# Biological data becomes computer literate: new advances in bioinformatics
## N Goodman

Bioinformatics is an art and science concerned with the use of computing in biological research areas such as genomics, transcriptomics, proteomics, genetics, and evolution. This review paints a broad picture of bioinformatics, drawing examples from genomic sequencing and microarray analysis. I highlight the role of bioinformatics at multiple points along the path from high-tech data generation to biological discovery.

**Addresses**
Freelance Bioinformatics Consultant, Seattle WA, USA
e-mail: natg@shore.net

**Abbreviation**
EST        expressed sequence tag

## Introduction
The term bioinformatics was coined by Hwa Lim in the late 1980s, and popularized in the 1990s through its association with the human genome project. It is a young field that is still defining itself. The term commonly refers to computational work in genomics, the many other new 'omics' that are sprouting, and neighboring research areas. The precise boundaries of the field are elusive, but as a general rule the closer a research area is to genomics, the more likely its computational aspects will be labeled bioinformatics. To muddy the waters further, some people distinguish bioinformatics from computational biology, using the latter to denote computational work whose agenda is clearly biological; it is often difficult to draw this distinction in practice, as much of what is done in the field is interdisciplinary and combines computational and biological expertise.

Bioinformatics is both an engineering art and a science. It encompasses the development of new computational methods and the application of those methods to solve biological problems. It also has a large service component in which computational resources, such as databases, are operated for the benefit of the research community.

Bioinformatics is a broad field that has a central role in many areas of biological research. These include genomics and, more specifically, genomic sequencing and mapping, genome annotation, and comparisons of multiple genomes. Bioinformatics is also essential in transcriptomics — the study of transcribed sequences, both full-length cDNAs and expressed sequence tags (ESTs) — and the analysis of gene expression data typically measured using DNA microarrays or some form of sample sequencing. It is also

crucial in proteomics for the analysis of protein sequences (e.g. to determine functional motifs), for the study of protein abundance (typically measured using two-dimensional gels or mass spectrometry), and the determination of protein structure either empirically or computationally. Bioinformatics is key in the analysis of protein–protein interactions and molecular pathways (the 'interactome') and in systematic studies of gene regulation (the 'regulome'). It also plays a vital role in genetics, both in the discovery of new molecular genetic markers, such as single nucleotide polymorphisms, and the use of these and other markers to dissect the genetic basis of disease and other phenotypes. Bioinformatics is also essential in studies of evolution and phylogeny.

In this overview, I will draw most of my examples from two biological areas that occupied the scientific headlines this year: genomic sequencing of higher organisms and the use of DNA microarrays to study gene expression. Many of the references are websites, rather than traditional publications, as these provide convenient entry points into the field.

This was the year of the vertebrate genome, with the publication of the draft human genome sequence by a public consortium [1•] and a private company [2•] in the early months of 2001. Substantial progress on the mouse genome was reported by both public [3•] and private [4] efforts during the year, and the draft and near-draft sequences for two pufferfish species were announced towards the end of the year [5•,6•]. The sequencing of a vertebrate genome is a huge project requiring several million to several tens of millions of sequencing runs. Bioinformatics is critical for the generation and analysis of such large datasets.

DNA microarrays were another hot technology this year [7••,8•]. A single microarray run generates between 100 000 and 1 million data values, while a typical series of experiments requires tens to hundreds of runs. Numerous software packages are available to analyze microarray datasets, but the development of new methods remains an active area of research.

## Breadth of the field: biological research areas
As mentioned in the introduction, bioinformatics touches a wide range of biological research areas. Although there are scientific commonalities across these areas, there are also major differences that affect bioinformatics methods.

For example, the problems of analyzing genomic sequences, transcribed sequences and protein sequences are similar in that they can all be described mathematically as sequences of letters, and all are subject to mutational

pressures (hence diverge in predictable ways across taxa). Beyond this, however, the methods and issues are quite different. With genomic sequences, a key issue is gene prediction [9••]. With transcribed sequences, a key issue is clustering of redundant sequences to coalesce all sequences that belong to the same gene [10•–12•]. For protein sequences, key issues are discovering functional motifs that are conserved across evolution, and the use of these motifs to functionally classify novel sequences [13•].

As we move beyond sequences to areas like gene expression, the diversity of methods becomes even more pronounced. In gene expression, a key issue is the clustering of genes that show similar patterns of expression [14••]; there is no analog to this in the sequence world.

