Bioinformatics

Introductory article

Wojciech Makalowski, *Pennsylvania State University, Philadelphia, Pennsylvania, USA*

Bioinformatics may be defined as the development and/or application of computational tools and approaches for expanding the use of biological data, including those to acquire, store, organize, archive, analyze or visualize such data.

Article contents

- Introduction
- Databases
- Sequence Analysis
- Nucleotide Sequence Assembly
- Gene Prediction
- Homology Search
- Protein Analysis
- Technical Aspects
- Education
- Conclusions

doi: 10.1038/npg.els.0005247

Introduction

Molecular biologists generate data at an unprecedented speed (**Figure 1**). For example, an average bacterial genome can be sequenced in one day by a sequencing center (factory) and the analysis of the expression of hundreds of genes in a single experiment has become routine. The demands and opportunities for interpreting the data are expanding more than ever. In this context, a new discipline – bioinformatics – has emerged as a strategic frontier between biology, computer science and statistics. Although the term 'bioinformatics' does not appear in the literature until around 1991, since the 1960s, long before anyone thought to label this activity with a special term, some evolutionarily oriented biologists (e.g. Margaret O. Dayhoff, Russell F. Doolittle, Walter M. Fitch and Masatoshi Nei) have been building databases, developing algorithms and making biological discoveries by sequence analysis. The field is not mature and therefore not very well defined. Almost anybody active in this field has a different definition of bioinformatics and the discussion of what does and what does not belong to it may go for hours.

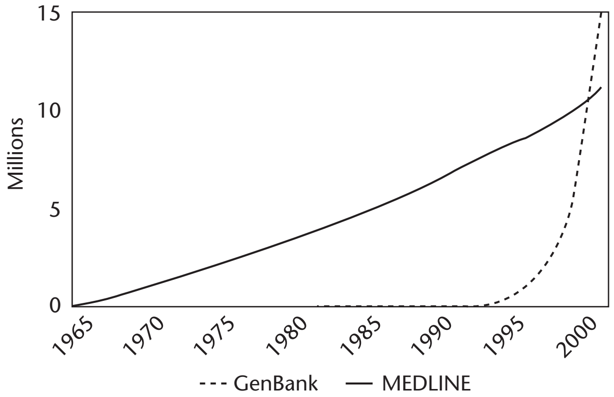

Figure 1 Growth of biological information measured by number of entries in two major databases: GenBank – a collection of annotated nucleotide sequences and MEDLINE – a collection of biomedical literature.

Classical bioinformatics

Fredj Tekaia at the Institut Pasteur offers this definition of bioinformatics: 'The mathematical, statistical and computing methods that aim to solve biological problems using deoxyribonucleic acid (DNA) and amino acid sequences and related information.' Most biologists talk about 'doing bioinformatics' when they use computers to store, retrieve, analyze or predict the composition or the structure of biomolecules, for example, nucleic acids and proteins. These are the concerns of 'classical' bioinformatics, dealing primarily with sequence analysis. It is a mathematically interesting property of most large biological molecules that they are polymers: ordered chains of simpler molecular modules called monomers. Each monomer molecule is of the same general class, but each kind of monomer has its own well-defined set of characteristics. Many monomer molecules can be joined together to form a single, far larger, macromolecule, which has exquisitely specific informational content and/or chemical properties. Therefore, the monomers in a given macromolecule of DNA or protein can be treated computationally as letters of an alphabet put together in preprogrammed arrangements to carry messages or do work in a cell.

New bioinformatics

The greatest achievement of biology, the Human Genome Project (HGP), has recently been completed. The project brought not only the ultimate blueprint of our genetic material but also rapid development of new technologies, for instance microarrays. As we enter the 'postgenomic' era, the nature and priorities of bioinformatics research and applications are changing, for instance statistic analysis plays a more prominent role.

Databases

Databases are an invaluable component of bioinformatics. They can be defined as large organized sets of data usually linked with software that allows database searching, extraction of the relevant information and the update of their content. A good database should enable easy access to the data and allow the extraction of information in a way that all required data will be extracted and none of unwanted data will be retrieved.

