INCORPORATING BIOINFORMATICS INTO BIOLOGICAL SCIENCE IN NEPAL: PROSPECTS AND CHALLENGES

Umesh Prasad Shrivastava

Department of Botany, T.U., TRM Campus. Birgunj

upshrivastava@gmail.com

Abstract

The huge amount of data created by proteomics and genomics studies worldwide has caused bioinformatics to gain prominence and importance for urgency to process and analyze those data. However, its multidisciplinary nature has created a challenge to meet the unique demand for specialist trained in both biology and computing. Several countries, in response to this challenge, have developed a number of manpower training programs. This review presents a description of the meaning, scope, history and development of bioinformatics with focus on prospects and challenges facing bioinformatics education worldwide. The paper also provides an overview of attempts at the introduction of bioinformatics; describes the existing bioinformatics scenario and suggests strategies for effective bioinformatics education for the sustainable growth and development in Nepal.

Key words

Bioinformatics, computational biology, challenges, strategies, propectes.

Introduction

Bioinformatics has emerged as an important discipline within the biological sciences that allows scientists to decipher and manage the vast quantities of data (such as genome sequences) that are now available. The ultimate goal of biological researches is to develop knowledge that will allow predictive approaches to all life sciences (Allen, 2004). In this regards, the data acquisition, management and analysis techniques of bioinformatics are key, as they provide a lens through which accumulated large-scale data set and knowledge can be viewed. The ultimate goal of bioinformatics is to present a complete representation of the cell and the

organism and to computationally predict systems of higher complexity, such as the interaction networks in cellular processes and the phenotypes of whole organism. Bioinformatics approaches have been found useful in several research areas which include disease detection, control and diagnosis, drug discovery and development, genome informatics, epidemiology, biomedical imaging and ecosystems modeling. A causeless combination of factors makes bioinformatics of particular interest for researchers worldwide (Degrave et al., 2001).

This wide applicability of bioinformatics can be described by following points:

- Computational approaches offer novel insights into pathogenic organisms, disease detection, control and diagnosis, epidemiology, genome informatics and ecological modeling of local relevance.

- Bioinformatics takes advantage on training in computer science, an area in which many developing countries including Nepal have made considerable investments.

- The necessary technology, including hardware (inexpensive personal computers), and support (e.g. for systems administration) is readily available worldwide.

- Computational biology has lower population density in the area of research. The demand for scholar of bioinformatics far outstrips supply, and training in this area will clearly be sustainable for national development.

- Computational biology is less affected by infrastructure and economics compared to other areas of biological research, providing an unusually level playing field.

Consequently, there is an obvious need to provide graduates in biosciences with generic, transferable skills in bioinformatics. This is responsible for various concerted efforts worldwide to standardize bioinformatics education through the development of appropriate curricula, incorporation of bioinformatics in undergraduate biological science education and organization of various train-the-trainers workshops and short courses. The skills in bioinformatics can enhance the economy of the nation with less capital serving worldwide situating in any corner of the world.

Meaning and scope of bioinformatics

Bioinformatics apply the principles of information sciences and technologies to complex life science data. Both bioinformatics and computational biology grew out of the traditional fields of health informatics, biostatistics, information sciences, mathematics, computer sciences and related engineering fields, molecular biology and genetics, and behavioral science (Allen, 2004). Moreover, the National Institutes of Health (NIH), USA defines bioinformatics as research, development, or application of computational tools and approaches for expanding the use of biological, medical, behavioral, or health data, including those to acquire, store, organize, archive, analyze, or visualize such data (National Institute of Health, 2000). NIH further describes computational biology as the development and application of data-analytical and theoretical methods, mathematical modeling, and computational simulation techniques to the study of biological, behavioral and social systems.

History and development of bioinformatics

The first microbial genome sequence was published in 1995 has seen an explosion in biological sequence data (Fleischmann et al., 1995; Sutcliffe and Hutchings, 2007).

For example, as at July 2007, the GOLD database (Liolios et al., 2006) lists 626 completed published genome projects and 2129 ongoing genome projects (excluding metagenomes). This vast volume of data (and those derived from the post-genomic technologies that have subsequently emerged, such as microarray technologies) represents both a wonderful resource but also a considerable and rapidly growing challenge to the biological scientists.As noted in a recent editorial (Delpech, 2006), this information is not as such knowledge and so the discipline of bioinformatics has emerged to aid in the collation, presentation and deciphering of this information. Moreover, many primary data publications contain some degree of bioinformatics content that must be critically evaluated (for example, sequence alignments and phylogenetic trees). It is thus increasingly important that pedagogy in the biological sciences keeps abreast of developments in this area and that students are educated in the handling and critical evaluation of such data.

