International Journal of Bioinformatics Research, ISSN: 0975–3087, Volume 1, Issue 2, 2009, pp-81-84

Bioinformatics is a key life science R & D activity

Srinivasa Rao V.*[1], Das S. K.[2], Nageswara Rao K.[3] and Kusuma Kumari E.[4]

[*1,3]Computer Science and Engineering, PVP Siddhartha Institute oOf Technology, Vijayawada, India, akrgvsr@gmail.com
[2] Department of Computer Science, Berhampur University, Berhampur, Orissa, India
[4] Electronics and Communication Engineering, Nova College of Engineering, Jangareddygudem, India

Abstract- Rapid advances in technologies like genomic as well as bioinformatics coupled with a unique collaboration between industry and academia are beginning to show the true potential for the human genome project to affect patient healthcare. By knowing the sequence of the human genome and beginning to unravel the location and sequence of all genes and their variants, scientists can establish a better understanding of the mechanisms for diseases, with subsequent availability of new treatments. Because of the vast amount of data coming out of the Human Genome Project, bioinformatics tools and databases have become an integral part of pharmacogenomic and disease susceptibility gene research.

Key words- bioinformatics, clinical, biomarker

Introduction

Bioinformatics play an important role in candidate gene identification, gene finding, SNP detection, genotyping and genetic analysis. Public sources of databases and tools abound, although it is sometimes difficult to determine the quality, consistency and sustainability of these sources. The data-management challenges arising from this heady sampling of the genome were making a strong impression, in both the public and private sectors, and the as-yet-unresolved (and highly charged) question of the patent ability of genes led to a land rush on intellectual property [27]. Bioinformatics data integration and tool standardization are critical to the success of association and linkage studies. The underlying data models accommodate the variability inherent in subject collections, the ability to trace the data source, and the automation and archival storage of analysis results. A fully traceable data source is important, as we are often faced with anomalies in data at a late stage that can be very time consuming to resolve in an infrastructure that does not facilitate data integration. The polymorphism database component includes data from public and proprietary sources. The subject phenotypes (a relevant measure of disease severity, disease progression and/or disease sub classification for disease genetics or a relevant measure of drug response for pharmacogenetics) and genotype components are fully integrated with the source databases. The subject database component also includes reference collections and allele frequency information needed for analysis. This model has proven useful in analyzing reasonably large datasets. The model is scalable to variations in volumes and expandable to accommodate a variety of markers. The performance for very high volumes (e.g. genome wide scans of a large population) is currently being investigated. SNPs are the most common markers for disease-gene and drug-response associations [2]. However, to detect association at a SNP near a complex disease gene, the appropriate SNPs must be chosen for analysis. In addition, the order and relationship of SNP markers is extremely important. The cost of doing high-density genome-wide association scans is still quite high, so, using a haplotype-based SNP map would maximize the information content and reduce the resource needs. The use of haplotypes has been discussed in great detail, including their benefits and limitations [45]. One limitation of haplotypes that needs to be considered is the fact that frequencies of most clinically significant AEs are low (< 5–10%) so the use of commonly occurring haplotypes (those with frequencies of at least 10%) may overlook important genetic associations [29]. Another approach that has been advocated to reduce the cost of genotyping is DNA pooling. Instead of analyzing SNPs from individual subjects, DNA from responders is pooled and compared with pooled DNA from control subjects. The advantages and disadvantages of this approach are reviewed in detail elsewhere [10].

Disease genetics and pharmacogenetics

Genotypic data can be combined with accurate phenotypic data and analyzed to determine the SNPs and/or haplotypes associated with disease susceptibility and/or drug response. A high-density genome association scan can be used to thoroughly evaluate the genes that modify a patient's response to medications (i. e. pharmacogenetics) and to push the limits of disease gene identification in appropriate populations (i.e. disease genetics). Examples of the use of the candidate gene approach and/or the whole genome scan approach are described below as they relate to disease genetics and pharmacogenetics.

