Introduction

High-throughput next-generation sequencing technologies foster new cutting-edge computing techniques in bioinformatics

Mary Qu Yang[1], Brian D Athey[2], Hamid R Arabnia[3], Andrew H Sung[4], Qingzhong Liu[4], Jack Y Yang[5], Jinghe Mao[6] and Youping Deng*[7]

Address: [1]National Human Genome Research Institute, National Institutes of Health (NIH), U.S. Department of Health and Human Services, Bethesda, MD 20892, USA, [2]National Center for Integrated Biomedical Informatics (NCIBI) and Center for Computational Medicine and Biology, University of Michigan, Ann Arbor, MI 48109, USA , [3]Department of Computer Science, University of Georgia, Athens, GA 30602-7404, USA , [4]Department of Computer Science, New Mexico Tech, Socorro, NM 87801, USA, [5]Harvard University, P.O. Box 400888, Cambridge, MA 02140-0888, USA, [6]Department of Biology, Tougaloo College, Tougaloo, MS 39174, USA and [7]Department of Biological Sciences, University of Southern Mississippi, Hattiesburg, MS 39406, USA

Email: Mary Qu Yang - yangma@mail.NIH.gov; Brian D Athey - bleu@umich.edu; Hamid R Arabnia - hra@cs.uga.edu; Andrew H Sung - sung@cs.nmt.edu; Qingzhong Liu - sung@cs.nmt.edu; Jack Y Yang - Dr.Yang@JHU.edu; Jinghe Mao - jmao@tougaloo.edu; Youping Deng* - Youping.deng@usm.edu

* Corresponding author

Abstract

The advent of high-throughput next generation sequencing technologies have fostered enormous potential applications of supercomputing techniques in genome sequencing, epi-genetics, metagenomics, personalized medicine, discovery of non-coding RNAs and protein-binding sites. To this end, the 2008 International Conference on Bioinformatics and Computational Biology (Biocomp) – 2008 World Congress on Computer Science, Computer Engineering and Applied Computing (Worldcomp) was designed to promote synergistic inter/multidisciplinary research and education in response to the current research trends and advances. The conference attracted more than two thousand scientists, medical doctors, engineers, professors and students gathered at Las Vegas, Nevada, USA during July 14–17 and received great success. Supported by International Society of Intelligent Biological Medicine (ISIBM), International Journal of Computational Biology and Drug Design (IJCBDD), International Journal of Functional Informatics and Personalized Medicine (IJFIPM) and the leading research laboratories from Harvard, M.I.T., Purdue, UIUC, UCLA, Georgia Tech, UT Austin, U. of Minnesota, U. of Iowa etc, the conference received thousands of research papers. Each submitted paper was reviewed by at least three reviewers and accepted papers were required to satisfy reviewers' comments. Finally, the review board and the committee decided to select only 19 high-quality research papers for inclusion in this supplement to *BMC Genomics* based on the peer reviews only. The conference committee was very grateful for the Plenary Keynote Lectures given by: Dr. Brian D. Athey (University of Michigan Medical School), Dr. Vladimir N. Uversky (Indiana University School of Medicine), Dr. David A. Patterson (Member of United States National Academy of Sciences and National Academy of Engineering, University of California at Berkeley) and Anousheh Ansari (Prodea Systems, Space Ambassador). The theme of the conference to promote synergistic research and education has been achieved successfully.

Introduction

With the advent of high-throughput next generation sequencing technologies, gigabases of sequence information can be obtained in just a few days. The technologies offer drastically faster and cost-effective sequence throughput and are vastly superior to shotgun sequencing due to the high volume of data and the drastically short time to sequence a whole genome or disease genome, but genome assembly is much more computational expansive. Therefore, the next generation sequencing technologies will foster enormous potential applications of supercomputing techniques in genome sequencing, epigenetics, metagenomics, personalized medicine, discovery of non-coding RNAs and protein-binding sites. Furthermore, Next-generation sequencing will substitute microarray, the mostly used technology in genomics and bioinformatics: Like next-generation sequencing, micro-arrays can be used to examine thousands of genes in one experiment and can be used to obtain gene profiles, but the drawback of Microarrays are based on *hybridization*. Gene expression levels are measured by fluoresce from hybridizes but quantification of the fluorescence of vast amount of spots on a chip are often unreliable from experiment to experiment. Furthermore many DNA samples can hybridize to more than one spot, thus, generating misleading results. Next-generation sequencing overcomes problems of Microarrays by generating actual sequence reads and is ideally detect genetic mutations. To measure the gene expression level reflected by the amount of a particular RNA molecule, simply tally up the number of sequence reads corresponding to that RNA molecule rather measuring an inaccurate fluorescent spot and trying to control for all sorts of experimental variation. Gene expression can be actually more accurately obtained by counting sequence reads. Next generation sequencing has distinct advantages to obtain regulatory markings in chromatin, and to identify neural regulatory protein binds in the genome, as well as to investigate differences between stem cells and differentiated cells such as cancerous cells, and to determine how gene regulatory network can be altered by an activated external signal. Next generation sequencing is cheaper, fast, and less time consuming but computational expensive. Given situations, the development of computational techniques is important for future bioinformatics data mining. To this end, the International Society of Intelligent Biological Medicine http://www.ISIBM.org works with academic conferences to promote the cutting edge research.