## From data to knowledge
The field is further broadened by the need for computation at several steps along the path from data generation to biological interpretation. A hallmark of omic biology is the use of powerful laboratory technologies to generate large, systematically collected datasets. The people who produce the data are experts in these technologies, whereas those who derive biological knowledge from the data are experts in specific biological problems — particular diseases, pathways, and so on — and make discoveries by combining data obtained using many different methods. A key challenge of bioinformatics is to bridge the considerable gap between technical data production and its use by scientists for biological discovery.

The first step along this path is to collect the data. Omic biology generally entails large laboratory projects that process thousands to millions of biological materials through a series of automated procedures. Software is needed to control the various automated instruments in such a laboratory, to extract data from these instruments, to make sure the correct materials are fed into each instrument at the correct time, and generally to keep track of what is going on. This area is broadly called laboratory informatics, and an important category of software is a laboratory information management system (LIMS). A LIMS includes a database that records the work that has been carried out in the laboratory, and a workflow management system that tracks the progress of the work. Laboratory informatics plays an essential role in genomic sequencing and microarray projects, but sadly there are few recent publications or websites covering this topic.

After data have been obtained from the laboratory, there is generally a need for computation to extract signal from the raw data. In genomic sequencing using current instruments, the raw data comprise a waveform called a chromatograph indicating the intensity of light emitted by each of four dyes at each position of the sequence. A computation called 'base calling' is needed to convert this waveform into the desired signal, namely the most likely DNA base at each position along with a measure of confidence in the

call. In a microarray experiment, the raw data comprise an image showing the intensity of light emitted across the surface of the microarray; a computation called 'image analysis' separates the image into spots, and calculates the intensity of light that can be attributed to the RNA bound to the probe at each spot.

The next step is a series of relatively simple, but systematic computations that check for obvious errors, and convert data from one format to another. This is often called a processing pipeline.

In genomic sequencing, the processing pipeline checks for errors such as clones that lack inserts, bacterial contamination, and low-quality sequence. The human genome papers [1•,2•] describe the processing pipelines used in those projects. One of the major sequencing centers, the Sanger Institute, provides additional details on its website [15•], along with a comprehensive listing of available software [16].

In microarray experiments, the processing pipeline has a similar shape, but the details, of course, are completely different, looking for errors such as failed hybridizations and degraded RNA samples.

The data at this point are still highly dependent on the detailed way they were produced, and cannot be readily interpreted by non-specialists. The next phase seeks to raise the data from the province of technology into the realm of science.

In genomic sequencing, this phase assembles the millions of short sequences produced in the laboratory into longer contiguous stretches, called contigs. The contigs are then placed onto a map of the genome, if possible, to provide a larger scale assembly of the sequence. As more long contigs are placed onto the map, the overall genome sequence starts to take shape and it becomes possible to attempt genome annotation to identify genes and other biologically interesting elements in the sequence. The annotation process can reveal errors in the assembly, for example, if the sequence of a known gene is scrambled; such cases require investigation by the annotators and may even necessitate more sequencing.

As this process iterates, the annotations become more solid and eventually the results are good enough to release for use by the general community. At this point, scientists at large can begin the open-ended process of analyzing the data to discover new biology.

In microarray experiments, this phase normalizes the data to reduce the impact of systematic measurement errors, and then filters the data to eliminate 'uninteresting' genes, typically those with expression levels that do not change very much over the course of the experiment. Traditionally, these issues have been addressed using ad hoc methods, such as global scaling of expression measurements and arbitrary cut-offs based on the magnitude of the expression change, called fold

change. There is considerable current research seeking to replace these ad hoc methods with robust statistical techniques; a complete survey of this field is beyond our scope, but a sampling of such software is available on the Web [17–22].

## More on genomic sequencing

The human genome papers [1•,2•] describe in some detail the assembly and annotation procedures used by the respective teams. Another important resource for genomic sequence assembly is the phred/phrap/consed suite [23•] used by many public sequencing centers to assemble data from bacterial artificial chromosomes and other large insert clones.

There are actually two distinct assemblies of the public human genome and three complete annotations publicly available. The assembly discussed in [1•] was produced by a team led by Haussler at the University of California at Santa Cruz; additional information on their approach can be found on the Web [24•]. The annotation discussed in the paper was produced by Ensembl [25•], a joint project between the European Bioinformatics Institute (EBI) and the Sanger Institute led by Birney. A separate annotation of the same assembly was reported by Haussler's team [26•]. A second assembly was carried out independently by Schuler and colleagues at the US National Center for Biotechnology Information (NCBI) who also performed their own annotation [27•]. In addition, proprietary annotations of the public data are available from several companies.