Disease genetics

In the past, disease genetics has focused on monogenic diseases such as Huntington's disease in which the expression of a particular variant of a single gene will, in the vast majority of cases, lead to disease. There are innumerable monogenic diseases, each of which affects only a small number of patients. In contrast, disease genetics research is now focused on identification of genes associated with common diseases (diseases affecting thousands or millions of people). These common diseases are multifactorial [i.e. dependent on complex interactions between numerous environmental factors and a number of alternative forms (alleles) of genes called disease susceptibility genes] and polygenic [35] . The overall goal of disease genetics is to identify how genetic variation can influence disease susceptibility and to improve our understanding of the molecular processes resulting in clinically overt disease. New treatments can then be designed to target these molecular processes to prevent and/or treat the disease. Typically, new disease susceptibility genes have been identified using a combination of linkage and association studies. The linkage studies involve collection of DNA samples and extensive clinical phenotypic data from multiple members of affected families. Markers are typed throughout the genome, and, using linkage analysis algorithms, chromosomal regions harboring disease genes are identified [36]. The regions are identified using highly informative markers on the basis of their chromosomal location by taking advantage of the meiotic process of recombination as apparent in families segregating for the disease [28]. Markers closest to the disease gene show the strongest correlation with disease patterns in families. These linkage studies allow identification of a region on a chromosome and large portions (1–20 cM) of the DNA (which may include 10–1000 genes) that may be linked to a specific disease. Candidate genes within the region can sometimes be inferred from the genome-wide databases that are currently available. Unfortunately, most of the few validated disease genes were not obvious candidates. Association studies are then conducted to identify the causative mutation responsible for the disease either using family-based association studies or unrelated case-control association studies. The key to success for linkage and association studies is the availability of high quality clinical information, available appropriate genotypic data and the ability to link such data (see above). Linkage and/or association studies have been reported to identify susceptibility genes for many therapeutic areas. The potential benefits of the human genome project are beginning to be realized with the availability of technology advances and bioinformatics tools. The identification of disease susceptibility genes and the development of many new treatments are the longer-term benefits. In the shorter term, the benefits will be the ability to predict those patients at risk for experiencing adverse reactions or patients with a high probability of experiencing improved efficacy (i.e. pharmacogenetics). As progress is made in the area of disease genetics and pharmacogenetics, our understanding of disease susceptibility and its interrelationship with drug response will improve, making targeted therapy (i. e. the right drug to the right patient) a reality.

Bioinformatics of proteomics for biomarker development

Mass spectrometry represents an important set of technologies for protein expression measurement. Among them, surface-enhanced laser desorption/ionization time-of-flight mass spectrometry (SELDI TOF-MS), because of its high throughput and on-chip sample processing capability, has become a popular tool for clinical proteomics. Bioinformatics plays a critical role in the analysis of SELDI data, and therefore, it is important to understand the issues associated with the analysis of clinical proteomic data [9]. Ball [17] used a model system to establish whether artificial neural networks could rapidly identify molecular ions of potential interst from a total data set of 100-120 000 data points derived from SELDI mass spectrometry data and they suggested that application of bioinformatic approach to larger cohorts of patient material could lead to identification of whose relative intensity profile accurately correlate to clinical parameters such as tumor staging and possibly events predicting patient responsiveness to particular forms of therapy.

On the basis of surface-enhanced laser desorption/ionization time-of-flight mass spectrometry (SELDI-TOF MS), Ciphergen's proteinchip system offers a single, unified, and high throughput platform for a multitude of proteomic research applications. Hu [22] developed and evaluated a proteomics approach to searching for new biomarkers and building diagnostic models. SELDI-TOF-MS Protein Chip was used to detect the serum protein patterns of 49 breast cancer patients, 51 patients with benign breast diseases, and 33 healthy women. The diagnostic models were developed and validated using bioinformatics tools such as artificial neural networks and discriminant analysis. Surface-enhanced laser desorption time of flight mass spectrometry (SELDI-TOF-MS) is an important proteomic technology that is immediately available for the high throughput analysis of complex protein samples. Over the last few years, several studies have demonstrated that comparative protein profiling using SELDI-TOF-MS breaks new ground in diagnostic protein analysis particularly with regard to the identification of novel biomarkers. Importantly, researchers have acquired a better understanding also of the limitations of this technology and various pitfalls in biomarker discovery. Bearing these in mind, great emphasis must be placed on the development of rigorous standards and quality control procedures for the pre-analytical as well as the analytical phase and subsequent bioinformatics applied to analysis of the data. To avoid the risk of false-significant results studies must be designed carefully and control groups accurately selected. In addition, appropriate tools, already established for analysis of highly complex microarray data, need to be applied to protein profiling data. To validate the significance of any candidate biomarker derived from pilot studies in appropriately designed prospective multi-center studies is mandatory; reproducibility of the clinical results must be shown over time and in different diagnostic settings. SELDI-TOF-MS-based studies that are in compliance with these requirements are now required; only a few have been published so far. In the meantime, further evaluation and optimization of both technique and marker validation strategies are called for before MS-based proteomic algorithms can be translated into routine laboratory testing [26].

