40
Ke, Chao-Hsuan	P22	61
Kim, Docyong	P13	52
Kim, Hanhae	O11	29
Kim, Junho	O06	21
Kim, Sungmin	P02	41
L		
Lai, Yi-Chun	P04	43
Lee, Haeseung	O14	35
Lee, Tak	O09	27
Liaw, Chyn	P12	51
Lu, Cheng-Tsung	O10	28

Author Index

Author	ID	Page NO.
M		
Mori, Ryota	P15	54
O		
Ozaki, Haruka	P07	46
P		
Praneenararat, Thanet	P11	50
S		
Sakata, Hayato	O01	13
Shin, Hyunjung	K03	26
Shrestha, Anish Man Singh	P18	57
Singh, Satyendra Pratap	P10	49
Su, Min-Gang	O08	24
T		
Taniguchi, Yuichi	K02	16
Tsuchiya, Takaho	P05	44
Tsuji, Junko	P19	58
W		
Wei, Bo	P27	66
Wu, Allison C.Y.	P16	55
Y		
Yang, Huanming	K01	12
Ye, Yanbo	P29	68
Yeh, Hsin-Ling	P26	65
Z		
Zeng, Chao	O02	14
Zheng, Han-Qin	P14	53
Zhou, Qi	O04	19

http://2012.ayrcob.org