According to Allen (2004), the recent evolution in bioinformatics as a recognized field resulted from technological developments in both biology and information technologies. It was further reported that the 1980s saw a growth of high-throughput techniques such as a high-speed nucleic acid sequencing and polymerase chain reaction that increased, geometrically, the volume of data generated by experiments in biology (Brass, 2000). Similarly, recent development in information technology has been reported to include new algorithms and methods of data retrieval and analysis, including tools to analyze data and technologies that facilitate efficient storage and retrieval of large data sets.

The growth of bioinformatics globally could also be attributed to the development of the Internet which enable creation and shared access to large repositories of biological data and offer a vehicle for rapid publication of research reports (Degrave et al., 2002). Early bioinformatics activities were also reported to involve mainly the analysis of single biological entities such as DNA and protein sequences or analysis of three-dimensional protein structure. However, contemporary approaches include methods for analysis of large numbers of genes and proteins simultaneously, such as the identification of clusters of related genes and networks of interacting proteins. Increase in the number of available complete genome sequences has expanded the scope of bioinformatics to providing conceptual and practical methods for detecting systemic functional behaviors of the cell and the organism (Kanehisa and Bork, 2003).

Prospects of bioinformatics

There is no doubt that the field of bioinformatics, or computational biology is an emerging discipline with many opportunities for future advancement in both pure and applied research, now these days. Computational biology/bioinformatics, as a new discipline, is an under-populated area of research. The demand for bioinformaticians globally far surpasses supply and training in this area (Allen, 2004). According to Hemminger and Anne-Bauers (2005) communication and other research interactions are based on computer networking, 'critical mass' issues are less critical: when it comes to bioinformatics, a world- wide community may be within reach. It is realized that equipping and running a bioinformatics teaching laboratory costs much less than equipping and running a molecular biology laboratory Counsell (2003). Most bioinformatics software is available for free to academics, as are many Internet- based bioinformatics services. The data from the public genome projects is also free to anyone with Internet access

According to Ranganathan (2005) for participating computational biology, the

necessary equipment costs can be relatively modest – a moderately powerful computer and periodic, low- speed Internet access. He further stated that the Linux operating system permits use of the personal computer as a powerful workstation, with access to a vast repository of public domain software for computational biology and other applications. Really, computational biology is foremost a people – and technology – intensive rather than capital-intensive discipline. It is thus less affected by infrastructure and economics than other areas of biological research, providing an unusually level playing field (Hack and Kendall, 2005). Overall, the requirements for a successful program include relatively inexpensive computers, periodic net access and well-trained, creative, hard-working scientists.

Advantages of capacity for cutting-edge research in bioinformatics and computational biology for a developing economy like Nepal are enormous. The onset of the genomics revolution in the late 80s globally raised high hopes for the development of applications in the fields of human, animal, plant, bacterial and viral genetics. According to Degrave et al. (2002), new bioinformatics tools targeting pathogens and their vectors have increased our understanding of the evolutionary processes and the delicate interplay of living beings with their environment. We can expect an immediate technological impact, not only on new diagnostics, therapeutics, and vaccine development, but also on our understanding of infectious disease mechanisms, pathogen–host interactions and transmission cycles. The international biomedical community, including scientists from disease-endemic countries (DECs), are increasingly taking advantage of the power of computing to analyze genomic data – not only because the quantity of raw data requires such approaches, but also because computational analysis allows the exploration of data in completely new ways. It is increasingly possible to carry out 'hypothesis-driven' research sitting at the computer.

Challenges facing bioinformatics education

It has long been recognized that bioinformatics/computational biology is a multidisciplinary field, operating at the interface between biology and computer science (Wickware, 2001; Friedman, 2004; Fetrow and John, 2006). World-wide, cooperation between professionals from the biological and computational 'sides' has not been as unstable as one might hope, as these disciplines use different languages to communicate. It can therefore take considerable time and effort to establish harmonious and fruitful interactions based on mutual trust, respect and understanding (Boyer, 2000). This problem is particularly acute in developing countries lacking a strong tradition of interdisciplinary collaboration.

