Data Analysis and Integration Tools in Biomedical Informatics: A Case Study in Aging Research

Hesham H. Ali
Department of Computer Science
UNO Bioinformatics Core Facility
College of Information Science and Technology
University of Nebraska at Omaha
Omaha, NE 68182-0116, USA
hesham@unomaha.edu

Abstract

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. This ranges from the effective storage of data and their associated data models, to the design of efficient algorithms to automate the data mining procedures, and also to the development of advanced software systems to support data integration. With more researchers taking on Bioinformatics projects that integrate theoretical and applied concepts from both Bioscience as well as Computational Sciences, Biomedical informatics is quickly emerging as the most exciting field of research in this century. In this tutorial, we present an overview of the state of discipline for Biomedical Informatics with a focus on the nature and diverse of the available data as well as data collection tools. We make a case for the need for smarter and more advanced data integration and data analysis tools. Such tools are desperately needed to connect the datasets and obtain useful information that can be used for better medical discoveries and patient care. We present examples of recently developed intelligent tools and expert systems that produced exciting results that could not have been obtained without such innovative integration. We then focus on a case study in aging research to illustrate the proposed integration and analysis tools.

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 the information stored in the available databases to extract new information as well as to understand various biological and medical phenomena.

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. This is particularly critical in managing biological and clinical data since relevant data is

available in different shapes and forms, and hence, employing all available data to extract meaningful properties is an enormous task. Heterogeneous data, obtained from microarrays, high throughput sequencers, mass spectrometry experiments and clinical records, can all be used to find potential correlations between genes/proteins and the susceptibility to have a particular disease. The proposed tutorial will address these issues with a particular focus on the following objectives:

1- Provide an overview of the exciting disciplines of Biomedical Informatics, including medical, public health and bio informatics with a focus on the interdisciplinary nature of these fields of study.

2- Introduce the main computational problems in biomedical research with a focus on data collecting and analysis related problems, then survey the current available algorithmic tools and address the advantages and the shortcoming of each tool.

3- Introduce the audience to the concept of intelligent data integrating and analysis tools. Such tools are critical to leverage data collected from different resources to produce useful information that can further advance biomedical research and has the potential lead to new discoveries directly related to patient care.

Background Knowledge Expected of the Participants

The tutorial is intended for bio-scientists and computational scientists who are interested in Biomedical Research and how to develop or use computational tools to solve data mining related problems. Although some basic background in biomedical sciences would be helpful, it is not necessary since the tutorial will provide a basic background of the needed concepts. Similarly, some basic background in algorithms would be useful but it is not necessary.

Topics to be Covered in the Proposed 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 1-6 below. The second part focuses on introducing the audience to new research projects with a focus on the concept of next generation data analysis and integration tools; that are Intelligent, Collaborative and Dynamic (ICD). Few examples of such tools will be discussed briefly, and then a focus on Aging Research will be covered in more details. In particular, studies related to mobility research and how mobility parameters can used to detect and predicts potential medical problems.. This part is covered in points 7-10 below.

1. Introduction to Biomedical Informatics
2. Brief discussion on the various aspects of Biomedical Informatics that include Bioinformatics, Medical Informatics, Public Health Informatics, and Biomedical Imaging.
3. Background – The Bioscience aspect and the computational perspective

4. Biomedical Informatics now – current state of the emerging discipline and overview of key Biomedical Research problems
5. Overview of selected current, first generation, data analysis tools
6. The need for next generation data integration and analysis tools; Intelligent, Collaborative and Dynamic (ICD) Tools.
7. Example of new data integration and analysis tools.
8. Case Study: Biomedical Informatics on Aging Research:
    - a) Mobility profiling and its application in predicting medical problems.
    - b) Correlation Networks and the identification of genes associated with aging.
9. Case Study 6: Intelligent Integrated Medical Data System (I2MeDS)
10. Translational Research and next steps in Biomedical Informatics

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,” Current Bioinformatics, 2012.
2. 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.
3. 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.

4. 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.
5. 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.
6. 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.
7. 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.
8. 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.
9. 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.
10. 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.
11. 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.
12. 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.
13. 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.
14. 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
15. 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.
16. 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.

17. 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.
18. 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.
19. 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.
20. P. Ciborowski and H. Ali, "Bioinformatics," a book chapter in, "Proteomics for Undergraduates," A. Kraj and J. Silberring (eds.), Wiley Inc., 2008.
21. 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.
22. 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.
23. 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.
24. 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.
25. 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.
26. X. Huang and H. Ali, "High Sensitivity RNA Pseudoknot Prediction," *Nucleic Acid Research*, 2007.
27. 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.
28. 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.
29. 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.
30. D Quest, K. Dempsey, D. Bastola, and H. Ali. An Automated Pipeline for Regulatory Motif Tool Assessment. *Computational Systems Bioinformatics (CSB)*, August 2006.
31. 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.

32. 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.
33. A. Churbanov, I. Rogozine, J. Deogun, and H. Ali, “Method of Predicting Splice Sites Based on Signal Interactions,” *Biology Direct*, 2006.
34. 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.
35. X. Deng and H. Ali, EXAMINE, “A Computational Approach to Reconstructing Gene Regulatory Networks,” *Journal of BioSystems*, 81:125-136, 2005.
36. 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.
37. 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.
38. 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.
39. 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.