Biological databases can be divided into primary and derivative databases. Primary databases store the original information submitted by experimentalists. Database staff organize but do not add additional or alter information submitted by a biologist. Three nucleic acid databases, DDBJ, EMBL and GenBank, are the best known examples of primary databases. Derivative databases may be human curated or computationally derived. In such a case, they usually represent a subset of primary database with added value, such as correction of the data or adding some new information. SWISS-PROT and UniGene represent human-curated and computer-generated derivative databases respectively.

Sequence Analysis

Development of technologies to determine the sequence of the nucleotides in nucleic acids or amino acids in proteins has created the need for sequence analysis. It is a process of investigating the information content of raw sequence data. This is not a trivial task as the information may be obscured from the researcher by long stretches of sequences that do not carry any information (this is particularly true for eukaryotic genomic sequences).

Nucleotide Sequence Assembly

The automation of DNA sequencing was a major contributing factor to the success of the HGP in the first place. Despite all improvements, DNA sequencing has one crucial limitation – no more than several hundreds of nucleotides can be read in a single reaction. To determine the sequence of a longer molecule, it has to be chopped into smaller pieces, sequenced and then the whole sequence of the long molecule has to be deduced based on the nucleotide sequence of shorter molecules. The process of the long sequence deduction is called sequence assembly. Both public and private sequencing projects used a shotgun technique to produce raw sequencing data. The difference lies in the size of the human genome fragments that are subjected to the shotgun sequencing process – about 200 kb long bacterial artificial chromosomes (BACs) in the public project and the whole human genome in the private one. The raw data, after some cleaning, are used to assemble longer contigs. First, all the sequences are checked for the overlaps. The information about overlapping fragments can be summarized in a graph structure. Then, the problem of sequence assembly is to find a path that covers all fragments. Unfortunately, this is an NP-hard problem (which means that it requires non-deterministic polynomial time to be solved) and therefore all known algorithms are relatively slow. Recently, a Eulerian approach (one of the algorithms that uses graph theory) has been proposed to solve the assembly problem but it remains to be shown if this approach is applicable to large eukaryotic genomes. One of the greatest challenges of the human genome assembly is the recently duplicated fragments of a genome. These fragments, which can be as long as 10 kb, have a sequence similarity higher than a sequencing error rate. Therefore, the assembly software treats them as an identical sequence. This causes the known problem of collapsing legitimate duplicons into one fragment. On the other hand, it is not uncommon that some fragments are 'artificially' duplicated by the assembly software, placing the same fragment in two distinct chromosomal locations.

Gene Prediction

The assembled genome is the starting point, not the goal, of the sequencing projects. The linear sequence is useless unless annotated. The annotation process reveals biological information hidden in the nucleotide (or amino acid) sequence. Only approximately 2% of the human genome codes for proteins, and the protein-coding regions are not contiguous – exons are interlaced with introns. Many techniques were developed to identify exons. Computational methods can be divided into homology-based or model-based. The former methods work well if a homologous sequence of a given gene is already known for other organisms. The latter do not require any previous knowledge about a given gene. Since many prokaryotic genomes have been sequenced in the last several years, the homology-based approach is a first choice for these organisms. In fact, most of the genes of the newly sequenced prokaryotic genomes can be determined in this way. Although for eukaryotic genomes model-based methods are still dominating, a comparative genomic approach can add valuable information to any method applied. None of the existing methods are perfect and we are far from solving the gene prediction problem. In practice, several methods should be applied and experimental validation is required. Luckily, it seems that computational methods can at least point to the protein-coding region, but precise boundaries of the exons are still quite difficult to predict. The major problem is that with increased sensitivity, specificity declines. Therefore, to avoid too many false positives we have to give up some sensitivity and allow a fraction of exons not to be discovered.

About 5% of mammalian genes do not code for proteins. In such a case, a transcribed molecule is a functional ribonucleic acid (RNA), for example, ribosomal RNAs (rRNAs), transfer RNAs (tRNAs) and small nuclear RNAs (snRNAs). Identifying these noncoding RNAs is much more complicated than identifying protein-coding genes. They are usually small, often present in multiple copies, have conserved a secondary structure rather than a linear sequence, and can be found in unusual locations, such as introns of other genes. A mathematical model called 'stochastic context-free grammars' (SCFGs) has been introduced as a framework for RNA structure and sequence alignment. A successful implementation of this approach is the program tRNAscan-SE, which detects tRNA genes with high sensitivity.

