Quantitative Biology 2013, 1(2): 175–181
DOI 10.1007/s40484-013-0015-1

QB

MEETING REPORT

Meeting report on RECOMB 2013 (the 17th Annual International Conference on Research in Computational Molecular Biology)

Xuegong Zhang[1,*] and Fengzhu Sun[2,1]

[1] MOE Key Laboratory of Bioinformatics and Bioinformatics Division/Center for Synthetic and Systems Biology, TNLIST, and Department of Automation, Tsinghua University, Beijing 100084, China
[2] Molecular and Computational Biology Program, University of Southern California, Los Angeles, CA 90089, USA
* Correspondence: zhangxg@tsinghua.edu.cn

Received May 20, 2013

RECOMB 2013 was successfully held in Tsinghua University, Beijing, China on April 7–10, 2013, hosted by the Bioinformatics Division and Center for Synthetic and Systems Biology, Tsinghua National Laboratory for Information Science and Technology (TNLIST). A total of about 500 professionals from both academia and industry from 29 countries and regions attended the conference and its RECOMB-Seq satellite workshop after the main conference.

The RECOMB conference series, with the full name of the Annual International Conference on Research in Computational Molecular Biology, was started in 1997 by Sorin Istrail, Pavel Pevzner and Michael Waterman. The conference aims at bringing together the computational, mathematical, statistical, and biological sciences, and bringing together researchers, professionals, students and industrial practitioners from all over the world for interaction and exchange of new developments in all areas of bioinformatics and computational biology. This year's conference features 6 keynote talks, and presentations of peer-reviewed high-quality research papers, selected highlight talks of recently published high-profile works, and poster presentations on the latest research progress. Figure 1 provides some sample snapshots of the conference.

THE KEYNOTE TALKS

The conference features 6 keynote speakers from world leaders in biological and life sciences. Dr. Deborah A. Nickerson (University of Washington, USA) started the conference with the applications of next generation sequencing (NGS) technologies to the study of human genomics, including the identification of genetic variants responsible for both rare and complex traits and the inference of evolutionary and demographic histories of protein coding mutations. Dr. Chung-I Wu (Chinese Academy of Sciences, China, and University of Chicago, USA) presented his recent work on the study of cell migration in hepatocellular carcinoma (HCC), which can be multifocal with several tumors in the same liver. He convincingly showed that in cell migration both invasion and dispersal are involved and they have different mechanisms. The work illustrated that selective advantage of migration is due to a mutual reinforcement of tumor-growth and cell-motility mutations, and the mutual reinforcement leads to a process of adaptive diversification that accelerates as tumors evolve. Dr. Takashi Gojobori (National Institute of Genetics, Japan) talked about the historical developments of various genomics databases in Japan and the relationships between genomics information and society. Dr. Scott Fraser (University of Southern California, USA) reported the exciting developments of intravital techniques for imaging combined with better sensors and molecular imaging to study the cellular and molecular events during complex biological processes including embryogenesis and organogenesis. These technologies make it possible to study the spatial and temporal heterogeneity of molecular events for a deep understanding of the biological systems. Dr. Nadia Rosenthal (Monash University, Australia, and Imperial College London, UK) reported her studies on the

Figure 1. Snapshots of RECOMB 2013. (A) Keynote talk by Dr. Nickerson; (B) The audience; (C) Poster Sessions; (D) Program committee members on the campus of Tsinghua University.

molecular mechanisms underlying tissue regeneration after injury. Using mouse as a model system, she presented results on the roles of growth factors and resident cells in the resolution of tissue injury, revealing the complex interactions between the molecular pathways and the cells of immune systems in tissue regeneration. Dr. Xiaoliang Sunney Xie (Harvard University, USA) gave the final keynote speech on single-molecule live-cell sequencing to study the stochastic effects of gene expression and genetic variation including single nucleotide polymorphisms, copy number variations, insertions/deletions, and translocations. In addition to these exciting technology developments, the keynote speakers also presented challenging and exciting computational problems associated with the various technologies.

