High Performance Computing in Biomedical Informatics

Hesham H. Ali
UNO Bioinformatics Core Facility
College of Information Science and Technology
University of Nebraska at Omaha
hali@unomaha.edu

Description

The last few years have witnessed significant developments in various aspects of Biomedical Informatics, including Bioinformatics, Medical Informatics, Public Health Informatics, and Biomedical Imaging. The explosion of medical and biological data requires an associated increase in the scale and sophistication of the automated systems and intelligent tools to enable the researchers to take full advantage of the available databases. The availability of vast amount of biological data continues to represent unlimited opportunities as well as great challenges in biomedical research. Developing innovative data mining techniques and clever parallel computational methods to implement them will surely play an important role in efficiently extracting useful knowledge from the raw data currently available. The proper integration of carefully selected/developed algorithms along with efficient utilization of high performance computing systems form the key ingredients in the process of reaching new discoveries from biological data. This tutorial focuses on addressing several key issues related to the effective utilization of High Performance Computing (HPC) in biomedical informatics research, in particular, how to efficiently utilize high performance systems in the analysis of massive biological data. A major key issue in that regard is how to develop innovative network models that allow researchers to integrate different types of biological data and extract useful knowledge out of all available datasets. Another major issue is how to design energy-aware parallel computational models for executing computationally-intensive biomedical applications on HPC systems. The integration between biomedical informatics and HPC will undoubtedly be a major driver in the next generation of biomedical research.

Objectives of the Tutorial

The field of Biomedical Informatics has been attracting a lot of attention in recent years. The massive size of the current available biological and medical databases and its high rate of growth have a great influence on the types of research currently conducted and researchers are focusing more than ever to maximize the use of these databases. Hence, it would be of great advantage for researchers to utilize High Performance Computing (HPC) system to explore the data stored in the available databases and extract new information that would lead to better understanding of various biological and medical phenomena.

The Biomedical Informatics domain is rich in applications that require extracting useful information from very large and continuously growing sequence of databases. The marriage between the bioinformatics domain and high performance computing is a natural one; particularly when it comes to the analysis of massive biological networks,

the problems in this domain tend to be highly parallelizable and deal with large datasets, hence using HPC is a natural fit.

In addition, from the IT point-of-view, the problem of efficiently collecting, sharing, mining and analyzing the wealth of information available in a growing set of the biological and clinical data has common roots in many IT applications. Addressing these issues require significant computational facilities; hence the need to integrate HPC research. How to efficiently manage the utilization of HPC systems in Biomedical Informatics is quickly emerging as one of the most urgent and critical problems in advancing biomedical research.

Topics to be covered in the Tutorial

The tutorial is designed for three hours and is divided into two parts, each scheduled for 80 minutes with a 20 minutes break. The first part covers the introduction, the background and an overview of key problems, algorithms and current tools in the area of Biomedical Informatics. The first part is covered in points 1-3 below. The second part focuses on introducing the audience to models for integrating HPC systems in Biomedical research with a focus on the concept of next generation data analysis and integration tools; that are Intelligent, Collaborative and Dynamic (ICD). The integration of HPC systems and Biomedical informatics using various network models will be presented, and then a focus on two specific case studies related to efficient utilization of HPC in biomedical research will be covered in details. These case studies are related to HPC energy-aware models and efficient parallel algorithms for sampling large biological networks. This part is covered in points 4-7 below.

1. Introduction to Biomedical Informatics - Brief discussion on the various aspects of Biomedical Informatics that include Bioinformatics, Medical Informatics, Public Health Informatics, and Biomedical Imaging.
2. Background – The Bioscience aspect and the computational perspective, the need for efficient HPC models for addressing key problems in Biomedical Informatics.
3. Biomedical Informatics now – current state of the emerging discipline and overview of key Biomedical Research problems, plus an overview of selected current, first generation, data analysis tools
4. The need for next generation data integration and analysis tools; Intelligent, Collaborative and Dynamic (ICD) Tools – A focus on advanced biological networks.
5. High Performance Computing (HPC) in Biomedical Informatics Research: current practices, pros and cons. A focus on HPC and new data integration and analysis tools.
6. Energy-aware scheduling in HPC: Case study – Scheduling models for computationally-intensive Bioinformatics applications on HPC systems.
7. HPC and the analysis of biological networks – Parallel Algorithms for filtering biological networks: Case Study - Correlation Networks and the identification of genes and cellular systems associated with HIV and aging research.

Background Knowledge Expected of the Participants

The tutorial is intended primarily for computational scientists who are interested in Biomedical Research and the impact of high performance computing in advancing Biomedical Informatics. Bio-scientists with some background in computational concepts represent another group of intended audience. Although some basic background in biomedical sciences would be useful, it is not necessary since the tutorial will provide a basic background of the needed concepts. Some basic background in algorithms would be helpful though.