An important aspect of annotation is gene prediction. Ouellette and colleagues [9••] present a detailed survey and comparison of gene prediction programs.

One surprising result from the annotation of the human genome sequence was the small number of genes found. All three public annotations and the private effort found some 30 000–40 000 genes, in sharp contrast to the estimates of 100 000 or more [28] that were widely believed before the genome sequence became available. Several papers have since appeared challenging various aspects of the annotation methodology and generally suggesting the gene count will ultimately be somewhat larger [29–31,32•,33].

The next major stage of genome annotation is comparison of genomes across a range of evolutionary distances. A large team led by Stubbs [34••] carried out a detailed comparison of human chromosome 19 with its corresponding regions in mouse; they were able to find confirmatory evidence for genes that were predicted to exist in the human genome annotation, as well as evidence for additional genes that were missed. Pufferfish is an excellent target for comparative analysis [35•] because of its compact genome (approximately 400 Mb compared with the 3 Gb of human or mouse) and low level of repetitive regions. It is likely to be especially valuable in identifying conserved regulatory sites. Early results comparing pufferfish with mouse have been recently reported [36]. A good overview of the informatics challenges in such comparisons is presented by Miller [37••].

## Software speciality

A final dimension of breadth concerns the variety of software issues that must be solved to create a usable computing system. This matter lies outside most biologists' range of experience and is often overlooked.

This includes databases to store the information, algorithms to analyze it, visualizations to help scientists understand the analysis, and user interfaces to afford scientists convenient access to all of these capabilities. Large software systems usually consist of many independently developed parts, and there is a need for data exchange mechanisms to move information among the components. Data integration is a related problem, but with the focus on combining information in scientifically valid ways. Workflow management is the software technology used for keeping track of tasks to be done in generating large datasets or in the automated analysis of such datasets. The technical aspects of these software issues are quite diverse, and involve very different computational expertise.

## Conclusions

The three dimensions of breadth — biological research area, the path from data to knowledge, and software speciality — combine to make bioinformatics a very broad field. Every point in this three-dimensional space corresponds to a unique bioinformatics problem. Naturally some points are more important than others, and as there are more points in space than investigators in the field, many important aspects are simply not addressed.

Bioinformatics will continue to evolve as new large-scale data-production technologies come into use. There is an easy trick to predict the future of this field: to know what is going to be hot in bioinformatics tomorrow, just look at which biotechnologies are coming on-line today.

## References and recommended reading

Papers of particular interest, published within the annual period of review, have been highlighted as:

• of special interest
•• of outstanding interest

1.  The Genome International Sequencing Consortium: **Initial**
•   **sequencing and analysis of the human genome.** *Nature* 2001, **409**:860-921.
This is the definitive publication on the public project to sequence the human genome. It describes the sequencing and analytical procedures that were employed, and biological insights gained from initial analysis of the data.

2.  Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG,
•   Smith HO, Yandell M, Evans CA, Holt RA *et al.*: **The sequence of the human genome.** *Science* 2001, **291**:1304-1351.
This is the definitive publication on the private project by Celera Genomics to sequence the human genome. Like [1•], it describes the procedures that were employed as well as biological results from the initial analysis.

3.  The Mouse Genome Sequencing Consortium: **Mouse Genome**
•   **Server.** URL: http://mouse.ensembl.org/
The official website of the public project to sequence the mouse genome.

4.  Celera Genomics: **Assembled and annotated human and mouse genomes**. URL: http://www.celera.com/genomics/commercial/ home.cfm?ppage=aahmgenomes

5.  DOE Joint Genome Institute, Fugu Genome Project: **Gene-rich**
•   **pufferfish DNA decoded.** URL: http://www.jgi.doe.gov/fugu/index.html
The official website of the US group that is leading the project to sequence the Japanese pufferfish, *Fugu rubripes*.

6.   Whitehead Institute: Center for Genome Research. *Tetraodon*
•    *nigroviridis* database. URL: http://www.genome.wi.mit.edu/annotation/
     tetraodon/
The official website of the major US group involved in sequencing the fresh-
water pufferfish *Tetraodon nigroviridis*.

