Bioinformatics Becomes Respectable

Stuart M. Brown

New York University School of Medicine, New York, NY, USA

BioTechniques 34:1124-1127 (June 2003)

Bioinformatics is a practical discipline that is concerned with the methods used by biologists to store and analyze their research data using computers. As such, it has received quite a bit of attention in the pages of *BioTechniques*. This has been very useful to biologists, since there have been few journals that publish articles that focus on the description of useful bioinformatics innovations. Often, major advances in bioinformatics are buried in the Materials and Methods sections of papers whose titles and primary focuses are on a basic biological problem. This makes it difficult for people to learn about these innovations if they are working on similar data analysis problems in unrelated branches of biology.

ORIGINS OF BIOINFORMATICS

Modern bioinformatics represents the convergence of two historical trends in biological research—the storage of molecular sequences in computer databases and the application of computational algorithms to the analysis of DNA and protein sequences. As a hybrid or interdisciplinary subject, bioinformatics is difficult to define. Most definitions suggest the union of information technology, computer science, and computational biology. I favor an operational definition: the application of computer technology for the acquisition, management, and analysis of biological information. Another definition might be “applied computational biology”.

The database aspect of bioinformatics traces its history to the Los Alamos Sequence Library, which was established in 1979 at the DOE Los Alamos National Laboratory (LANL) and to Margaret Dayhoff’s Atlas of Protein Sequence and Structure (1), first published in print form in 1965, now online as the Protein Information Resource (PIR) (<http://pir.georgetown.edu/>).

Many of the algorithms that constitute the core of the current bioinformatics tool kit originated in the 1970s and 1980s as theoretical computational biology. However, when they are pressed into service as data processing subroutines, then they are properly considered bioinformatics.

I would like to acknowledge some of the milestones of algorithm development by a few of the computational biology pioneers who have made this field possible:

- 1970** - The Needleman-Wunsch algorithm for sequence comparison (global alignment) (2)
- 1978** - Chou and Fasman - prediction of protein secondary structure from amino acid sequence (3)

- 1979** - The Staden package for the assembly of overlapping fragments in DNA sequencing projects (4,5)
- 1981** - The Smith-Waterman algorithm for sequence alignment (local alignment) (6)
- 1982** - Kyte and Doolittle - plotting protein hydropathy (7)
- 1987** - Feng and Doolittle - multiple sequence alignment (8)
- 1987** - Saitou and Nei - the neighbor-joining method for reconstructing phylogenetic trees (9)
- 1988** - Pearson and Lipman - FASTA - fast heuristic similarity searching of a sequence database (10)
- 1990** - Altschul et al. - The BLAST program is implemented at the NCBI (even faster heuristic similarity searching of sequence databases) (11)

BIOINFORMATICS IN THE SCIENTIFIC LITERATURE

Before 1985, the term “bioinformatics” is not indexed as a keyword or in any title or abstract in the PubMed/Medline database. The term bioinformatics does not appear in any articles in 1991 or 1992. Three articles published in 1993 include the term, 9 in 1994, 10 in 1995, and it takes off exponentially from there (Figure 1).

Early bioinformatics articles most often appeared in the journals *Computer Applications in the Biological Sciences* (founded in 1985), *The Journal of Molecular Biology*, or *Nucleic Acids Research*. *Computer Applications in the Biological Sciences* changed its name to *Bioinformatics* in 1998 and now serves as a leading journal in the field. *Nucleic Acids Research* still plays a major role, publishing an annual issue devoted to entirely to biological databases. There are also a number of newer journals that now offer bioinformatics as their primary focus: *Applied Bioinformatics* (2002), *Bioinform* (1997), *Briefings in Bioinformatics* (2000), *Journal of Computational Biology* (1994), *Journal Of Biological Systems* (1993), and *Journal of Bioinformatics and Computational Biology* (2003) (see Table 1 for more resources).

BIOINFORMATICS TRAINING

Bioinformatics has long suffered from a “two cultures” problem in the interaction between computer scientists and biologists. This has been evident in small matters such as the

Table 1. Bioinformatics Resources

Professional Societies
International Society for Computational Biology Bioinformatics.org
International Conferences
IEEE Computer Society Bioinformatics Conference O'Reilly Bioinformatics Technology Conference Asia-Pacific Bioinformatics Conference Intelligent Systems for Molecular Biology (ISMB) International Conference on Computational Molecular Biology (RECOMB)

style of presentations at conferences (PowerPoint™ for the biologists, handwritten transparencies for the mathematicians) and in larger matters such as the lack of a common vocabulary to discuss biological data and the methods by which it should be analyzed. There has been some progress in these areas over the last 5–10 years as some public databases and computer programs have become widely popular. The U.S.