Brief Bio Sketch of the Instructor

Hesham H. Ali is a Professor of Computer Science and the Lee and Wilma Seaman Distinguished Dean of the College of Information Science and Technology (IS&T), at the University of Nebraska at Omaha (UNO). He is also the director of UNO Bioinformatics Core Facility that supports a large number of biomedical research projects in Nebraska. He has published numerous articles in various IT areas including scheduling, distributed systems, wireless networks, and Bioinformatics. He has also published two books in scheduling and graph algorithms, and several book chapters in Bioinformatics. He is currently serving as the PI or Co-PI of several projects funded by NSF, NIH and Nebraska Research Initiative (NRI) in the areas of wireless networks and Bioinformatics. He has been leading a Bioinformatics Research Group at UNO that focuses on developing innovative computational approaches to identify and classify biological organisms. The research group is currently developing new graph theoretic models for assembling short reads obtained from high throughput instruments, as well as employing a novel correlation networks approach for integrating and analyzing large heterogeneous biological data associated with various biomedical research areas. He has also been leading two funded projects for developing secure wireless infrastructure and using wireless technologies to study mobility profiling for aging research.

References

1. K. Dempsey and H. Ali, “On the Discovery of Cellular Subsystems in Correlation Networks using Centrality Measures,” to appear in Current Bioinformatics, 2013.
2. J. Warnke and H. Ali, “An efficient and scalable graph modeling approach for capturing information at different levels in next generation sequencing reads,”, BMC Bioinformatics, 14: 11-S7, 2013.
3. J. Banwait, H. Ali, and D. Bastola, “Optimization of miRNA-mRNA relationship prediction using biological features,” To appear in the international Journal of Computational Biology and Drug Design, 2013.
4. O. Bonham-Carter, H. Ali, and D. Bastola, “A base composition analysis of natural patterns for the pre-processing of metagenome sequences,” BMC Bioinformatics 14:11-S5, 2013.
5. S. West, K. Dempsey, S. Bhowmick, and H. Ali, “Analysis of Incrementally Generated Clusters in Biological Networks Using Graph-Theoretic Filters and Ontology Enrichment,”, International Workshop on Incremental Clustering, Concept Drift and Novelty Detection (IclaNov), held in conjunction with the International Conference in Data Mining (ICDM 2013) Dallas December 7-10, 2013.

6. J. Warnke and H. Ali, “A Tolerance Graph Approach for Domain-Specific Assembly of Next Generation Sequencing Data,”, Proceedings of the Biological Data Mining and its Applications in Healthcare (BioDM), held in conjunction with the International Conference in Data Mining (ICDM 2013) Dallas December 7-10, 2013.
7. K. Dempsey, I. Thapa, D. Bastola, and H. Ali, “On Mining Biological Signals using Correlation Networks,” The Third International Workshop on Data Mining in Networks (DaMNet), held in conjunction with the International Conference in Data Mining (ICDM 2013) Dallas December 7-10, 2013.
8. R. Khazanchi, K. Dempsey, I. Thapa, and H. Ali, “On Identifying and Analyzing Significant Nodes in Protein-Protein Interaction Networks," The Third International Workshop on Data Mining in Networks (DaMNet), held in conjunction with the International Conference in Data Mining (ICDM 2013) Dallas December 7-10, 2013.
9. K. Dempsey, S. Bhowmick, and H. Ali. Function-preserving filters for sampling in biological networks. 2012 Int Conference on Computational Science (ICCS 2012). June 4-6, 2012: Omaha, NE.
10. K. Dempsey, K. Duraisamy, S. Bhowmick, and H. Ali. The Development of Parallel Adaptive Sampling Algorithms for Analyzing Biological Networks. 11th IEEE International Workshop on High Performance Computational Biology (HiCOMB 2012). May 21, 2012: Shanghai, China.
11. K. Dempsey, I. Thapa, D. Bastola and H. Ali, “Identifying Modular Function via Edge Annotation in Gene Correlation Networks using Gene Ontology Search,” Proceedings of the Second Workshop on Integrative Data Analysis in Systems Biology (IDASB), held in the 2011 IEEE International Conference on Bioinformatics & Biomedicine (BIBM 2011), Atlanta, Georgia, USA, Nov. 12-15, 2011.
12. K. Dempsey, H. Ali, “Evaluation of Essential Genes in Correlation Networks using Measures of Centrality. 4th Annual 2011 BIBM Workshop on Bio-molecular Network Analysis, Atlanta, Georgia, November 12-15, 2011.
13. K. Duraisamy, K. Dempsey, H. Ali and S. Bhowmick, “A Noise Reducing Sampling Approach for Uncovering Critical Properties in Large Scale Biological Networks,” Proceedings of the 2010 Workshop International Workshop on High Performance Computing Systems for Biomedical, Bioinformatics and Life Sciences (BILIS 2011), held in conjunction with The 2011 International Conference on High Performance Computing & Simulation (HPCS 2011), Istanbul, Turkey, July 4- 8, 2011.
14. K. Dempsey, K. Duraisamy, H. Ali, S. Bhowmick, “A Parallel Graph Sampling Algorithm for Analyzing Gene Correlation Networks,” Proceedings of the 11th International Conference on Computational Science (ICCS 2011), Tsukuba, Japan, June 1-3, 2011.
15. H. Geng, J. Iqbal, W. Chan, H Ali. Virtual CGH: an integrative approach to predict genetic abnormalities from gene expression microarray data applied in lymphoma *BMC Medical Genomics*, 4:32, April 2011.
16. K. Dempsey, B. Currall, R. Hallworth and H. Ali, “A New Approach for Sequence Analysis: Illustrating an Expanded Bioinformatics View through Exploring Properties of the Prestin Protein,” a book chapter in, “Handbook of Research on Computational and Systems Biology: Interdisciplinary Applications,” IGI Global, 2011.
17. K. Dempsey, S Bonasera, D. Bastola and H. Ali, “A Novel Correlation Networks Approach for the Identification of Gene Targets, Proceedings of the 44th Hawaii International Conference on System Sciences (HICSS-44), Kauai, January 4-7, 2011.
18. K. Dempsey, B. Currall, R. Hallworth and H. Ali, “An intelligent data-centric approach toward identification of conserved motifs in protein sequences,” Proceedings of the 2010 ACM

