Bioinformatics – a new era

Mark S. Boguski
National Center for Biotechnology Information, National Library of Medicine, NIH, Bethesda, MD 20894, USA.
boguski@ncbi.nlm.nih.gov

Today, most graduate students and postdocs would find it difficult to imagine a time when sequence databases and search tools were not a ubiquitous and accessible part of the research landscape[1]. I'm not *that* old but, even during my first year of college, the only sequence 'database' available was a thin book containing 65 tRNA and 5S RNA sequences (mostly from microorganisms) together with a smattering of short samples of RNA from a few bacteriophages and viruses[2]. The first mammalian mRNA sequence to be determined, rabbit β globin mRNA, made the cover of *Cell* in 1977, and papers describing the coding sequence, the 5′ untranslated region and the 3′ untranslated region merited three separate publications in the issue[3–5]! Students entering graduate school at that time might still expect to receive a PhD for cloning and sequencing a single cDNA – now you can't get one for sequencing a million!

GenBank[6] was not set up until 1982 (Ref. 7) and, during the early days, was distributed to university computing centers four times per year on magnetic tapes. By the time I needed to do my first homology search in graduate school, GenBank contained a whopping 2427 sequences (compared with ~2 532 359 available today), most of which were typed in manually from journals (or from printed, hard-copy submissions) by GenBank curators. Luckily, against all odds, I got an informative 'hit' with my very first query sequence[8] and that experience profoundly altered the future direction of my career.

The term 'bioinformatics' is a fairly recent invention, not appearing in the literature until around 1991 and then only in the context of the emergence of electronic publishing[9]. I think that the current concept of bioinformatics was best described as the convergence of two technology revolutions: the explosive growth in biotechnology, paralleled by the explosive growth in information technology[10]. This is illustrated, in an uncanny way, by the fact that both the size of GenBank and the power of computers have been doubling at about the same rate (every 18–24 months) for many years (Fig. 1).

The term bioinformatics still carries with it enough hype to make investigating 'biology with computers' seem like the cutting edge. However, some of my role models when I was a graduate student (Margaret O. Dayhoff, Russell F. Doolittle, Walter M. Fitch and Andrew D. McLachlan) had been building databases, developing algorithms and making biological discoveries by sequence analysis since the 1960s (see Refs 1, 11), long before anyone thought to label this activity with a special term (if anything, it was called 'molecular evolution'). Even a relatively new kid on the block, the National Center for Biotechnology Information (NCBI), is celebrating its 10th anniversary this year, having been written into existence by US Congressman Claude Pepper and President Ronald Reagan in 1988 (Ref. 12). So bioinformatics has, in fact, been in existence for more than 30 years and is now middle-aged. This is, of course, a time of life for reinvention and renewal, and I will describe some of the challenges and opportunities that this discipline faces today.

Practitioners and training

Bioinformatics is still a somewhat nebulous term that can mean anything from bar-coding samples in an industrial laboratory to hypothesis-driven research. Apart from the obvious data management applications of bioinformatics, computational biology research is divided into two main schools: the analysis and interpretation of data and the development of new algorithms and statistics (you will find examples of both schools in this guide). Most current practitioners are still self-taught because university departments of computational biology do not yet exist. Other types of training programs are limited in number and scope[13], although several excellent, practical, short courses [such as the one at the Cold Spring Harbor Laboratory (http://nucleus.cshl.org/meetings/98c-ecg.htm)] are offered periodically. Therefore, the supply of 'skilled labor' in bioinformatics is inadequate, and most organizations must be willing to provide on-the-job training. Market forces have responded to this labor supply problem by the creation of several small companies that provide bioinformatics products and services, mostly for industry, so large companies now have the option of outsourcing some of their bioinformatics needs to third parties.

Nevertheless, there is an urgent need to train the next generation in a more formal, academic manner by establishing training programs in university departments with a 'critical mass' of faculty and adequate financial support. Appropriate training at the undergraduate level is also to be fostered and supported. One encouraging sign is the influx of physicists into biology.

For those who want to teach themselves, several excellent books stressing both practical[14] and theoretical[15–17] aspects of computational biology have recently appeared. I would also recommend that computational biologists obtain a working

0167-7799/98/$ – see front matter. Published by Elsevier Science.