Bioinformatics for clinical decision support systems

One of the most promising areas in bioinformatics is computer-aided diagnosis, where a computer system is capable of imitating human reasoning ability and provides diagnoses with an accuracy approaching that of expert professionals. This type of system could be an alternative tool for assisting dental students to overcome the difficulties of the oral pathology learning process. Borra [7] developed an open decision-support system based on Bayes' theorem connected to a relational database using the C++ programming language; developed software was tested in the computerization of a surgical pathology service and in simulating the diagnosis of 43 known cases of oral bone disease. The simulation was performed after the system was initially filled with data from 401 cases of oral bone disease. The system allowed the authors to construct and to manage a pathology database, and to simulate diagnoses using the variables from the database. The integration of patient-specific genomic information into the electronic medical record (EMR) will create many opportunities to improve patient care. Key to the successful incorporation of genomic information into the EMR will be the development of laboratory information systems capable of appropriately formatting molecular diagnostic and cytogenetic findings in the EMR. Due to the lack of granular genomics-related content in existing medical vocabularies, the adoption of new standards for describing clinically significant genomic information will be an important step toward recognizing the genome-enabled EMR [21]. Appropriate capture of patient-specific genomic results in the EMR will generate new opportunities to utilize this information in clinical decision support, including automated response to pharmacogenomic - based risks.

Conclusion

Recognizing the importance of the information technology for pursuing advanced research in modern biology and biotechnology, a bioinformatics programme, envisaged as a distributed database and network organization Distributed Information Centers located in universities and research institutions are fully engaged in R&D task. The computer communication network, linking all the bioinformatics centers, is playing a vital role in the success of the bioinformatics R&D research development. Database development, R&D activities in bioinformatics, human resource development and a variety of services in support of biotechnology R&D programmes and projects, has made the programme very popular and useful to the scientific community.

References

[1] Anna T. (2006) *FEBS Let.*580 2928-2934.

[2] Apweiler R., et al., (2004) *Nucleic Acids Res* 32, pp. 115–119.

[3] Argraves G.L., Jani S., Barth J.L., Argraves W.S. (2005) *BMC Bioinformatics*,6:287.

[4] Baker P.G., Goble C.A., Bechhofer S., Paton N.W., Stevens R. and Brass A. (1999) *Bioinformatics*, 15: 510-520

[5] Benson et al., (2004) *Nucleic Acids Res*,32:23–6.

[6] Berman et al., (2000) *Nucleic Acids Res,* 28 pp. 235–242.

[7] Borra R.C., Andrade P.M., Corrêa L., Novelli M.D. (2007) *Eur J Dent Educ,*1:87-92.

[8] Brown P.O., and Botstein D. (1999) *Nat Genet 21 (Suppl 1),* 33–37.

[9] Nicole White C.,' Daniel W. Chan and Zhen Zhang (2004) *Clin Biochem 37,* 636-641.

[10] Chanock S. (2001) Disease Markers 17: 89–98.

[11] Charles D. (2006) *Pharmacology & Therapeutics*, 12, 677-700.

[12] Debouck C. and P.N. Goodfellow (1999) *Nat Genet 21 (Suppl 1),* 48–49.

[13] Eysenbach G.,Thomas L Diepgen (1998) *BMJ* , 317:1496-1502.

[14] Felsenstein J. (1989) *Cladistics* 5:164–6.

[15] Fischer H.P. (2005) *Biotechnol. Annual Rev*,1-68.

[16] Friedman C.P., et al., (2004) *J Am Med Informatics Assoc* 11 (3), 167–172.

[17] Ball G., Mian S., Holding F., Allibone R. O., Lowe J., Ali S., Li G., McCardle S., Ellis I. O., Creaser C. and Rees R. C. (2002) Bioinformatics, 18 (3), 395-404.