The different multidisciplinary areas within bioinformatics relate to biological molecules and therefore require knowledge in the fundamentals of biochemistry, molecular cell biology, genetics, thermodynamics, biophysics, and statistical mechanics. According to Pearson (2001) and Campbell (2003) students are required to apply analytical capabilities obtained from their knowledge of computer science, mathematics, and statistical principles to sift the deluge of data produced by genomic and proteomic studies. Fetrow and John (2006) reported that the multidisciplinary nature does not only transcend the established areas of sciences, it also forces integration of knowledge and cross-field utilization of techniques where researchers use it interchangeably.

Consequently, teaching bioinformatics requires a specialist educator with in-depth knowledge of all the different components: mathematics, biology and computer science (Wickware, 2001). This is quite a dashing task, therefore most universities, lacking specialist and experience bioinformatics staffs resort to interdisciplinary and cross faculty teaching. This is logical route as utilizing experts in their

respective areas will not only ensure a well-taught course but also teach the necessary breadth and depth.

However, inter-faculty teaching raises the issue of 'ownership' and the placement of the course in an institution. The management has to determine which faculty or department to house the course, facilities to use, in biology, mathematics or computer science departments. Consequently, this causes concerns on the teaching and learning culture, as it is obvious each discipline has its own inherent culture. Though teaching and research in bioinformatics has been reported to be less cost-intensive compared to other biological disciplines, in most developing countries, even minor costs can be an impediment. It is also recognized that there is often a serious need for maintenance and support of computing hardware, software and networks. Funds to expedite network upgrading and/or the financing of dedicated lines are often an excellent investment.

Pearson (2001) and Pevzner (2004) advocate another challenge in the area of teaching bioinformatics is the depth of the topic. Their report indicated that broad introduction to bioinformatics without the necessary depth will produce bioinformatics technicians rather than bioinformatics scientists. The authors also highlighted the importance of teaching principles of algorithm, statistics and creating a biologically motivated problems-based learning in order to effectively teach bioinformatics.

Most biologists are comfortable in using software such as BLAST and are contented simply by either finding a match or not, without even understanding the underlying principles behind it (Kumar, 2005). This treatment of bioinformatics merely as computational tools is prone to erroneous assumptions if derived from flawed understanding of the algorithm behind the tools. This is further compounded by the rise in publication of bioinformatics textbooks that are cook-book

styled and protocol-centric. Hence, failure to produce a course with the necessary depth will produce students severely lacking in skills for pursuing careers in bioinformatics. The necessary depth of the curricula is achievable and commensurate to an undergraduate course time span (3 or 4 years). It has been suggested that this can be achieved by paring down the non- related biological subjects and focusing on cellular molecular biology, genetics and gene expression studies (Maojo and Kulikowski, 2003).

Bioinformatics education in Nepal

Like other modern biological science disciplines, bioinformatics is new to Nepal. The pioneer university in Nepal is Tribhuvan University, now these days many other universities are also working; presently no any University runs bioinformatics as a independent discipline either at undergraduate or postgraduate levels. Whereas, in my knowledge, Central Department of Botany, Tribhuvan University now has initiated bioinformatics as a part of the curriculum of M. Sc. of Botany from last year. Perhaps, a few universities in Nepal have introduced undergraduate course in biotechnology either as a sole subject or in combination with biochemistry. The present bioinformatics status by search across database for Nepal can be viewed in Fig 1, which shows a very poor condition of the database. It is evident with the fig. 1 that there is no any data of Expressed Sequence Tag record (EST), Genome Survey Sequence record (GSS), structure, taxonomy, genome structure variation (dbVar), Pathways and systems of interacting molecules (BioSystems), eukaryotic homology groups (HomoloGene), gene-oriented clusters of transcript sequence (UniGene), conserved protein domain database (CDD), integrated data for clone resources (Clone), marker and mapping data (UniSTS), Epigenetic maps and dataset (Epigenomics), and a collection of related protein sequences (Protein clusters). It is realized that attempts to introduce bioinformatics are the present demand
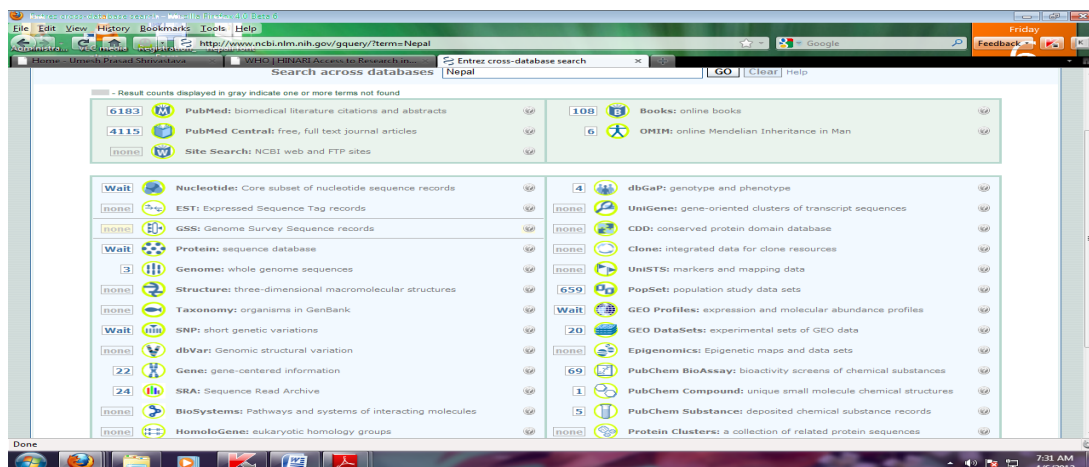
Fig 1: Search cross database for Nepal retrieved on 6th April, 2012.