National Center for Biotechnology Information (NCBI; www.ncbi.nlm.nih.gov) has had a particularly large role in creating a common set of terminology since its GenBank® database, ENTREZ query engine, and BLAST sequence similarity searching tool have become the international standards. It is now possible to speak to a roomful of either experimental molecular biologists or computational biologists and use terms such as “GenBank accession number” or “BLAST e-value” without pausing to define them.

As many more research projects have been designed to gather large amounts of DNA sequences, protein mass spectrometry profiles and large-scale measurements of mRNA gene expression, bioinformatics skills have been in increasing demand. The number of positions has exceeded the number of qualified people, both for biologists with computer skills and for computer scientists with an understanding of the terminology and problems of molecular biology. To meet the need for bioinformaticians, government and private research foundations have begun funding bioinformatics training programs. There are now approximately 50 institutions in the United States that offer graduate training in bioinformatics, including 40 Ph.D. programs (12). However, since most of these training programs were founded in the past five years, there is an interesting discontinuity between students and fac-

20th Anniversary Perspective

ulty. Almost no one who is now teaching bioinformatics has a degree in this subject. The community of professional bioinformaticians—those who teach the courses, write the grant proposals, and serve on the grant review panels—is composed of biologists who are self-taught in computer science methods and computer scientists (as well as some mathematicians, statisticians, and the occasional physicist) who are self-taught in molecular biology. While such a diversity of perspectives has its advantages, it must be disconcerting for the students when they realize that they are being trained for a profession that has no coherent traditions and almost no structure. The various training programs are taught by faculty drawn from both the biology and computer science departments and offer a curriculum that is a hodgepodge collection of best guesses from the available resources at each institution that do not reflect any common philosophy or core curriculum. Hopefully, as these training programs mature, they will develop a coherent set of concepts that will be taught across different schools, and the graduates will enter the ranks of bioinformatics professionals and faculty as a bridge between the biology and computer science cultures.

In our ongoing work at New York University to develop a bioinformatics/computational biology training program, we have encountered several challenges that must be faced at each institution as it develops a bioinformatics concentration. First, bioinformatics is extremely problematic as an undergraduate major. It is simply too much material to require for a student to learn in four years. It is not possible to develop a thorough grounding in either biology or computer science if one is forced to take a full set of courses in both departments. And what of basic writing, social science, and other requirements for a well-rounded undergraduate education? Another difficulty is the diversity of backgrounds of the potential students. It seems obvious that a bioinformatics graduate program must admit both Biology and CompSci/Math majors as well as the occasional physicist and economist. So how can we provide the necessary

background training in math and computer science for the biologists and biology training for the quantitative students without creating two completely different tracks—which would lead to two different kinds of graduates and would exacerbate the “two cultures” problem? In fact, the goal should be to force students from diverse backgrounds to spend as much time working together as possible to develop collaborative skills and a common vocabulary that would allow bioinformatics to solidify as a single coherent discipline. At NYU, our current plan to address this problem of prerequisites is a sort of summer school intensive program where we will fill in as much of the missing background material for each student as possible in two sets of short courses: (i) biology for quantitative scientists and (ii) math and programming skills for biologists.

A PLACE AT THE TABLE

In just the past couple of years, bioinformatics seems to be gaining a new legitimacy as an academic discipline. I offer some of our experiences at NYU School of Medicine as typical of events and trends at many biomedical research institutions. NYU School of Medicine has had a bioinformatics core facility for more than 15 years. Ten years ago, its main functions were to maintain a mainframe computer (running VMS!) that provided GCG (Wisconsin Package) services to faculty and students and to teach a graduate course in bioinformatics methods. Occasionally, we would be called on to provide some consulting on a particularly thorny sequence analysis problem—a difficult alignment or phylogenetic tree, defining a new motif, etc.

However, about 2 or 3 years ago, we began to notice an increasing number of requests for assistance in the analysis of large sets of DNA sequences—particularly various types of bulk expressed sequence tag sequencing projects that involved differential or subtracted cDNA libraries. The bioinformatics

Figure 1. The growth of the terms “bioinformatics” and “computational biology” in PubMed-indexed scientific articles from 1985 to 2002.

group is now being drawn into earlier phases of projects, getting first crack at the data as it comes off the sequencers: to cluster, analyze, annotate, and organize the sequences before the laboratory-based scientists tried to discern the meaning of their experiments. This has led to more co-authored papers and some additional funding for the Bioinformatics center as a sub-contractor or collaborator on various grants.

