



## **Incorporating bioinformatics into the biochemistry curriculum**

**Nyerhovwo J. TONUKARI**

*Department of Biochemistry, Delta State University, P.M.B. 1, Abraka, Nigeria*

Received 16 May 2005

MS/No BKM/2005/025, © 2006 Nigerian Society for Experimental Biology. All rights reserved.

---

### **Abstract**

Bioinformatics is fast becoming an integral part of the biological sciences, and it is enhancing the use of genome data and the associated computational resources in basic and applied research in biotechnology and biomedical sciences. Efforts to increase the ability of African scientists to process and analyse genomic and post-genomic data ought to begin with the introduction of bioinformatics into the curriculum. Most web-based genomic databases and sequence analysis sites are essentially free, and these can be utilized for teaching as well as research. This presentation highlights some introductory bioinformatics tools that are freely available and can be quickly understood by higher level undergraduate as well as postgraduate students.

**Keywords:** Introductory bioinformatics, resources, teaching curriculum

---

**E-mail:** tonukari@deltastate.net **Tel:** 08028342603

## INTRODUCTION

Bioinformatics studies how information is stored, reproduced, and used by living systems. It is not an overstatement to say that bioinformatics is what biochemistry and molecular biology is evolving to become in the 21st century. The subject matter of bioinformatics has existed as long as there has been molecular biology, but it has emerged as a distinct discipline only in the last two decades. This was prompted by a veritable explosion of information triggered by the innovations of recombinant DNA and DNA sequencing, which have culminated in the human genome project (HGP). As the time needed to store, annotate, and analyze this stream of information has grown from a laboratory pastime to a full-time job, bioinformatics has, by necessity, emerged as a separate and vital field of study. There is expected to be a dramatic increase in demand for practitioners in this field, who are already in short supply<sup>1</sup>.

The driving force behind the growth of bioinformatics is the rapid accumulation of a rich array of data of diverse types. In order to access and exploit this information, molecular biologists have become increasingly dependent on computational approaches to analyze these data sets - that is, the goal of bioinformatics. Functional characterizations of thousands of gene products from many species are described in the published literature. These are extremely valuable for characterizing the functions not only of these gene products, but also of their homologs in other organisms<sup>2</sup>. The importance of bioinformatics to the future of biology is reflected by its inclusion among the new topics covered in recent molecular and cellular biology textbooks. The challenge for faculty is to find ways to introduce undergraduate biology students to the basic concepts, tools, and databases in bioinformatics<sup>3</sup>.

These rapid advances in the area of bioinformatics and molecular biology create opportunities and challenges for researchers and lecturers, because maintaining the high level of skills and competencies necessary to teach the access, retrieval, and use of relevant information

is becoming more and more complex. For students to effectively use this vast repository of information, they ought to be introduced to bioinformatics software and databases. Even lecturers must engage in continuous learning to stay abreast of relevant scientific advances, especially in the disciplines of genetics and molecular biology.

## BIOINFORMATICS EDUCATION

A student of bioinformatics will need to develop proficiency in both biology and computer science in order to be competitive in today's job market<sup>3</sup>. The goal should be for them to acquire the ability to work from DNA sequence to protein structure and function and back again. Many of us developed this capability during the course of our research training in a piecemeal fashion. It would be advantageous for these undergraduate students to enter graduate programs or industry with an effective working knowledge of bioinformatics. It is also essential that these students gain a deeper and more critical understanding of bioinformatics methods and their applications, so that they have an appreciation of the strengths and limitations of the tools they use in their work<sup>3</sup>.

However, it should be noted becoming a bioinformaticist is not the career goal of most of our students. The majority of our students are interested in pursuing careers in industries or the health sciences. For these students, the objective is not to prepare them for a career in bioinformatics, but to give them literacy in the basic methods and applications of bioinformatics. It is likely that these students will access this information in the workplace or as private citizens<sup>3</sup>.

At present, the major obstacle to bioinformatics education presently is lack of access to an Internet-connected computer. More than 99% of our students do not have their own personal computers and accessing the internet via cyber cafés (the only available option) does not come cheap.

While it is certainly not necessary that students become computer scientists in order to

effectively use bioinformatics programs, an acquaintance with the process of designing, writing, and testing a simple bioinformatics program will demystify the process while providing some understanding of the ways in which even simple computer programs can go astray. Therefore, although my recommendations do not include specialized bioinformatics tasks like annotations, PERL programming and algorithms design, an elective course in bioinformatics at the postgraduate level can deal with these.