of Nepal. Various central departments of Tribhuvan University and other campuses such as TRM Campus, Birgunj, should start bioinformatics. Kathmandu University has potential to initiate this discipline also. Short term training or workshop should be arranged for the meaningful competence for research and teaching in bioinformatics. Effective teaching and learning of bioinformatics in Nepal is predominantly affected by shortage of qualified and properly trained trainers, lack of funds and unavailability of basic infrastructure for bioinformatics training and research. These infrastructures include adequate power supply (or alternative power supply), Internet connectivity and computer systems. While the problem of lack of computer systems is gradually fizzling away, acquisition and maintenance of fast and effective Internet connection and provision of adequate power supply is still a challenge in Nepal. It is obvious that load shading in power supply is a challenging task now these days in Nepal, which can be fulfilled by using alternate sources.

Moreover, the few bioinformatics and computational biology experts available in Nepal work do not work coherently with each other. This lack of network has negatively impact rapid development of the discipline and the quality and quantity of research in bioinformatics in Nepal.

Strategies for effective bioinformatics education in Nepal

Provision of adequate infrastructure and training the trainers are the primary areas requiring critical attention. All over the world, there is already considerable activity in terms of ongoing regular training on bioinformatics and Nepal is lagging behind in this regards. In some advanced countries, the introduction of bioinformatics as an undergraduate programme has been advocated (Doom et al., 2003; Burhans and Skuse, 2004). However, since the problem in most cases is country specific, proffered solutions should be country specific as well.

For Nepal, both long and short term approaches to human-capacity development in bioinformatics are advocated here. On a short term basis, it is important to support the training courses that provide skills and training to life scientists on one hand and computer scientists on the other hand. This could be in form of train-the-trainer and multiplication courses using standardized training materials and tools. This is expected to sensitize scientists in

developing economies like Nepal to the potential of computational biology and provide a screening mechanism to identify gifted researchers for further trainings. On the long run, more training at masters, doctoral and postdoctoral levels should be made a priority. It is expected that the cost of such trainings might be high, since such trainings may not be available locally. However, possibilities for cross registration could be used with students registering locally but receiving some of their training at more advanced institutions.

In order to encourage local computational biology training programmes, and help develop networks and/or databases in developing economies like Nepal, it has been advocated that it is important to establish locally appropriate strategies such as the development of instructional web-based, distant-learning courses. It has been suggested that such networks would aid in disseminating course curriculum materials, and promote and support undergraduate and Master's-level training programmes in such disciplines as molecular genetics, applied computer science and statistics which are all important aspects of a research career in computational biology. It is also important that a National Biotechnology and Bioinformatics Network be established to enhance research activities in bioinformatics and biotechnology, and accelerate productivity. Malaysia conceptualized and introduced a similar network in 1997 with the aim to give R & D support to researchers to conduct biotechnology and bioinformatics researches. It also allows sharing of resources and usage of a dedicated bioinformatics architecture which eventually improve research in Malaysia (Pongor and Landsman, 1999). A similar network is available in India and South Africa. It is also imperative that Nepal identifies with Indian Bioinformatics Network and the International Conference

on Bioinformatics that Nepal can achieve long term sustainable database for national progress. Consequently, the database for rich biodiversity of Nepal can be made by the trained bioinformatics scientists after such type of training and knowledge.