OVERVIEW OF SUBMISSIONS

The RECOMB2013 had three tracks of submissions. The proceeding track called for extended abstracts on September 3, 2012 with a deadline of October 7, 2012. A total of 167 submissions were finally submitted and every submission was assigned to three program committee (PC) members who reviewed the extended abstracts personally or asked external reviewers for evaluation. As a result, 152 of the submissions received at least 3 reviews and 15 submissions received 2 reviews. The submissions came from 26 countries or regions with the majority of the authors coming from the United States, Germany, China, Canada, Israel and France (Table 1). The papers that received controversial reviews were put under extensive discussion among relevant PC members before decisions were reached. Finally, 32 extended abstracts were selected for oral presentation at RECOMB 2013 with an acceptance rate of 19%.

The highlight track called for abstracts based on papers published after October 2012 or in press at the time of submission and were already linked on the journal web site with a deadline of February 10, 2013. A total of 48 submissions were received and 9 of them were selected to be presented orally at the conference. The submissions came from 17 counties with the majority of the authors

coming from the United States, China, Canada, Israel and Russian Federation (Table 1). The talks were selected by the PC chair and the executive committee members based on the impact of the work on the field, the likelihood that the work makes a good presentation and the relevance to biological and biomedical research in general.

The poster track invited abstracts of ongoing and published researches that used mathematical, statistical and computational approaches to solve biological and biomedical problems. A total of 194 submissions were received. A group of poster committee members read and evaluated the submissions and gave suggestions to some of the posters. The poster authors came from 27 countries with the majority coming from China, United States, Canada and Germany (Table 1). The posters were presented at two separate sessions at RECOMB 2013. The poster track is an essential and integral part of the conference.

PROCEEDING AND HIGHLIGHT TALKS

In the following, we summarize the proceeding and highlight talks at RECOMB 2013.

Sequence assembly and sequence comparison: New high-throughput sequencing technologies continue to call for faster and more accurate algorithms for sequence assembly and comparison, two fundamental problems in computational biology. Based on the observations of improved high quality in Illumina sequencing reads and at most one insertion/deletion in most reads, Shijian Chen et al. (Chinese Academy of Sciences, China) designed a fast alignment program for mapping high-quality Illumina

Table 1. The number of authors from different countries or regions for the proceeding, highlight, and poster tracks, respectively, at RECOMB 2013

Proceeding track		Highlight track		Poster track	
United States	279	United States	55	China (the mainland)	269
Germany	40	China (the mainland)	20	United States	154
China (the mainland)	31	Israel	12	Canada	42
Canada	31	Russian Federation	12	Germany	28
Israel	18	Brazil	8	Netherlands	18
France	17	Canada	7	United Kingdom	18
Singapore	12	Switzerland	7	Hong Kong, China	14
Russian Federation	12	Hong Kong, China	6	Singapore	14
Japan	12	Korea, Republic of	6	Russian Federation	13
Hong Kong, China	10	United Kingdom	5	Taiwan, China	13
Australia	9	Ireland	4	France	12
India	6	Iran, Islamic Republic of	3	Poland	12
United Kingdom	5	Poland	2	Switzerland	11
Netherlands	5	Australia	1	Finland	9
Finland	5	Germany	1	Brazil	8
Turkey	3	Netherlands	1	Korea, Republic of	8
Korea, Republic of	3	Singapore	1	Israel	5
Spain	2			Denmark	4
Poland	2			Australia	3
Norway	2			Iran, Islamic Republic of	3
Tunisia	1			Ireland	3
Switzerland	1			India	2
Romania	1			Greece	1
Italy	1			Qatar	1
Brazil	1			Romania	1
Belgium	1			Spain	1
				United Arab Emirates	1

reads to reference genomes. Single-cell genome sequencing projects are usually characterized by high non-uniform coverage of reads across the genome and elevated rate of chimeric reads and read-pairs making sequencing assembly of single cells challenging. Sergey Nurk et al. (University of California at San Diego, USA) presented algorithms for the identification of chimeric edges and the resolution of complex bulges in de Bruijn graphs resulting in significant improvement of single cell sequencing assembly.

Although molecular sequences are generated rapidly, most of the new sequences are similar to existing ones and thus, the amount of novel sequence information is growing much more slowly. In a highlight talk, Po-Ru Loh et al. (Massachusetts Institute of Technology, USA) showcased their recent work on "compressive genomics" by compressing the sequencing data exploiting the redundant information to reduce data storage requirement and to increase computational speed for sequence analysis.