Homology Search

One of the first questions asked by a biologist after a new nucleotide molecule has been sequenced is whether a sequence is already known. The best way to find this out is to compare a new sequence to all known nucleotide sequences. With the rapid growth of repository databases, an exhaustive comparison of the database is practically impossible, for example comparison of 1000 nt sequence with the whole GenBank (2×10^{11} nt) would take 5555 h (232 days) on a single CPU computer. The key idea of the homology search is that most of the sequences in the database do not match a query sequence (the sequence that is used for the database search). Therefore, finding a fast way to eliminate sequences that do not match would speed up the whole process of the database comparison. There are several methods for the database screening, but two of them, FASTA and BLAST, have dominated the whole field (the original paper describing BLAST algorithm with almost 14 000 citations is by far the most cited paper in the history of biology). The influence of both programs is so profound that all database searching programs have converged to a basic format: a graphical description of the results, a list of top scoring sequences from the database and a series of alignments for some of the top scoring sequences. The common strategy of the database searching program is fast screening to eliminate unrelated sequences followed by a complete local alignment of top scoring sequences. The important part of the process is a statistical assessment of the obtained alignment, in other words estimating the probability of obtaining a given alignment by chance. Different flavors of programs can compare nucleotide or amino acid sequence against nucleic acid or protein databases. Interestingly, because of the larger alphabet and more sophisticated scoring systems, protein-based searches are usually more sensitive than nucleotide searches.

Protein Analysis

Once the amino acid sequence of a protein is determined, a biologist would like to know the function of that protein in a cell. Protein function can be inferred based on an overall similarity to other proteins of known function or based on the domains or motifs present in a previously analyzed protein. Finally, since structure is the most evolutionarily conserved feature of a protein, structural similarity between proteins may suggest functional similarity. Protein similarity searches, for example using BLAST, against a comprehensive database is the easiest and most frequently utilized way to infer function. However, there are some traps awaiting in the similarity searches. Annotation of a protein in a database might be faulty. For example, the annotation of a double-functioning protein may be missing one of its functions, two different proteins may have two different names in the database, the annotation of one protein may be mistakenly propagated on other proteins. A protein may also change its function during the evolution. For instance, birds' lysozyme and mammalian lactalbumin are homologous despite their completely different function. Finally, protein function may depend on the place where it is expressed.

Common domains or motifs shared between proteins with relatively low (less than 20%) similarity may also shed light on unknown protein function. Domains, such as DNA binding domains, are functional units located in a particular region of a protein that perform a specific function. Functional domains usually coincide with structural domains. Motifs are sequence patterns common among proteins performing a given function. A given motif presence does not infer a common ancestry of proteins bearing it, in other words, two proteins may possess the same motif because of convergent evolution. Several computational methods exist to define motifs in amino acid sequences from the same protein family. The simplest way is to look for the conserved amino acids in columns of a multiple sequence alignment of the family. Other methods do not require multiple sequence alignment and include machine learning approach, the hidden Markov models (HMM) and Gibbs sampling approach. Newly sequenced genomes bring a high number of proteins with unknown function. For instance, out of 25 498 predicted genes in a weed thale cress (*Arabidopsis thaliana*), only 30% have known function and only 9% of proteins have their function confirmed experimentally. (***See*** Gibbs Sampling and Bayesian Inference; Hidden Markov Models.)