Conclusion

In this regards, plans to inaugurate a Bioinformatics Training Network (BTN) for Nepal can be taken as consideration by Tribhuvan University, TRM Campus, Birgunj for the national interest. In addition to centrally organized and coordinated regular training activities of the network, each node will be challenged to evolve their own training strategies to meet the needs of their immediate environment. It is also envisaged that the proposed network will be able to secure sufficient fund from within and outside Nepal to promote and finance its activities.

References

Allen, G.K. (2004). Bioinformatics: new technology models for research, education, and services. *Educause Centre for Applied Research Bulletin 8*, 1–9.

Boyer, R. (2000). The new biochemistry: blending the traditional with the other. *Biochemistry and Molecular Biology Education 28*, 292–296.

Brass, A. (2000). Bioinformatics education–a UK perspective. *Bioinformatics 16*, 77–78.

Burhans, D. T., Skuse, G. R. (2004). The role of computer science in undergraduate bioinformatics education. *SIGCSE Bulletin 36*, 417–421.

Campbell, A.M. (2003). Public access for teaching genomics, proteomics, and bioinformatics. *Cell Biology Education 2*, 98–111.

Counsell, D. (2003). A review of bioinformatics

education in the UK. *Briefingsin Bioinformatics 4*, 7–21.

Degrave, W., Huynh, C. H., Roos, D., Oduola, A., Morel, M. C. (2002). Bioinformatics for disease endemic countries: opportunities and challenges in science and technology development for health. *IUPAC Journal 1*, 1–7.

Delpech, R., 2006. Bioinformatics and school biology. *Journal of Biological Education 40*, 147–148.

Doom, T., Raymer, M., Krane, D., Garcia, O. (2003). Crossing the interdisciplinary barrier: a baccalaureate computer science option in bioinformatics. *IEEE Transactions on Education 46*, 387–393.

Fetrow, J. S., John, D. J., 2006. Bioinformatics and computing curriculum: a new model for interdisciplinary courses. *ACM SIGCSE Bulletin 38*, 185–189.

Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., et al. (1995). Whole-genome random sequencing and assembly of *Haemophilus influenza* Rd. *Science 269*, 496–512.

Friedman, C. P. (2004). Training the next generation of informaticians: the impact of 'BISTI' and bioinformatics; are port from the American College of Medical Informatics. *Journal of American Medical Informatics Association 11*, 167–172.

Hack, C., Kendall, G. (2005). Bioinformatics: current practice and future challenges for life science education. *Biochemistry and Molecular Biology Education 33*, 82–85.

Hemminger, B. M., Anne-Bauers, T. L. (2005). Survey of bioinformatics programs in the United States. *Journal of the American Society for Information Science and Technology 56*, 529–537.

Kanehisa, M., Bork, P. (2003). Bioinformatics in post-sequence era. *Nature Genetics Supplement 33*, 305–310.

Kumar, A. (2005). Teaching systems biology: an active-learning approach. *Cell Biology Education 4*, 323–329.

Liolios, K., Tavernarakis, N., Hugenholtz, P., Kyrpides, N. C. (2006). The Genomes on Line Database (GOLD) v.2: a monitor of genome projects worldwide. *Nucleic Acids Research 34*, D332–D334 (database issue).

Maojo, V., Kulikowski, C. A. (2003). Bioinformatics and medical informatics: collaboration on the road to genomic medicine? *Journal of the American Medical Informatics Association M1305*.

Pearson, W. R. (2001). Training for bioinformatics and computational biology. *Bioinformatics 17*, 761–762.

Pevzner, P. A. (2004). Educating biologists in the 21stcentury: bioinformatics scientists versus bioinformatics technicians. *Bioinformatics 20*, 2159–2161.

Pongor, S., Landsman, D. (1999). Bioinformatics and the developing world. *Biotechnology and Development Monitor 40*, 10–13

Ranganathan, S. (2005). Bioinformatics education–perspectives and challenges. *PloS Computational Biology 1*, 52.

Sutcliffe, I. C., Hutchings, M. I. (2007). Putative lipoproteins identified by bioinformatics genome analysis of *Leifsonia xyli* subsp. *xyli*, the causative agent of sugarcane ratoon stunting disease. *Molecular Plant Pathology 8*, 121–128.