Population genomics: Various genomic variation data including single nucleotide polymorphisms (SNPs), copy number variations (CNVs), insertions/deletions (Indels), etc., are now widely available for a large number of individuals. Such genomic variation data make it possible to find genetic basis of complex traits, to construct *cis*- and *trans*-regulation networks, and to infer the evolution and demographic histories of population samples. To achieve these objectives, significant challenges remain, including the design of appropriate models and statistics, efficient computational methods to calculate the statistics, and appropriate approaches for the adjustment of multiple testing. Itamar Eskin et al. (Tel-Aviv University, Israel) considered the problem of estimating the fraction of genetic materials of each individual in pooling studies using individual genotypes and the allele frequencies in a pool. The method can also be used to study the abundance levels of microbial organisms with known genome sequences in metagenomics studies. Emrah Kostem and Eleazar Eskin (University of California at Los Angeles, USA) designed a novel two-stage testing procedure that identifies all significant SNP and gene expression level associations without testing all SNP-expression pairs. In the first stage, a small fraction of informative SNPs are tested, and regions that may potentially contain significant associations are identified. In the second stage, more SNPs in regions identified in the first stage are tested. The resulting algorithm can speed up the current available ones by at least 75 folds.

The consideration of hidden relationships of individuals in genome wide association studies is important to control false positive and false negative results. However, such hidden relationships among the individuals are generally unknown and must be estimated from the genomics data. Dan He (IBM T.J. Waterson Research, USA) considered the problem of pedigree reconstruction based on genotype data. They developed an inheritance path based approach that efficiently uses the dynamic programming algorithm to infer the relationships among individuals in the same generation. The algorithm runs in linear time with respect to the number of generations and thus can be used for pedigree reconstruction for large pedigrees. Jesse M. Rodriguez et al. (Stanford University, USA) developed a novel computationally-efficient method, PARENTE, to detect related pairs of individuals and their shared haplotypic segments based on genotypes.

Evolutionary genomics: Understanding the evolutionary history of population samples and gene families has been a core problem for a long time. In evolutionary studies, the comparison of a gene tree with its species tree under a reconciliation framework can help the understanding of gene family evolution including gene duplication, loss and transfer. Although the maximum parsimony solution to the Duplication-Loss (DL) reconciliation is unique, there are multiple solutions to the maximum parsimony Duplication-Transfer-Loss (DTL) reconciliation problem. Mukul S. Bansal et al. (Massachusetts Institute of Technology, USA) presented an effective and scalable computational method to deal with the fundamental DTL problem. Roy Ronen et al. (University of California at San Diego, USA) studied the important traditional problem of identifying loci responsive to selective pressure and proposed a support vector machine based approach using site frequency spectrum to solve the problem. The new approach was used to identify loci adaptive to hypoxia in drosophila and it was shown that the Notch pathway plays important role in hypoxia tolerance. Yufeng Wu (University of Connecticut, USA) reported a new exact algorithm to construct the most parsimonious phylogenetic network that displays a given set of gene trees. Jesper Jansson et al. (Kyoto University, Japan) presented a deterministic algorithm to build the majority rule consensus tree for a set of conflicting phylogenetic trees.

Cancer genomics: Cancer is a complex disease under extensive studies, and high-throughput DNA and RNA sequencing data for individual and mixtures of cancer cells are available. Layla Oesper et al. (Brown University, USA) described an algorithm, Tumor Heterogeneity Analysis (THetA), to infer the number of distinct cell subpopulations and their fractions directly from high-throughput DNA sequencing data from a mixture of tumor samples. Raheleh Salari et al. (Stanford University, USA) developed an algorithm to infer the phylogenetic tree of multiple related tumor tissue samples and then to infer somatic SNPs based on the phylogenetic tree. It was shown that the accuracy of the inferred SNPs is higher than current available methods that do not use the

phylogenetic relationships among the samples. Dong-Yeon Cho and Teresa M. Przytycka (National Institutes of Health, USA) proposed a probability method to study the heterogeneity of cancer cells by leveraging the phenotype similarity among the different cancer cells measured by the similarity between the gene expression profiles. Fabio Vandin (Brown University, USA) studied the problem of identifying sequence variants associated with cancer survival. They showed that the traditional widely used *log-rank* test for survival analysis based on the approximate *p*-value distribution is not adequate for genomics studies, and reported a fully polynomial time approximation scheme (FPTAS) to calculate the *p*-value of the *log-rank* test.