## RECOMMENDATIONS

Today, the typical bioinformaticist is by necessity largely self-taught. Although it has always been true that a savvy molecular biologist could take computer science courses, organized programs of study that combine life and information sciences have arisen only in the last decade, and bioinformatics is still an oddity in the university curriculum. The central difficulty in establishing a bioinformatics curriculum is not only that the subject is cross-disciplinary, but that the disciplines it unites are so disparate in method, outlook, and culture<sup>1</sup>. Despite the barriers between the life and computing sciences, successful programs have risen, and are growing steadily mostly in the United States.

In Nigeria today, establishing an independent bioinformatics department with its own specialized faculty, is not yet possible at most institutions; instead, it is necessary to incorporate aspects of bioinformatics into the curriculum.

Computer course taken by 200 level Biochemistry students in some universities is very unrealistic for the modern world. Students are taught some Fortran programming without ever touching a computer. What is urgently needed is an understanding of the common office and internet applications like Microsoft Word, Excel, Access, Powerpoint which are the word processing, statistical analysis, data entry and presentation softwares. Students also need to know how to search the web as well as use online applications which are numerous. Therefore the Computer Science Department

should be petitioned to move into the 21<sup>st</sup> century.

As part of the Biochemical Techniques or Molecular Biology Course, which is taught in most universities during the third year, students ought to learn how to search and retrieve specific gene sequences from GenBank as well as translate these sequences both manually and using Internet-based softwares. Students will obtain good knowledge of introns, the genetic code, as well as translation start sites. For example, students can be given handout containing the DNA sequence of a gene. As they work together in groups to carry out the conceptual translation of the gene from the nucleotide sequence, they occasionally make mistakes, making it necessary for the group to backtrack or even start over. It impresses them to see how precise the cellular machinery must be to faithfully duplicate and express the information in DNA. As they work through this exercise, they also encounter what is often a surprise—an intron that interrupts a codon. Once again, they see how precise the cell must be, in terms of splicing the transcribed RNA.

Other authors have suggested that faculty develop new exercises that reflect the growth and development of bioinformatics as a discipline and the ways in which it can contribute to biology and medicine. For example, students could use on-line tools to analyze the gene makeup of a relatively small genome, such as those of bacteria or viruses<sup>3</sup>.

At the final year students can learn how to design primers during lessons on the application of the polymerase chain reaction which is part of Genetic Engineering Course. They can also work in group to do sequence alignment for the construction of degenerate primers. At this level, students should be very comfortable with Internet search for journals as well as information on proteins, DNA sequence and genomes.

For a PhD as usual, the emphasis is on innovation and creative thinking, whereas the expectation for the MSc-level scientist is to be highly competent in applying a base of

knowledge, and to be able to keep that base of knowledge current. For those interested in a career in independent research, the obvious road is a department that offers a PhD, and the expectation should be for a curriculum that is relatively heavy on computation and light on molecular biology<sup>1</sup>. However, the bulk of introductory bioinformatics to be taught M.Sc. level should include topics that can be easily applied in research such as search, BLAST, restriction mapping, translation, sequence alignment and motif analysis.

### **Search**

Today, scientific data are inevitably digitized, stored in a wide variety of formats, and are accessible over the Internet. Scientific discovery increasingly involves accessing multiple heterogeneous data sources, integrating the results of complex queries, and applying further analysis and visualization applications in order to collect datasets of interest<sup>4</sup>. Established in 1988 as a national resource for molecular biology information, NCBI (<http://www.ncbi.nlm.nih.gov>) creates public databases, conducts research in computational biology, develops software tools for analyzing genome data, and disseminates biomedical information.

### **Blast**

The recent availability of whole-genome sequences and large numbers of protein-coding regions from high-throughput cDNA analysis has fundamentally changed experimental biology. These efforts have provided huge databases of protein sequences, many of which are of unknown function<sup>5</sup>. Deciphering the functions of these myriad proteins presents a major intellectual challenge. BLAST (Basic Local Alignment Search Tool) is a database searching program developed at the National Center for Biotechnology Information (NCBI) in the US. The BLAST programs are widely used tools for searching protein and DNA databases for sequence similarities<sup>6</sup>. BLAST searches for regions of local similarity between a query sequence and database sequence entries.

### **Primer design**

There are also programs designed to select primers for Polymerase Chain Reactions or for

DNA sequencing. These programs select pairs of primers or primers from a single strand of DNA. The program takes into consideration sequence specificity, melting temperature, secondary and primer annealing. Upper and lower limits for any or all of these parameters, for either the product or the primers, can be modified with an extensive array of options. Several programs that can be used to design primers have been documented by Abd-Elsalam<sup>7</sup>.