The recent increase in the popularity of microarrays and other genomics technologies (large scale single nucleotide polymorphism genotyping, bulk genome sequencing of microorganisms from patient samples, etc.) has led to a huge surge in requests for bioinformatics consultations, letters of support for grant applications, and a new respect for bioinformatics within the administrative structure of our university. It has become clear to even the most reluctant administrators that institutional strategies for Cancer Research, Clinical Trials, and new "Research Buildings" need to include plans for the use of genomics technologies and that genomics requires bioinformatics support. I have been called into more meetings with Deans and Center Directors in the past 6 months than in all of the previous 6 years of my tenure here.

In the past few months, bioinformatics has been given a seat at the table of the NYU Clinical Research Advisory Committee. This group approves (or disapproves) all research projects that deal with clinical trials and human subjects. Now, in addition to prior review for safety, as well as biostatistics, proposed projects will be vetted for their bioinformatics needs to ensure that the data analysis plans are reasonable and to give my group some prior warning as the projects are initiated. By taking bioinformatics into account in the research planning process, it is becoming legitimized and accepted as an essential component of the research process, much as biostatistics was incorporated into clinical research 20–30 years ago.

If the exponential growth of bioinformatics papers shown in Figure 1 continues for several more years, then grant funding for bioinformatics research is likely to follow a similar trajectory. At most universities, increased funding means faculty recruitment, more graduate students, and more course offerings. This suggests that bioinformatics is going to fill an increasingly significant role as an academic discipline as well as its now traditional role as an essential research support service.

On the other hand, Lincoln Stein has recently suggested (O'Reilly Bioinformatics Technology Conference, Keynote, February 2003) that bioinformatics is a tool and not a scientific discipline—because it is defined by its techniques rather than by a fundamental biological problem. However, this distinction has not been a major impediment to the Departments of Radiology or Surgery at our medical school. In Stein's view, bioinformatics, like Molecular Biology (or the use of PCR), will be absorbed into mainstream biology so that almost every biologist will become skilled at using computational tools for the analysis of their data. I think that this is not likely. Yes, every graduate student now uses BLAST on the NCBI Web site, but how many of them can build even a simple database? Is this likely to change in the next 10 years? Perhaps there is a compromise position where new fields and departments will be defined around new problems and approaches that are fundamentally made possible by bioinformatics—such as the new Department of Genome Sciences at the University of Washing-

ton School of Medicine. And there will still be a need for skilled people to develop new databases and new algorithms as genomics and other fields develop new technologies that produce new types of data in huge amounts. I think those students in the many new bioinformatics training programs will have excellent career prospects.

REFERENCES

1. Dayhoff, M.O., R.V. Eck, M.A. Chang, and M.R. Sochard. 1965. Atlas of Protein Sequence and Structure, vol. 1. National Biomedical Research Foundation, Silver Spring, MD.
2. Needleman, S.B. and C.D. Wunsch. 1970. A general method applicable to the search for similarities in the amino acid sequences of two proteins. *J. Mol. Biol.* 48:443–453.
3. Chou, P.Y. and G.D. Fasman. 1978. Prediction of the secondary structure of proteins from their amino acid sequence. *Adv. Enzymol.* 47:145–148.
4. Smith, T.F. and M.S. Waterman. 1981. Identification of common molecular subsequences. *J. Mol. Biol.* 147:195–197.
5. Kyte, J. and R.F. Doolittle. 1982. A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.* 157:105–132.
6. Staden, R. 1979. A strategy of DNA sequencing employing computer programs. *Nucleic Acids Res.* 7:2601–2610.
7. Staden, R. 1980. A new computer method for the storage and manipulation of DNA gel reading data. *Nucleic Acids Res.* 8:3673–3694.
8. Feng, D.F. and R.F. Doolittle. 1987. Progressive sequence alignment as a prerequisite to correct phylogenetic trees. *J. Mol. Evol.* 25:351–360.
9. Saitou, N. and M. Nei. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* 4:406–425.
10. Pearson, W.R. and D.J. Lipman. 1988. Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci. USA* 8:2444–2448.
11. Altschul, S.F., G. Warren, W. Miller, E.W. Myers, and D.J. Lipman. 1990. Basic local alignment search tool. *J. Mol. Biol.* 215:403–410.
12. Zatz, M.M. 2002. Bioinformatics training in the USA. *Brief. Bioinform.* 4:353–360.

Address correspondence to Dr. Stuart M. Brown, Research Computing Resource, New York University School of Medicine, 550 First Ave., New York, NY 10016, USA. e-mail: stuart.brown@med.nyu.edu