Solid tumor cells are generally heterogeneous, containing both cancer and normal cells, which makes the interpretation of their genomic profiles challenging. In a highlight talk, Yinyin Yuan et al. (University of Cambridge, UK) presented histopathology images, gene expression profiles, and DNA copy number variation data for a group of breast cancer patients, and developed a computational approach to integrate these data to accurately predict patient survival.

Epigenomics: Chromatin interactions are essential for gene regulation since they bring distal genomic regulatory sequences spatially close to their regulatory targets. ChIA-PET is a molecular experimental technique measuring chromatin interactions between genomic sites bound by a particular protein. Christopher Reeder and David Gifford (Massachusetts Institute of Technology, USA) presented a probabilistic mixture model based approach to accurately predict chromatin interactions. Michael Stevens et al. (Washington University in St Louis, USA) introduced a new conditional random fields-based method, MethylCRF, to predict DNA methylation levels at single CpG resolution integrating methylated DNA immunoprecipitation (MeDIP-Seq) and methylation sensitive restriction enzyme sequencing (MRE-Seq) data.

Two highlight talks are related to epigenomics. The ribosome profiling technique (Ribo-Seq) makes it possible to map the locations of translating ribosomes on mRNAs. Audrey M. Michel et al. (University College Cork, Ireland) identified a triplet periodicity pattern of ribosome protected fragments (RPF) when they are aligned to mRNA. They developed a computational method for detecting transitions between reading frames that occur during programmed ribosomal frame-shifting or in dual coding regions where the same nucleotide sequence codes for multiple proteins in different reading frames using the triplet periodicity pattern. Sheng Zhong (University of California at San Diego, USA) presented both experimental and computational results of epigenomic markers in human, mouse, and pig and studied the conservation of the epigenomic markers. His group also developed a finite mixture model of HMMs to cluster genomic sequences based on the similarity of temporal changes of multiple epigenomic marks during a cellular differentiation process and applied the method to study differentiation of mouse embryonic stem (ES) cells.

Gene expression regulation: Transcription factors (TFs) regulate the expression of target genes by binding to specific sites of the *cis*-regulatory regions of the target genes. Although members of some protein families recognize similar DNA patterns, they do not bind to the same locations *in vivo*. Using features reflecting the DNA binding specificities of putative co-factors of two TFs: c-Myc and Mad2, Alina Munteanu (Alexandru I. Cuza University, Romania) and Raluca Gordân (Duke University, USA) designed a classification approach to distinguish the targets of the TFs.

Three highlight talks are related to the regulation of gene expression. Yitzhak Friedman et al. (The Hebrew University of Jerusalem, Israel) presented their work on miRNA combinatorial regulation on gene expression and developed a systematic approach to identify pairs and triplets of miRNAs that maximally affect cellular pathways. Bartek Wilczynski et al. (EMBL Heidelberg, Germany, and University of Warsaw, Poland) described a Bayesian network approach to integrate diverse information including transcription factor recruitment to multiple *cis*-regulatory modules, insulator binding and histone modification during *drosophila* development to predict spatial and temporal aspects of gene expression. Raheleh Salari et al. (Stanford University, USA) studied the contributions to expression variation at the transcription and the translation stages. In particular, they studied how sequence features related to translation efficiency increase noise strength and how such increase compares to the amplification associated with the TATA box.

Mass spectrometry: Mass spectrometry (MS) has been widely used for peptide sequencing. Most *de novo* sequencing algorithms depend on the prior knowledge of the fragmentation properties such as ion types and their propensities of specific MS/MS spectra. Kyowon Jeong et al. (University of California at San Diego, USA) provided a universal tool, UniNovo, for *de novo* peptide sequencing that does not depend on specific fragmentation techniques. Xiaowen Liu et al. (University of California at San Diego, USA) presented a fast method to identify both expected and unexpected multiple post-translational modifications (PTMs) based on top-down MS data. Mingxun Wang and Nuno Bandeira (University of California at San Diego, USA) described a novel method, Spectral Library Generating Function, for assigning statistical significance to spectrum-spectrum matches obtained from spectral library searches of proteomic tandem mass spectra.