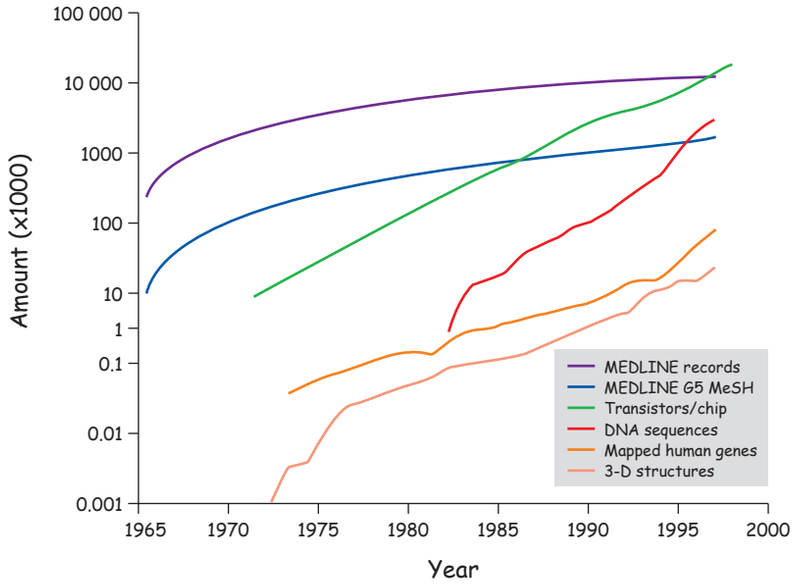

Fig. 1. Cumulative growth of biomedical information and computing power. MEDLINE (purple line) is the bibliographic database of the US National Library of Medicine (http://www.nlm.nih.gov/Entrez/medline.html) and currently contains >10 million records derived from published articles in >3900 biomedical journals. Articles categorized under the 'G5' Medical Subject Heading (MeSH) of 'molecular biology and genetics' (blue line) total nearly 1 million. The total number of DNA sequence records in GenBank (red line) is >2.5 million (data from ftp://ncbi.nlm.nih.gov/genbank/gbrel.txt). Note that there are now more sequence records in GenBank than there are related publications in the literature, indicating an important milestone and an increasing gap in our lack of understanding of the functions of these sequences. Hopefully 'functional genomics' technologies will help us to narrow this gap. The line representing the number of transistors per chip (green line) refers to Intel™ microprocessors and illustrates Moore's Law, which refers to the exponential growth rate of computing power (data obtained from http://www.physics.udel.edu/http://wwwusers/watson/scen103/intel.html). The number of mapped human genes (orange line) is currently >30 000 (Ref. 26 and http://www.ncbi.nlm.nih.gov/genemap). The number of three-dimensional protein structures in the Protein Data Bank (pdb) (pink line) is currently ~7500 (http://www.pdb.bnl.org).

knowledge of molecular evolution and structural biology in order to understand and interpret sequence data fully.

Sequences and beyond

Most of the articles in this guide concern the analysis of sequence data and, for a long time, bioinformatics has been virtually synonymous with sequence data management and analysis. This activity has reached an impressive new level in comparative genomics applications (see the article by James Lake and Jonathan Moore on pp. 22–23). However, I believe that sequence analysis alone is too limited a definition and scope for the field of computational biology. I will return to this issue later, but first I will consider some trends and challenges in the investigation and explication of sequence data.

Even before the emergence of complete genomes and large expressed-sequence tag (EST) surveys[18], GenBank [and its counterparts, the European Molecular Biology Library (EMBL) Data Library and the DNA Database of Japan] had become a large, complex and internally redundant sequence archive that required considerable experience and/or a primer[19] to use it most efficiently and effectively. Many practical tips on this subject are discussed in this guide. One of the challenges and necessities for the future is to reorganize and streamline the data for more efficient use. The NCBI, for example, has undertaken a 'reference genes' project, in part, to address this issue (J. Ostell, pers. commun.). Countervailing forces, however, are making this task more arduous and more urgent. The daily release of 'unfinished' sequences onto numerous Web sites makes it extremely difficult for biologists to maintain a current and comprehensive view of available data, despite the fact that 'consumer guides' to these Web sites have been published and are reasonably helpful[20]. This situation is going to get even worse, as there is a movement towards a vast acceleration in the production of unfinished, fragmentary (assembled shotgun) sequence data from genome sequencing laboratories[21]. Biologists will require new software tools (and experimental validation resources) to maximize the utility of these data.

Even for 'finished' (i.e. highly accurate, contiguous sequence) data, there are serious issues with annotation that affect our ability to rely on consistent, up-to-date and quality-assured information about genes and genomes. These issues have been described recently[22] and it is unnecessary to reiterate them here. Suffice it to say that annotation, particularly gene prediction, remains a challenging problem for genome interpretation (see the article by David Haussler on pp. 12–15).

New directions

I believe that the most exciting frontier is at the interface between computational and high-throughput experimental biology. For many years, there has been an 'impedance mismatch' between the rapid output of computational predictions and the ability of traditional experimental methods to test and verify these predictions. Through the development and application of new gene expression technologies, for example, 'wet bench' biologists can produce functional information about gene products almost as rapidly as computational biologists can analyse the underlying genomes (see the article by Michael Brownstein *et al.* on pp. 27–29). There are tremendous challenges and opportunities to be found here, and much new biology to be discovered.