For proteins without similarity to any known protein, further computational analysis has to be done in order to obtain more biological information on analyzed proteins. It is possible to predict transmembrane proteins based on physicochemical properties of amino acids forming a peptide chain. Statistical analysis can help find a signal peptide. Since the function of a protein is determined by its three-dimensional structure, structurally similar proteins are likely to perform similar function. As the number of known protein structures continue to expand (as of 20 August 2002, there are 18 488 biological macromolecule structures deposited in Protein Data Bank (PDB)), the chance that a new gene product folds like a known structure will continue to increase. Paradoxically, the database noise continues to grow as well, and finding the true matching fold may not be a trivial task. One of the techniques to find similar structure is threading. This method is based on the assumption that proteins are in a state of minimum free energy, and that this energy may be computed for any given structure. The energy computation takes into account the compatibility of different amino acids at each position in the structure. Given a function that can evaluate the compatibility of a sequence with a structural template, threading algorithms attempt to minimize this function by considering various possible sequences to structure alignments.

Technical Aspects

Most of the tasks in bioinformatics have to handle large data sets and the HGP is not an exception. Very few of them could be easily completed on the personal computers and operating systems usually favored by biologists, that is, MacOS on an Apple Macintosh or MS Windows on an Intel-based PC. Consequently, Unix platforms are first choice for bioinformaticians and most of the software related to the HGP was developed under this operating system. The major features of Unix that determined its popularity among bioinformaticians are: support of high-end hardware with large memory and robust multithreading, support for large (>2 GB) data files, availability of efficient scripting and data manipulation languages (pearl, awk, python and unix shell are widely used for that) and finally general stability of the operating system – a well-maintained Unix machine needs rebooting only for major hardware updates. The major providers of high-end computers and operating system to the bioinformatics community have been Sun Microsystems, Silicon Graphics Inc. and Compaq. IBM has recently invested in hardware and software development for the life sciences as well. Linux, a freeware version of Unix running on Intel-based machines, has become very popular among bioinformaticians especially for distributed computing (PC clusters or farms).

PC farms, because of the large numbers in which they are manufactured, have a better cost/performance ratio than high-end machines. The main drawback of this approach is that the system management of PC clusters is quite difficult and they are prone to failure if components are not chosen carefully. Nevertheless, commercial solutions to the problem exist and one can buy ready-to-run PC clusters. Also, commercial versions of some crucial software, ready to run on PC clusters, are available as well. These include BLAST, hidden Markov models, sensitive sequence alignment (implementations of Smith–Waterman algorithm) and gene discovery systems. The Grid computing will be likely the next step in the distributed computing within the ever CPU-hungry bioinformatics community, especially that major players such as Sun Microsystems make a serious effort to bring effortless solutions to this approach.

Although the internet was not invented for the HGP, it did provide the right infrastructure at the right time. Biologists depend heavily on the internet for data dissemination and collaborative work. The websites provide easy and convenient access to the data and visualization of the data analysis results. Many bioinformatics tools and HGP results are easily accessible to the whole community, including those using personal computers. (See **Table 1** for a non-comprehensive list of bioinformatics resources on the web.)

Education

Despite the huge demand for the trained bioinformaticians, until recently there were no formal training programs in bioinformatics. Practically, all bioinformaticians are either computer scientists who have some idea about biology, or biologists who know how to program. Interestingly, molecular evolutionary geneticists make an especially easy transition, most likely because of their background and training. Some physicists, statisticians and mathematicians moved to the field attracted by either interesting, difficult to solve biological problems or simply by better job prospects. There is an ongoing discussion of who makes a better bioinformatician – retrained biologists