RNA and chromosome structure: Microarray and RNA-Seq gene expression data showed clearly that the genomes of all organisms are pervasively transcribed, resulting in large quantities of non-coding RNAs (ncRNAs). Thus, the studies of RNAs become increasingly important in recent years. Sebastian Will et al. (University of Freiburg, Germany) described a simultaneous folding and alignment algorithm for RNA secondary structures that solves a decade-old problem of devising a practicable Sankoff-style algorithm for RNA secondary structure prediction. Evan Senter et al. (Boston College, USA) developed a Fast Fourier Transform (FFT)-based scheme to accelerate the search for RNA conformational switches. The use of FFT led to a substantial improvement in performance in both space and time. Vladimir Reinharz et al. (McGill University, Canada) described a dynamic programming approach to compute the probability that a specific position has a particular nucleotide given a secondary structure, a sequence and a prescribed maximal number of mutations. This approach can potentially be used to identify errors in deep sequencing data for structured RNAs.

Within the nucleus, chromosomes are not linear, instead, they form complex three dimensional structures affecting many chromosomal mechanisms including gene regulation, DNA replication, epigenetic modifications, and maintenance of genome stability. Hi-C data can provide contact frequency information among different parts of the genomes. Zhizhuo Zhang et al. (National University of Singapore, Singapore) presented a deterministic method, ChromSDE, that employs semi-definite programming (SDP) techniques to find the best structure fitting the observed data from Hi-C experiments.

Protein structure: The determination of protein structure in solution and structural changes when a protein interacts with other proteins is important for understanding its function. The depth of residues in a protein structure is a key useful attribution for protein structure modeling and function annotation. However, available methods for calculating residual depth is time consuming and not efficient. Dong Xu et al. (Sanford-Burnham Medical Research Institute, USA) presented a fast and accurate Euclidean distance transform (EDT)-based method to compute the depth of residues in proteins. They showed that the method is 2.6 times faster than the state of the art currently available methods. Chittaranjan Tripathy et al. (Duke University, USA) developed two algorithms for global protein fold determination incorporating the information from residual chemical shift anisotropies (RCSAs) in addition to residual dipolar coupling (RDC) from NMR experiments. They derived analytical solutions for the peptide plane orientations and torsion angles based on RDC and RCSA, and then used the analytical results to solve for the protein structure. Proteins interact through their domains, and some interaction hot spots, small subsets of binding surfaces that accounts for the majority of binding free energy, are usually present. Lei Deng et al. (Tongji University, China) presented a computational method to predict protein binding hot spots using structural properties of the proteins. When proteins interact with other proteins, conformation changes can usually be observed. Fei Guo et al. (City University of Hong Kong, China) proposed a flexible approach, FlexDoBi, for detecting conformational changes of proteins given protein interactions.

Molecular networks: Molecules perform their functions by interacting with other molecules to form networks including gene regulation, protein-protein physical and genetic interactions, and metabolic networks. Kriti Puniyani and Eric P. Xing (Carnegie Mellon University, USA) developed a network inference algorithm integrating multiple different data sources assuming the same underlying relationships between the nodes. The algorithm uses the semi-parametric Gaussian copula to model the distribution of the different data sources with different copulas sharing the same covariance matrix and estimates the model in the high dimensional scenario. Mohammad Javad Sadeh et al. (University of Regensburg, Germany) presented a method to infer signaling networks from gene expression profiles of gene silencing assays. They addressed the question of which features of a network can be confounded by incomplete observations and designed a test for inferring such non-confoundable characteristics of signaling networks. Ali A. Faruqi et al. (Imperial College, UK) constructed an ancestral state of metabolic networks for 141 bacteria and analyzed the reaction gains and losses for these bacteria with respect to their lifestyles and pathogenic nature. They showed that there exist statistically significant relationships between metabolism and lifestyle in some cases. Network motifs, subnetworks that are over-represented in a network, may represent biological functional units. However, the observed network from high-throughput experiments is usually incomplete and contains many false positive edges. Ngoe Hieu Tran et al. (National University of Singapore, Singapore) developed a powerful method to correct link errors in estimating directed and undirected network motifs in noisy networks.