7.   The Chipping Forecast. *Nat Genet Suppl* 1999,
••   **21**:1-60.
This supplement to *Nature* is now a bit out-of-date, but provides a good
overview of microarray technology and analysis.

8.   Grills G, Griffin C, Massimi A, Lilley K, Knudtson K, VanEe J:
•    **2000/2001 ABRF Microarray research group study: a current
     profile of microarray laboratories.** URL: http://abrf.org/ResearchGroups/
     Microarray/EPosters/MARG_poster_2001.pdf
Annual survey of microarray practice. Not a proper scientific survey, but
provides a glimpse of products and methods being used in the field.

9.   Rogic S, Mackworth AK, Ouellette FBF: **Evaluation of gene-finding
••   programs on mammalian sequences.** *Genome Res* 2001,
     **11**:817-832.
An excellent review and comparison of seven leading gene prediction programs:
FGENES, GeneMark, Genie, Genscan, HMMgene, Morgan, and MZEF. The
article evaluates these programs on a test dataset using several different metrics.
Broadly, the comparison indicates that the leading programs have improved
considerably over the past five years, but even the best programs still make too
many mistakes to be used for fully automatic genome annotation.

10.  National Center for Biotechnology Information: **UniGene.**
•    URL: http://www.ncbi.nlm.nih.gov/UniGene/
The official website of UniGene, the *de facto* gold standard database of
clustered ESTs and other transcribed sequences.

11.  The Institute for Genomic Research: **TIGR Gene Indices.**
•    URL: http://www.tigr.org/tdb/tgi.shtml
This website provides access to databases of clustered ESTs and other
transcribed sequences from a wide range of organisms.

12.  South African National Bioinformatics Institute: **The STACK project.**
•    URL: http://www.sanbi.ac.za/
The website for a leading research project and database of clustered transcripts.

13.  European Bioinformatics Institute: **InterPro.**
•    URL: http://www.ebi.ac.uk/interpro/
The official website of the most comprehensive database of protein motifs.
The site also provides a database showing the motifs in known proteins, and
tools for finding motifs in new sequences.

14.  Quackenbush J: **Computational analysis of microarray data.**
••   *Nat Rev Genet* 2001, **2**:418-427.
An in-depth and very current survey of computational methods for micro-
array analysis. It covers the entire microarray process from probe selection
to data analysis. The paper also includes a detailed discussion of clustering
methods, and illustrates the methods on a sample dataset.

15.  The Wellcome Trust Sanger Institute: **Production sequencing
•    software.** URL: http://www.sanger.ac.uk/Software/sequencing/
This website describes in some detail the processing pipeline used by the
major genome sequencing center at the Sanger Institute. This is a must-read
for anyone setting up a large-scale sequencing operation.

16.  The Wellcome Trust Sanger Institute: Software.
     URL: http://www.sanger.ac.uk/Software/

17.  University of California at Berkeley. **SMA: statistics for microarray
     analysis.** URL: http://www.stat.berkeley.edu/users/terry/zarray/
     Software/smacode.html

18.  The Jackson Laboratory: **MA-ANOVA programs for microarray data.**
     URL: http://www.jax.org/research/churchill

19.  University of Wisconsin: **GEDA: gene expression data analysis.**
     URL: http://www.biostat.wisc.edu/geda/eba.html

20.  University of Wisconsin: **Microarray.zip.**
     URL: http://www.stat.wisc.edu/~yandell/statgen/tr1031.html

21.  Institute for Systems Biology: **Variability and ERror assessment
     and SAM: significance of array measurement.**
     URL: http://www.systemsbiology.org/VERAandSAM/?id=yvfw4

22.  University of California at Irvine: **National Center for Genome
     Resources. Cyber T.** URL: http://genebox.ncgr.org/genex/cybert/

23.  The Phred/Phrap/Consed System Home Page:
•    URL: http://www.phrap.org/
The official website for a suite of software that is widely used by sequencing
groups to analyze chromatograms from raw reads, assemble raw reads into
longer contiguous regions, and to edit and review the assemblies. The suite
includes phred for base calling, phrap for sequence assembly, consed for
visual review and editing of assemblies, autofinish to automate the choice of
sequencing reads needed to close gaps and fix problems in the assembly,
and swat and crossmatch for sequence comparison.