### **Restriction mapping**

Restriction mapping programs finds and displays restriction endonuclease cleavage sites in DNA sequences. Most of the commercially available restriction endonucleases are listed in the default data file used by the mapping programs. A good restriction mapping program is available at <http://www.firstmarket.com/cutter/cut2.html>.

### **Translation**

For nucleotide sequences you can automatically translate in any forward reading frame or reverse, complement or reverse and complement a sequence. Programs to translate a nucleotide sequence as well as to reverse and complement a nucleotide sequence are available at <http://www.ebi.ac.uk/emboss/transeq>, [http://www.ualberta.ca/~stothard/javascript/rev\\_comp.html](http://www.ualberta.ca/~stothard/javascript/rev_comp.html) and <http://searchlauncher.bcm.tmc.edu/seq-util/Options/sixframe.html>.

### **Sequence alignment**

Understanding gene function, gene regulation, gene networks, phylogenetic studies and other aspects of evolution all depend on accurate nucleic acid and protein sequence alignment<sup>8</sup>. The assembly of a multiple sequence alignment (MSA) has become one of the most common tasks when dealing with sequence analysis. Programs to import, manage and align sequences and to analyse the properties of DNA, RNA and proteins are essential for every biological laboratory, some which are available at [http://npsa-pbil.ibcp.fr/cgi-bin/npsa\\_automat.pl?page=npsa\\_clustalw.html](http://npsa-pbil.ibcp.fr/cgi-bin/npsa_automat.pl?page=npsa_clustalw.html), <http://www.ccmbindia.org/koshika/software/online/softwareonline-nucleicacidseqanalysis.html> and [http://www.ch.embnet.org/software/BOX\\_form.html](http://www.ch.embnet.org/software/BOX_form.html).

## Motif

Motif is program that searches protein sequences for defined sequence patterns (such as consensus sequences for active sites in enzymes, structurally conserved sites, and phosphorylation sites). The program uses a local data file that contains the sequence pattern database. Without doing detailed wet laboratory studies and relying on the PSORT analysis (<http://psort.nibb.ac.jp>), we were able to obtain in earlier studies that the *Cochliobolus carbonum* Snf1 protein is located in the nucleus and has a protein kinase function<sup>9</sup>.

## CONCLUSION

There is at present an unprecedented increase in biological data available from Internet-accessible databases. Clearly we are still at the very beginning of an exciting era of molecular biology, and bioinformatics will contribute heavily to the expansion of our understanding of the molecular basis of life in the years ahead. Bioinformatics is one of the key sciences of the new century, and demand for specialists in the area is strong and growing in the advanced countries. The opportunity to train this new generation of scientists is one that we dare not ignore especially as computing facilities is becoming increasingly affordable.

## REFERENCES

1. **Zauhar, R. J. (2001)** University bioinformatics programs on the rise. *Nature* **19**:285-286.
2. **Raychaudhuri, S., Chang, J. T., Sutphin, P. D., and Altman, R. B. (2004)** Associating Genes with Gene Ontology Codes Using a Maximum Entropy Analysis of Biomedical Literature. *Genome Res.* **12**:203-214.
3. **Honts, J. E. (2003)** Evolving strategies for the incorporation of bioinformatics within the undergraduate Cell Biology curriculum. *Cell Biol. Educ.* **2**:233-247.
4. **Lacroix, Z., Raschid, L., and Eckman, B. A. (2004)** Techniques for optimization of queries on integrated biological resources. *J. Bioinform. Comput. Biol.* **2**:375-411.
5. **Shapiro, L. and Harris, T. (2000)** Finding functions through structural genomics. *Curr. Opin. Biotechnol.* **11**:31-35.
6. **Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D. J. (1997)** Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**:3389-402.
7. **Abd-Elsalam, K. A. (2003)** Bioinformatic tools and guideline for PCR primer design. *Afr. J. Biotechnol.* **2**:91-95.
8. **Chain, P., Kurtz, S., Ohlebusch, E. and Slezak, T. (2003)** An applications-focused review of comparative genomics tools: capabilities, limitations and future challenges. *Brief. Bioinform.* **4**:105-123.
9. **Tonukari, N. J., Scott-Craig, J. S. and Walton, J. D. (2000)** The *Cochliobolus carbonum* SNF1 Gene Is Required for Cell Wall-Degrading Enzyme Expression and Virulence on Maize. *Plant Cell* **12**:237-247.