Research presentations

Lianjiang Wang, Jack Y. Yang and Mary Qu Yang [1] presented a novel computational approach to predict DNA-binding residues from protein sequence information using random forests. The method can be used in analysis a large amount of genomic data from high-throughput next generation sequencing and provide insight in protein-DNA interactions. Yadong Wang, Jack Y. Yang, Yunlong Liu et al [2]. developed a method for reconstructing Gene Regulatory Network using Slice Pattern Model. Such model could be applied in studying data from next generation sequencing technology. Qingzhong Liuet al [3] presented a comprehensive comparison of feature selection and classification for MALDI-MS data. Xin Wang, Yunlong Liu et al. [4] presented Genome-wide prediction of cis-acting RNA elements regulating tissue-specific pre-mRNA alternative splicing, Chacko et al. [5] presented a comprehensive splicing graph analysis of alternative splicing patterns in chicken and compared to human and mouse. Osborne, et al. [6] discussed annotations of the Human Genome with Disease Ontology. Valdimir N. Uversky, Zoran, Keith Dunker et al [6] systematically studied Unfordomics for human diseases and demonstrated links to protein intrinsic disorder with diseases. Corresponding author Uversky presented this important research and discoveries in the keynote lecture. Jianhua Ruan, Youping Deng, Weixiong Zhang [8] presented an ensemble learning approach to the problem of reverse-engineering transcriptional regulatory networks using time-series gene expression data. Hong Zhou et al. [9] studied energy profile and secondary structure that impact shRNA Efficacy. Taheri et al. [10] studied RBT-GA as a example of their novel metaheuristic for solving the multiple sequence alignment problem. Gu et al. [11] reported their studies in genomic and systems evolution in vibrionaceae species. Midic, A Keith Dunker et. al. [12] systematically studied protein disorder in the human diseasome and showed unfoldomics in human genetic diseases, Uversky delivered keynote speech summarizing their seminal research. Kim et al. [13] studied PDA as an example of automatic and comprehensive analysis program for protein-DNA complex structures. Huang et al. [14] studied differences in duplication age distributions between human GPCRs and their downstream genes from a network prospective. Zhang et al. [15] presented their research in reverse engineering module networks by PSO-RNN hybrid modeling. Jake Chen [16] talked about their Online database for comprehensive human annotated and predicted protein interactions. P. Ghosh, Joe Zhang et al. [17] formulated a computational model to predict the binding rate for the siRNA-RISC complex formation reaction. Lichtenberg et. al [18] studied promoters in human DNA repair gene pathways. Zhongxue Chen, Monnie McGee, Qingzhong Liu et al. [19] studied a distribution-free convolution model for background correction of oligonucleotide microarray data. The above 19 research papers were selected by the committee based on peer-reviews. The selected papers covered broad range of research fields; mainly include gene regulation, gene regulatory network construction, protein-DNA interaction predication and modeling, microarray data analysis, evolution study, dis-

ease study and biology database. The authors presented a variety of approaches and tools that can be used in analyzing next generation sequencing data. Meanwhile the researchers can greatly benefit from quality and quantity of data generating by next generation sequencing technology.

Future meeting
The next annual conference will be held in same location in Las Vegas, Nevada on July 13–16, 2009. The web site, http://www.world-academy-of-science.org, contains further information on future meetings. The meeting is a large international conference with more than two thousand attendees from more than 80 countries annually. The conference continuously aims at promoting computational science research and education in biomedical sciences. For articles from the 2008 conference, please see http://www.biomedcentral.com/1471-2164/9?issue=S1, and also http://www.biomedcentral.com/1471-2164/9?issue=S2 for further reading.