24.  GigAssembler: an algorithm for the initial assembly of the human
•    genome working draft. URL: http://genome.cse.ucsc.edu/goldenPath/
     algo.html
This website describes the sequence assembly method used by Haussler
and colleagues to assemble the human genome.

25.  European Bioinformatics Institute and the Sanger Institute: **Project
•    Ensembl.** URL: http://www.ensembl.org/
This website provides a detailed explanation of the assembly and annotation meth-
ods used by this group to create its widely used version of the human genome.

26.  University of California at Santa Cruz: **Human genome project
•    working draft.** URL: http://genome.cse.ucsc.edu/
This website provides access to the annotation of the human genome
carried out by Haussler and colleagues.

27.  National Center for Biotechnology Information: contig assembly and
•    annotation process. URL: http://www.ncbi.nlm.nih.gov/genome/
     guide/build.html
The website of Project Ensembl who performed the official annotation of
the public human genome sequence.

28.  Liang F, Holt I, Pertea G, Karamycheva S, Salzberg SL, Quackenbush J:
     **Gene index analysis of the human genome estimates
     approximately 120,000 genes.** *Nat Genet* 2000, **25**:239-240.

29.  Salzberg SL, White O, Peterson J, Eisen JA: **Microbial genes in the
     human genome: lateral transfer or gene loss?** *Science* 2001,
     **292**:1903-1906.

30.  Gopal S, Schroeder M, Pieper U, Sczyrba A, Aytekin-Kurban G,
     Bekiranov S, Fajardo JE, Eswar N, Sanchez R, Sali A *et al*: **Homology-
     based annotation yields 1,042 new candidate genes in the
     *Drosophila melanogaster* genome.** *Nat Genet* 2001, **27**:337-340.

31.  Karlin S, Bergman A, Gentles AJ: **Genomics: annotation of the
     Drosophila genome.** *Nature* 2001, **411**:259-260.

32.  Hogenesch JB, Ching KA, Batalov S, Su AI, Walker JR, Zhou Y,
•    Kay SA, Schultz PG, Cooke MP: **Related articles a comparison of
     the Celera and Ensembl predicted gene sets reveals little overlap
     in novel genes.** *Cell* 2001, **106**:413-415.
Compares the annotations of the public and private human genome efforts.
Reveals that even though the two annotations contain similar numbers of
genes, the specific genes in the two sets are rather different.

33.  Wright FA, Lemon WJ, Zhao WD, Sears R, Zhuo D, Wang J-P, Yang H-Y,
     Baer T, Stredney D, Spitzner J *et al*: **A draft annotation and overview
     of the human genome.** *Genome Biol* 2001, 2:research0025.

34.  Dehal P, Predki P, Olsen AS, Kobayashi A, Folta P, Lucas S, Land M,
••   Terry A, Ecale Zhou CL, Rash S, Zhang Q *et al*: **Human
     chromosome 19 and related regions in mouse: conservative and
     lineage-specific evolution.** *Science* 2001, **293**:104-111.
A groundbreaking, detailed comparison of human chromosome 19 and its
homologous 15 regions in mouse. The investigators aligned the human and
mouse sequences and found more than 12 000 highly conserved sequence
elements, which they termed conserved sequence elements. They found that
80% of exons from known genes on human chromosome 19 are conserved
in the mouse, but these comprise only about 40% of all conserved sequence
elements. Another 25% of conserved sequences are similar to ESTs or
genes from other species. Of the remaining conserved sequences, about
85% are within 5 kb of known genes, suggesting that they are either novel
exons or regulatory elements. They also found that single-copy genes are
very highly conserved, while multicopy gene families show more variability.

35.  Venkatesh B, Gilligan P, Brenner S: **Fugu: a compact vertebrate
•    reference genome.** *FEBS Lett* 2000, **476**:3-7.
Explains the rationale for sequencing pufferfish and how this data might be
used to augment research on other organisms.

36.  Clarke D, Elgar G, Clark MS: **Comparative analysis of human 19p12–13
     region in Fugu and mouse.** *Mamm Genome* 2001, **12**:478-483.

37.  Miller W: **Comparison of genomic DNA sequences: solved and
••   unsolved problems.** *Bioinformatics* 2001, **17**:391-397.
An insightful essay on the new genre of sequence analysis problems that
remain to be solved in the era of whole genome sequencing. These include
better methods for aligning two or more long genomic sequences, rigorous
means of evaluating these alignments, more informative ways of visualizing
them, and improved gene prediction programs.