

Bioinformatics

— a new era

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 landscape¹. 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 viruses². 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 issue³⁻⁵! 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!

GenBank⁶ 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 sequence⁸ 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 publishing⁹. I think that the current concept of bioinformatics was best described as the convergence of two technology

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

revolutions: the explosive growth in biotechnology, paralleled by the explosive growth in information technology¹⁰. 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 scope¹³, 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 practical¹⁴ and theoretical¹⁵⁻¹⁷ aspects of computational biology have recently appeared. I would also recommend that computational biologists obtain a working

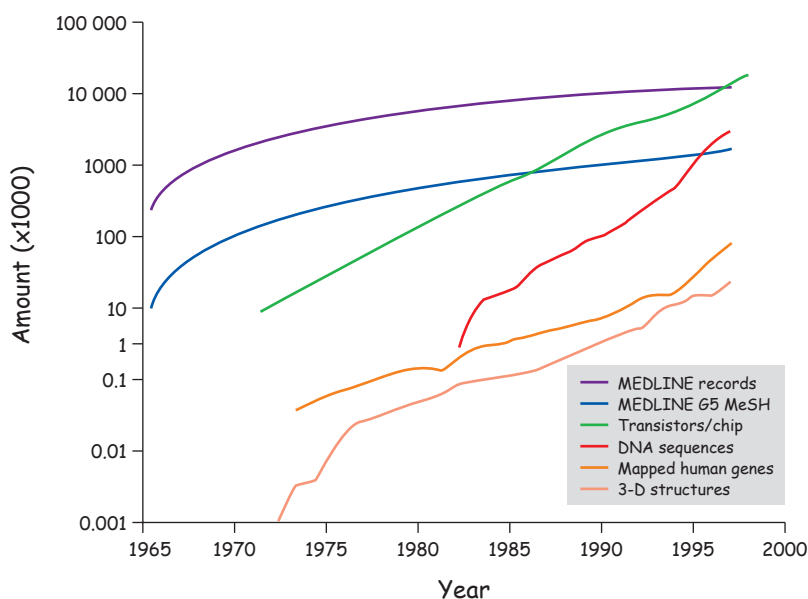


Fig. 1. Cumulative growth of biomedical information and computing power. MEDLINE (purple line) is the bibliographic database of the US National Library of Medicine (<http://www.nlm.nih.gov/Entrez/medline.html>) and currently contains >10 million records derived from published articles in >3900 biomedical 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://www.users/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 exploitation of sequence data.

Even before the emergence of complete genomes and large expressed-sequence tag (EST) surveys¹⁸, 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 primer¹⁹ 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 helpful²⁰. 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 laboratories²¹. 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 recently²² 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'²² SNPs (single nucleotide polymorphisms) upon us²³, 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}.

TRENDS GUIDE TO BIOINFORMATICS

Biological data, and DNA sequence data in particular, are accumulating at a phenomenal rate. By around 2005, it is likely that the DNA sequence of the complete human genome will have been determined. Although this achievement might seem an end in itself, in reality it is only the beginning. In order to exploit the wealth of DNA sequence and other biological data, a new science has arisen that fuses biology with mathematics and computer science – 'bioinformatics'.

To find the genes within the genomic sequence is a massive task in itself. Once apparent, otherwise uncharacterized coding regions must be assigned a function. Thereafter, the interactions between genes and gene products must be understood at all levels, not merely in the context of the pathways within and between cells but also in terms of the evolution of gene families within and between species. These questions can all be addressed using bioinformatics.

Bioinformatics touches all of biology, and straightforward access to data via the Internet means that a wealth of information is available, literally at our fingertips. However, the newcomer to bioinformatics might be discouraged because of the initially daunting computational and mathematical content. Others might simply be confused by the language and terminology of bioinformatics. To help overcome these barriers, the *Trends Guide to Bioinformatics* examines the background to this novel and rapidly evolving scientific discipline. A series of tutorials, written by expert authors, clearly explains the concepts and provides practical examples of how the immense store of data can be exploited. Technical terms are highlighted in each article and defined in the glossary on pp. 32–33. Topics covered range from retrieving and aligning sequences, to predicting structure and function of gene products. Whether you are interested in molecular structure or taxonomy of organisms, the *Trends Guide to Bioinformatics* is an essential tool.

- 1 Bioinformatics – a new era
Mark Boguski
- 3 Text-based database searching
Fran Lewitter
- 7 Fundamentals of database searching
Stephen Altschul
- 9 Practical database searching
Steven Brenner
- 12 Computational genefinding
David Haussler
- 15 Multiple-alignment & -sequence searches
Sean Eddy
- 18 Protein classification & functional assignment
Kay Hofmann
- 22 Phylogenetic analysis & comparative genomics
James Lake and Jonathan Moore
- 24 Databases of biological information
Minoru Kanehisa
- 27 Functional genomics
Michael Brownstein, Jeffrey Trent and Mark Boguski
- 30 The future of bioinformatics
Janet Thornton
- 32 Glossary

COPYRIGHT INFORMATION



©1998 Elsevier Science. All rights reserved. This supplement and the individual contributions contained in it are protected under copyright by Elsevier Science. See the box in the accompanying *Trends* journal for further terms and conditions that apply to the copyright. Except as outlined in the terms and conditions, no part of this publication may be reproduced, stored in a retrieval system or transmitted in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, without prior written permission of the publisher.

Guest Editors

Steven Brenner, Stanford University, CA, USA
and Fran Lewitter, Whitehead Institute for
Biomedical Research, Cambridge, MA, USA

Editors

Mark Patterson and Michaela Handel

Editorial Staff

Caroline Ash, Meran Owen, Rob Brines and
John Pettigrew

Editorial Administrator

Helen Steele

Production Designer

Naomi Wright

Publisher

Peter Desmond