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
All authors provided professional services to the conferences and contributed in writing this introductory article. All authors reviewed and agreed on the content of this introductory article.

Acknowledgements

References
1. Wang LJ, Yang JY, Yang MQ: **Prediction of DNA-binding residues from protein sequence information using random forests.** *BMC Genomics* 2009, **10(Suppl 1):**S1.
2. Wang Y, Wang G, Yang B, Tao H, Yang JY, Deng Y, Liu Y, *et al.*: **Reconstruct Gene Regulatory Network using Slice Pattern Model.** *BMC Genomics* 2009, **10(Suppl 1):**S2.
3. Liu Q, Sung AH, Qiao M, *et al.*: **Comparison of feature selection and classification for MALDI-MS data.** *BMC Genomics* 2009, **10(Suppl 1):**S3.
4. Wang X, Wang K, Radovich M, Liu Y, *et al.*: **Genome-wide prediction of cis-acting RNA elements regulating tissue-specific pre-mRNA alternative splicing.** *BMC Genomics* 2009, **10(Suppl 1):**S4.
5. Chacko E, Ranganathan S: **Comprehensive splicing graph analysis of alternative splicing patterns in chicken, compared to human and mouse.** *BMC Genomics* 2009, **10(Suppl 1):**S5.
6. Osborne JD, Flatow J, Holko M, *et al.*: **Annotating the Human Genome with Disease Ontology.** *BMC Genomics* 2009, **10(Suppl 1):**S6.
7. Uversky VN, Oldfield CJ, Midic U, Dunker AK, *et al.*: **Unfordomics of human diseases: linking protein intrinsic disorder with diseases.** *BMC Genomics* 2009, **10(Suppl 1):**S7.
8. Ruan J, Deng Y, Perkins E, Zhang W: **An ensemble learning approach to reverse-engineering tran-scriptional regulatory networks from time-series gene expression data.** *BMC Genomics* 2009, **10(Suppl 1):**S8.
9. Zhou H, Zeng X: **Energy profile and secondary structure impact shRNA efficacy.** *BMC Genomics* 2009, **10(Suppl 1):**S9.
10. Taheri J, Zomaya AY: **RBT-GA: A novel metaheuristic for solving the multiple sequence alignment problem.** *BMC Genomics* 2009, **10(Suppl 1):**S10.
11. Gu J, Neary J, Cai H, *et al.*: **Genomic and systems evolution in Vibrionaceae species.** *BMC Genomics* 2009, **10(Suppl 1):**S11.
12. Midic U, Oldfield CJ, Dunker AK, Obradovic Z, Uversky VN, *et al.*: **Protein Disorder in the human diseasome: unfoldomics of human genetic diseases.** *BMC Genomics* 2009, **10(Suppl 1):**S12.
13. Kim RG, Guo JT, *et al.*: **PDA: an automatic and comprehensive analysis program for protein-DNA complex structures.** *BMC Genomics* 2009, **10(Suppl 1):**S13.
14. Huang Y, Zheng Y, Su Z, Gu X: **Differences in duplication age distributions between human GPCRs and their downstream genes from a network prospective.** *BMC Genomics* 2009, **10(Suppl 1):**S14.
15. Zhang Y, Xuan J, *et al.*: **Reverse Engineering Module Networks by PSO-RNN Hybrid Modeling.** *BMC Genomics* 2009, **10(Suppl 1):**S15.
16. Chen JY, Mamidipalli SR, Huan T: **HAPPI: an Online Database of Comprehensive Human Annotated and Predicted Protein Interactions.** *BMC Genomics* 2009, **10(Suppl 1):**S16.
17. Ghosh P, Dullea R, Fischer JE, *et al.*: **Comparing 2-nt 3' overhangs against blunt-ended siRNAs: a systems biology based study.** *BMC Genomics* 2009, **10(Suppl 1):**S17.
18. Lichtenberg J, Jacox E, Yang MQ, Elntiski L, Welch L, *et al.*: **Promoters in Human DNA repair Gene Pathways.** *BMC Genomics* 2009, **10(Suppl 1):**S18.
19. Chen Z, McGee M, Liu Q, *et al.*: **A Distribution-Free Convolution Model for Background Correction of Oligonucleotide Microarray Data.** *BMC Genomics* 2009, **10(Suppl 1):**S19.