International Conference on Bioinformatics and Computational Biology (BCB 2010), Niagara Falls, New York, August 2-4, 2010.

19. R. Sengupta, D. Bastola and H. Ali, "Classification and Identification of Fungal Sequences Using Characteristic Restriction Endonuclease Cut Order," *Journal of Bioinformatics and Computational Biology*, Volume 8, Number 6, 2010.
20. S. Pawaskar and H. Ali, "A Dynamic Energy-Aware Model for Scheduling Computationally Intensive Bioinformatics Applications," Proceedings of the 2010 Workshop on Optimization Issues in Energy Efficient Distributed Systems (OPTIM 2010), held in conjunction with The 2010 International Conference on High Performance Computing & Simulation (HPCS 2010), Caen, Normandy, France, June 28- July 2, 2010.
21. D. Quest and H. Ali, "The Motif Tool Assessment Platform (MTAP) for Sequence-Based Transcription Factor Binding Site Prediction Tools," a Book Chapter in," Computational Biology of Transcription Factor Binding: Methods and Protocols," Springer, 2010.
22. H. Zhou, H. Ali, J. Youn, Z. Zhang, "A Hybrid Wired and Wireless Network Infrastructure to Improve the Productivity and Quality Care of Critical Medical Applications", the International Conference on Complex Medical Engineering (CME 2010), Gold Coast, Australia, July 2010
23. S. Vaidya, J. Youn, H. Ali, N. Bahl, and D. Singh, "Real-Time Fall Detection and Activity Recognition Using Wireless Sensors," International Conference on Networking and Information Technology (ICNIT-2010), Manila, Philippines. June 2010.
24. J. Youn, H. Ali, H. Sharif, and B. Chhetri, "RFID-Based Information System for Preventing Medical Errors," The Sixth Annual International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services, Toronto, Canada, July 2009.
25. R. Sengupta, D. Bastola and H. Ali, "Characteristic Restriction Endonuclease Cut Order for Classification and Identification of Fungal Sequences," Proceedings of the 2009 IEEE Computer Society Bioinformatics Conference (CSB 2009), Stanford University, August 10-12, 2009.
26. N. Sharma, J. Youn, N. Shrestha and H. Ali, "Direction Finding Signage System using RFID for Healthcare Applications," Proceedings of The International Conference on BioMedical Engineering and Informatics (BMEI2008), Sanya, Hainan, China, May 27-30, 2008.
27. J. Uher, D. Sadofsky, J. Youn, H. Ali, H. Sharif, J. Deogun, and S. Hinrichs, "I2MeDS: Intelligent Integrated Medical Data System," Proceedings of The International Conference on BioMedical Engineering and Informatics (BMEI2008), Sanya, Hainan, China, May 27-30, 2008.
28. P. Ciborowski and H. Ali, "Bioinformatics," a book chapter in, "Proteomics for Undergraduates," A. Kraj and J. Silberring (eds.), Wiley Inc., 2008.
29. X. Deng , H. Geng and H. Ali, "A Hidden Markov Model Approach to Predicting Yeast Gene Function from Sequential Gene Expression Data," *The International Journal of Bioinformatics Research and Applications*, 2008:4(3):263-273.
30. D. Quest, K. Dempsey, M. Shafiullah, D. Bastola, and H. Ali. MTAP: A Motif Tool Assessment Pipeline for Automated Assessment of De Novo Regulatory Motif Discovery Tool. *BMC Bioinformatics*, August 2008.
31. D. Quest, K. Dempsey, M. Shafiullah, D. Bastola, and H. Ali. A Parallel Architecture for Regulatory Motif Algorithm Assessment. *HiCOMB 2008: Seventh IEEE International Workshop on High Performance Computational Biology*, April 14th 2008.
32. X. Deng , H. Geng and H. Ali, "Cross-platform Analysis of Cancer Biomarkers: A Bayesian Network Approach to Incorporating Mass Spectrometry and Microarray Data," *Journal of Cancer Informatics*, 2007.
33. A. Sadanandam, M. Varney, L. Kinarsky, H. Ali, R. Lee Mosley, R. Singh, "Identification of Functional Cell Adhesion Molecules with a Potential Role in Metastasis by a Combination of *in*