Finally, I don't think that computational biologists should ignore the fields of genetic epidemiology and evolutionary genetics. In the past, these have been rather small and insular specialties. But, once again, new technology is shortly going to be 'raining down'[2] SNPs (single nucleotide polymorphisms) upon us[23], and statistical and computational analysis of the relationships between detailed genotypes and complex phenotypes will play a very large part in the future of mammalian and plant biology[24,25].

References

1 Doolittle, R.F. (1997) *J. Mol. Med.* 75, 239–241
2 Barrell, B.G. and Clark, B.F.C. (1974) *Handbook of Nucleic Acid Sequences*, Joynson–Bruvvers
3 Baralle, F.E. (1977) *Cell* 10, 549–558
4 Efstratiadis, A., Kafatos, F.C. and Maniatis, T. (1977) *Cell* 10, 571–585
5 Proudfoot, N.J. (1997) *Cell* 10, 559–570
6 Benson, D.A. *et al.* (1998) *Nucleic Acids Res.* 26, 1–7
7 Smith, T.F. (1990) *Genomics* 6, 701–707
8 Boguski, M.S. *et al.* (1984) *Proc. Natl. Acad. Sci. U. S. A.* 81, 5021–5025
9 Franklin, J. (1991) in *The Future of the Medical Journal* (Lock, S.P., ed.), BMJ Press
10 Murray-Rust, P. (1994) *Curr. Opin. Biotechnol.* 5, 648–653
11 Dayhoff, M.O. (1969) *Sci. Am.* 221, 86–95
12 Cook-Degan, R. (1994) *The Gene Wars: Science, Politics and the Human Genome*, Norton and Co.
13 Austin, M.J.F. (1998) *Current Status of Bioinformatics Training Programs and Related Activities*, Merck Genome Research Institute, West Point, PA, USA
14 Baxevanis, A. and Ouellette, B.F.F. (1998) *Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins*, John Wiley & Sons
15 Baldi, P. and Brunak, S. (1998) *Bioinformatics: The Machine Learning Approach*, MIT Press
16 Durbin, R. *et al.* (1998) *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*, Cambridge University Press
17 Gusfield, D. (1997) *Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology*, Cambridge University Press
18 Boguski, M.S. (1995) *Trends Biochem. Sci.* 20, 295–296
19 Ouellette, B.F. and Boguski, M.S. (1997) *Genome Res.* 7, 952–955
20 Pruitt, K.D. (1997) *Genome Res.* 7, 1038–1039
21 Venter, J.C. *et al.* (1998) *Science* 280, 1540–1542
22 Wheelan, S.J. and Boguski, M.S. (1998) *Genome Res.* 8, 168–169
23 Chakravarti, A. (1998) *Nat. Genet.* 19, 216–217
24 Schafer, A.J. and Hawkins, J.R. (1998) *Nat. Biotechnol.* 16, 33–39
25 Schork, N.J., Cardon, L.R. and Xu, X. (1998) *Trends Genet.* 14, 266–272
26 Deloukas, P. *et al. Science* (in press)

Text-based database searching

As the amount of biologically relevant data is increasing at such a rapid rate, knowing how to access and search this information is essential. There are three data retrieval systems of particular relevance to molecular biologists – Entrez, Sequence Retrieval System (SRS) and DBGET.

Fran Lewitter
Whitehead Institute for Biomedical Research, Nine Cambridge Center, Cambridge, MA, USA.
lewitter@wi.mit.edu

The amount of biological information accessible via the World Wide Web (WWW) is truly astonishing, and the volume of data is increasing at a fast pace. It is important for the bench scientist to have easy and efficient ways of wading through the data and finding what is important to his or her research. Although one can browse the data, a far more efficient access method is to perform a search. Depending on the type of data at hand, there are two basic ways of searching: using descriptive words to search text databases or using a nucleotide or protein sequence to search a sequence database. This article focuses on the former; see the articles by Stephen Altschul (pp. 7–9) and Steven Brenner (pp. 9–12) for information about sequence-based searching.

Here, I will discuss three tools – Entrez, the Sequence Retrieval System (SRS) and DBGET – that allow text searching of multiple molecular biology databases and provide links to relevant information for entries that match the search criteria (see the URLs box). Examples of basic and advanced search strategies are also included. Although many databases that can be accessed with text-based searching will not be discussed here, the search strategies presented are broadly applicable and can be used to search many organism-specific resources, such as the *Saccharomyces* Genome Database (SGD)[1] and the Mouse Genome Database (MGD)[2].

These retrieval systems are indispensable to the scientist in search of information. In using any of these systems, queries can be as simple as entering the accession number of a newly published sequence or as complex as searching multiple database fields for specific terms (see Box 1 for search concepts). The advantage of Entrez, SRS and DBGET is that they not only return matches to a query, but also provide handy pointers to additional important information in related databases. The three systems differ in the databases that they search and the links they make to other information.