Several highlight talks also addressed various aspects of molecular networks. Qiangfeng Cliff Zhang et al. (Columbia University, USA) showed that geometric relationships between protein structures can be used to accurately predict protein-protein interactions on a genome-wide scale, and he developed an algorithm combining structural information with non-structural clues to predict protein interactions with comparable quality to high-throughput experiments. Shihua Zhang et

al. (Chinese Academy of Sciences, China) developed a matrix factorization based approach to understand combinatorial control at multiple levels including DNA methylation, gene expression, and microRNA expression and applied the method to The Cancer Genome Atlas (TCGA) data.

AWARDS AND FELLOWSHIPS

With the generous donation of some sponsors, RECOMB 2013 was able to set three tracks of fellowships to conference attendees, especially students. They are: the ISCB student travel fellowship supported by the International Society for Computational Biology (ISCB), the NSF student travel fellowship supported by the US National Science Foundation (NSF), and the Tsinghua/TNLIST travel fellowship supported by Tsinghua University and the Tsinghua National Laboratory for Information Science and Technology. These fellowships partially supported the traveling for a total of 60 students and postdoc fellows.

RECOMB has the tradition of setting a Test of Time Award, the Best Paper Award and Best Poster Award. The Test of Time Award was selected for papers published 12 years ago at the RECOMB conference that have the highest influence to the field. This year's Test of Time Award goes to Drs. Pavel A. Pevzner, Haixu Tang, and Michael S. Waterman for their work "A New Approach to Fragment Assembly in DNA Sequencing" which formulated the sequence assembly problem to a graph searching problem, and to Drs. Jeremy Buhler and Martin Tompa for their work "Finding motifs using random projections", both presented at RECOMB 2001. The Best Paper Award of RECOMB 2013 goes to the paper "Genome-wide survival analysis of somatic mutations in cancer" by F. Vandin, A. Papoutsaki, B.J. Raphael and E. Upfal, and the Best Poster Award of RECOMB 2013 goes to two poster presentations, "Implementation of efficient haplotype matching using suffix array based methods" by T. Ilicic and R. Durbin, and "Functional distinctive CTCF bindings revealed by a novel motif discovery pipeline" by R. Fang and Z. Zhang. The Best Poster Award was supported by the journal *Nucleic Acids Research*, and the Best Paper Award was supported by the journal *Quantitative Biology*.

THE RECOMB-SEQ SATELLITE WORKSHOP

The RECOMB conference series has a series of Satellite Workshops on a variety of specific areas. RECOMB-Seq, the Annual RECOMB Satellite Workshop on Massively Parallel Sequencing is an annual forum for exploring the computational aspects of research, development and novel applications of high-throughput sequencing in life sciences. The Third Annual RECOMB Satellite Workshop on Massively Parallel Sequence or RECOMB-Seq 2013 was held at Tsinghua University on April 11–12, 2013, immediately following the main RECOMB 2013 conference.

The RECOME-Seq 2013 workshop features 3 keynote speakers: Dr. Colin Collins (University of British Columbia, Canada) presented their investigation on prostate cancer with extensive applications of next-generation sequencing and bioinformatics. Dr. Bing Ren (University of California, San Diego and Ludwig Institute for Cancer Research, USA) brought to the audience the latest progress on the study of the 3D organization of the genome with next-generation DNA sequencing. Dr. Inna Dubchak (DOE Joint Genome Institute, USA) presented the existing work and future challenges for efficient visualization of next-generation sequencing data.

After peer-review by the program committee (chaired by Haixu Tang, Indiana University, USA and Tao Jiang, University of California, Riverside, USA) on the 41 submissions, a total of 23 papers were accepted for oral presentation at the two-day workshop, with authors from China, Korea, France, Spain, Finland, Russia, Turkey, Australia, Canada and USA. The papers covered a broad range of topics on high-throughput sequencing and its applications. About 200 participants attended the workshop. Figure 2 is a snapshot of the workshop.

(About the authors: Xuegong Zhang is the Conference Chair of RECOMB 2013, and Fengzhu Sun is the Program Committee Chair of RECOMB 2013.)

Figure 2. A snapshot of the RECOMB-Seq 2013 Workshop.