Table 1 Some bioinformatics resources on the web

Site	Address	Description
General sites and compilations of resources		
NCBI	http://www.ncbi.nlm.nih.gov/	Three homes of nucleic acid repository databases and gateways to many resources
EBI	http://www.ebi.ac.uk/index.html	
DDBJ	http://www.ddbj.nig.ac.jp/	
KEGG	http://www.genome.ad.jp/kegg/kegg3.html	Kyoto encyclopedia of pathways and maps
NAR	http://www3.oup.co.uk/nar/database/c/	Compilation of databases
ExPASy	http://www.expasy.ch/	ExPASy Molecular Biology Server
TreeLife	http://tolweb.org/tree/phylogeny.html	Tree of life taxonomy project
BCM Launcher	http://searchlauncher.bcm.tmc.edu/	Launches numerous nucleotide and protein tools
EMBL Tools	http://www.embl-heidelberg.de/Services/index.html	Various Computational Services at EMBL-Heidelberg
The Biology WorkBench	http://workbench.sdsc.edu/	Unified area for searching and analysis at San Diego Supercomputer Center (requires a free account)
MDC	http://www.bioinf.mdc-berlin.de/	Bioinformatics tools at Delbruck Center
Posnania	http://posnania.biotec.psu.edu/	Bioinformatic tools server at Penn State
Database searches and alignment tools		
BLAST	http://www.ncbi.nlm.nih.gov/BLAST	NCBI gateway to db searches
FASTA	http://www.ebi.ac.uk/fasta33/	EBI web interface to FASTA searches
BlastU	http://www.proweb.org/proweb/Tools/WU-blast.html	BLAST query against user FASTA database
Blat	http://genome.cse.ucsc.edu/cgi-bin/hgBlat?command = start&db = hg12	Quick alignment of query to human genome assembly
WABA	http://www.cse.ucsc.edu/ ~ kent/xenoAli/xenAliTwo.html	Genomic to genomic comparisons
Pipmaker	http://bio.cse.psu.edu/pipmaker/	Per cent identity plot alignment of two genomic sequences
ClustalW	http://www.ebi.ac.uk/clustalw/	General purpose multiple sequence alignment program
Mulatalin	http://prodes.toulouse.inra.fr/multalin/multalin.html	Aligns multiple sequences with flexible formatting
Genome browsers and 'complete' genomes		
Genome Browser	http://genome.cse.ucsc.edu/cgi-bin/hgGateway?db = hg12	Human genome assembly and track browser at UCSC
Ensembl	http://www.ensembl.org/	Ensembl Genome Browser
Genomic Biology	http://www.ncbi.nlm.nih.gov/Genomes/index.html	Genomic resources at NCBI
euGenes	http://iubio.bio.indiana.edu:8089/man/	Five species map browser and 37 049 human gene reports
Vista	http://sichuan.lbl.gov/ids-bin/VistaInput	Visualization tool for long genomic alignments
TIGR CMR	http://www.tigr.org/tigr-scripts/CMR2/CMRHomePage.spl	Comprehensive microbial resource and genome browser
Sanger	http://www.sanger.ac.uk/HGP/links.shtml	Human genome project at Sanger Centre
Jax	http://www.informatics.jax.org/	Mouse genomics at Jackson Laboratory
RatMap	http://ratmap.gen.gu.se/	Rat genome database at Goteborg
Zfin	http://zfin.org/cgi-bin/webdriver?MIval = aa-ZDB_home.apg	Zebra fish information network and database project
Exofish	http://www.genoscope.cns.fr/externe/tetraodon/	Genome analysis of *Tetraodon nigroviridis*
Flybase	http://flybase.bio.indiana.edu/	Database of *Drosophila* genomics at Indiana
WormBase	http://www.wormbase.org/	Nematode mapping, sequencing and phenotypic repository