[18] Giannadakis N., Rowe A. , Ghanem M. and Guo Y. (2003) *Inform Sci—Inform Comput Sci: An Int J* 155, 199–226.

[19] Guidance for Industry and FDA Staff: Class II Special Controls Guidance Document: Drug Metabolizing Enzyme Genotyping System. Available from: www.fda.gov/cdra

[20] Heller M.J. (2002) *Annu Rev Biomed Eng* 4 : 129–153

[21] Hoffman M.A. (2007) *J Biomed Inform.*, 40(1):44-6.

[22] Hu, Suzhan Zhang, , Jiekai Yu, Jian Liu and Shu Zheng (2005) *The Breast*, 14: 250-255

[23] Jeanette J. McCarthy and Rolf Hilfiker (2000) *Nat. Biotechnol* .,18, 505 – 508.

[24] Kane M.D., Jeffrey L. Brewer (2007) *J. Biomed. Informatics*, 40 67-72.

[25] Kanehisa M. and Goto S. (2000) *Nucleic Acids Res* 28, 27–30.

[26] Kiehntopf M., Siegmund R., Deufel T. (2007) *Clin Chem Lab Med*, 45:1435-49

[27] Kiley T.D. (1992) *Science,* 257:915–918.

[28] Kruglyak L. (1999) *Nat. Genetics* 22: 139 – 144.

[29] Lai E., et al., (2002) *Nat. Genet*. 32: 353.

[30] Lander et al., (2001) *Nature* 409 (6822), 860–921.

[31] Leo P.M. et al., (2004) *Bioinformatics Italian Society Meeting (BITS 2004) Padova*.

[32] Lynne Neufeld and Martha Cornog (1999) *J. of the American Society for Information Science* 37: 183 – 190.

[33] Madden T.L., Tatusov R.L., Zhang J. (1996) *Methods Enzymol.*, 266: 131-41.

[34] Marth et al., (2001) *Nat. Genet*. 27: 371–372

[35] Middleton et al., (2000) *Community Genetics* 3: 198–203

[36] Monica stoll et al., (2000) *Genome Res.*, 10, 473-482.

[37] Neil Kumar et al., (2006) Drug. Dis. Today. 11, 806-811.

[38] Oinn, et al., (2004) *Bioinformatics* 20, 3045–3054.

[39] Rice P., Longden I. and Bleasby A. (2000) *Trends Genet* ,16, 276–277.

[40] Rolf Backofen and David Gilbert (2001) Constraints, 6: 141-156.

[41] Sackett D., Rosenberg W., Gray J., et al., (1996) *BMJ*, 312: 71-72.

[42] Sanober Shaikh, Robert W. Kerwin (2002) *British Journal of Clinical Pharmacology* 54 (4), 344–348.

[43] Shayne Cox Gad (2005) *Drug Discovery Handbook, Wiley-Interscience*, ISBN 0-71-21384-5.

[44] Stajich J.E. and Lapp H. (2006) *Brief Bioinform.*, 7:287-96.

[45] Stephens J. C. (1999) *Molecular Diagnosis* 4: 309–317.

[46] Tang et al., (2005) *BMC Bioinformat* 6 p. 69.

Biography

Dr. V. Srinivasa Rao received the degree M. Tech in Computer Science and Technology from Andhra University. He received the Ph.D. degree in Computer Science and Engineering from the Berhampur University. Currently, he is a Professor at PVP Siddhartha Institute of Technology, Vijayawada, A.P, India

Dr. S. K. Das presently working in the department of Computer Science, Berhampur University, Orissa, India.

Dr. K. Nageswara Rao received the M.Tech degree in Computer Science and Engineering from the Andhra University. He received the Ph.D. degree in Computer Science and Engineering from Andhra University. Currently, he is a Professor and Head at PVP Siddhartha Institute of Technology, Vijayawada, India.

Smt. E. Kusuma Kumari received the degree M. Tech in Electronics and Instrumentation Engineering from JNT University Hyderabad. Currently she is a Associate Professor in ECE Dept at Nova College of Engineering, Jangareddygudem, India. Her research interests are Antenna communications related and in Bioinformatics Instrumentation. Now she is doing Ph.D in Electronics and Communication Engineering.