vivo Phage Display and *in silico* Analysis," *OMICS: A Journal of Integrative Biology*, Vol. 11, No. 1: 41-57, March 2007.

34. X. Huang and H. Ali, "High Sensitivity RNA Pseudoknot Prediction," *Nucleic Acid Research*, 2007.
35. N. Sharma, J. Youn, N. Shrestha and H. Ali, "Direction Finding Signage System using RFID for Healthcare Applications," Proceedings of The International Conference on BioMedical Engineering and Informatics (BMEI 2008), Sanya, Hainan, China, May 27-30, 2008.
36. J. Uher, D. Sadofsky, J. Youn, H. Ali, H. Sharif, J. Deogun, and S. Hinrichs, "I2MeDS: Intelligent Integrated Medical Data System," Proceedings of The International Conference on BioMedical Engineering and Informatics (BMEI 2008), Sanya, Hainan, China, May 27-30, 2008.
37. H. Geng, H. Ali and J. Chan, "A Hidden Markov Model Approach for Prediction of Genomic Alterations from Gene Expression Profiling," Proceedings of the fourth International Symposium on Bioinformatics Research and Applications (ISBRA), Atlanta, Georgia, May 6-9, 2008.
38. D Quest, K. Dempsey, D. Bastola, and H. Ali. An Automated Pipeline for Regulatory Motif Tool Assessment. *Computational Systems Bioinformatics (CSB)*, August 2006.
39. H. Geng, X. Deng and H. Ali, "MPC: a Knowledge-based Framework for Clustering under Biological Constraints," *Int. J. Data Mining and Bioinformatics*, Volume 2, Number 2, 2007.
40. X. Deng, H. Geng, D. Bastola and H. Ali, "Link Test — A Statistical Method for Finding Prostate Cancer Biomarkers," *Journal of Computational Biology and Chemistry*, 2006.
41. A. Churbanov, I. Rogozine, J. Deogun, and H. Ali, "Method of Predicting Splice Sites Based on Signal Interactions," *Biology Direct*, 2006.
42. X. Deng, H. Geng, and H. Ali, "Joint Learning of Gene Functions--A Bayesian Network Model Approach". *Journal of Bioinformatics and Comp. Biology*, Vol. 4, No. 2, pp. 217-239, 2006.
43. X. Deng and H. Ali, EXAMINE, "A Computational Approach to Reconstructing Gene Regulatory Networks," *Journal of BioSystems*, 81:125-136, 2005.
44. A. Churbanov, M. Pauley, D. Quest and H. Ali, "A method of precise mRNA/DNA homology-based gene structure prediction," *BMC Bioinformatics*, 6:261, 2005.
45. A. Mohamed, D. Kuyper, P. Iwen, H. Ali, D. Bastola and S. Hinrichs, "Computational approach for the identification of Mycobacterium species using the internal transcribed spacer-1 region," *Journal of Clinical Microbiology*, Vol. 43, No. 8: 3811-3817, 2005.
46. A. Churbanov, I. Rogozin, V. Babenko, H. Ali and E. Koonin, Evolutionary conservation suggests a regulatory function of AUG triplets in 5'UTRs of eukaryotic genes, *Nucleic Acid Research*, 33(17), pp. 5512-20, Sep 2005.
47. H. Geng, X. Deng and H. Ali, "A New Clustering Algorithm Using Message Passing and its Applications in Analyzing Microarray Data," The Fourth International Conference on Machine Learning and Applications (ICMLA'05), pp. 145-150, 2005.