Table 1 Continued

Site	Address	Description
SGD	http://genome-www.stanford.edu/Saccharomyces/	Yeast genome database
TAIR	http://arabidopsis.org/	The *Arabidopsis* information resource and tool center
GOLD	http://igweb.integratedgenomics.com/GOLD/	Monitors genome projects worldwide
Gene prediction and feature finding		
GenomeScan	http://genes.mit.edu/genomescan.html	Predicts genes incorporating protein homology
RGD	http://rgd.mcw.edu/	Comparative mapping tool for rat–mouse–human synteny
FGENEH	http://genomic.sanger.ac.uk/gf/gf.shtml	Splice sites, coding exons, gene models, promoter and poly A
Softberry	http://www.softberry.com/berry.phtml	Nucleotide sequence analysis, genes, promoters
TwinScan	http://genes.cs.wustl.edu/query.html	Gene prediction for eukaryotic genomic sequences
WebGene	http://www.itba.mi.cnr.it/webgene/	Tools for analysis of protein-coding gene structure
GrailEXP	http://grail.lsd.ornl.gov/grailexp/	Exons, genes, promoters, poly A, CpG islands at ORNL
GeneWise2	http://www.sanger.ac.uk/Software/Wise2/genewiseform.shtml	Compares protein to genomic with introns and frameshifts
Genie	http://www.fruitfly.org/seq_tools/genie.html	Finds multiexon genes only, trained on human
AAT	http://genome.cs.mtu.edu/aat.html	Analysis and annotation tool for finding genes
HMMgene	http://www.cbs.dtu.dk/services/HMMgene/	Prediction of vertebrate and nematode genes
NetGene2	http://www.cbs.dtu.dk/services/NetGene2/	Neural network predictions of splice sites
GeneMark	http://www.ebi.ac.uk/genemark/	Gene prediction server at EBI
GeneID	http://www1.imim.es/geneid.html	Gene prediction of selected signals and exons
RNAGenie	http://gobi.lbl.gov/~steveh/rnagene/	Locates RNA genes from secondary structures
Vienna RNA	http://www.tbi.univie.ac.at/~ivo/RNA/	Seven RNA analysis tools
PromoterScan	http://bimas.dcrt.nih.gov/molbio/proscan/index.html	Predicts promoters using homologies with Pol II promoters
ORF Finder	http://www.ncbi.nlm.nih.gov/gorf/gorf.html	Open reading frame finder for ORFs of cutoff size
Emboss	http://www.ebi.ac.uk/emboss/cpgplot/	Predicts CpG islands and isochores
UTR	http://bighost.area.ba.cnr.it/BIG/UTRHome/	Untranslated regions of eukaryotic mRNAs
AltSplice	http://141.80.80.48/splice/	Database of alternate splicing in expressed sequence tags and disease genes
Webcutter	http://www.firstmarket.com/firstmarket/cutter/cut2.html	Finds restriction endonuclease sites
FSED	http://ir2lcb.cnrs-mrs.fr/d_fsed/fsed.html	Frameshift error detection in new sequences
ERR_WISE	http://www.bork.embl-heidelberg.de/ERR_WISE/	Detection of frameshift sequencing errors
TfScan	http://bioweb.pasteur.fr/seqanal/interfaces/tfscan.html	Scans DNA sequences for transcription factors
RepeatMasker	http://ftp.genome.washington.edu/cgi-bin/RepeatMasker	Finds retroposons and repeats
Censor	http://www.girinst.org/Censor_Server-Data_Entry_Forms.html	Finds repeated elements, Repbase updates
Tandem	http://tandem.biomath.mssm.edu/trf/trf.html	Tandem repeat finder in DNA
TandemRepeats	http://c3.biomath.mssm.edu/trf.submit.options.html	Finds adjacent imperfect repeat patterns in DNA
Tnoco	http://www.biophys.uni-duesseldorf.de/local/TINOCO/tinoco.html	Suggests secondary structure in RNA or DNA
RNA_align	http://www.csd.uwo.ca/~kzhang/rna/rna_align.html	Aligns two RNA species from secondary and tertiary structures
Protein tools		
SwissTools	http://www.expasy.ch/tools/	Large collection of protein tools and databases
SAPS	http://www.ebi.ac.uk/saps/	Statistical analysis of protein sequences

Table 1 Continued

Site	Address	Description
Signalp	http://www.cbs.dtu.dk/services/SignalP/	Predicts signal peptide cleavage sites
TMpred	http://www.ch.embnet.org/software/TMPRED_form.html	Predicts transmembrane segments and orientation
TMHMM	http://www.cbs.dtu.dk/services/TMHMM/	Predicts transmembrane helices
TMAP	http://www.mbb.ki.se/tmap/index.html	Predicts membrane proteins from multiple alignment
COILS	http://www.ch.embnet.org/software/COILS_form.html	Prediction of coiled-coil regions in proteins
PREDATOR	http://www-db.embl-heidelberg.de/jss/servlet/de.embl.bk.wwwTools.GroupLeftEMBL/argos/predator/predator_info.html	Secondary structure from multiple sequences
SSCP	http://www.bork.embl-heidelberg.de/SSCP/	Predicts helix, strand and coil from composition
STRIDE	http://www-db.embl-heidelberg.de/jss/servlet/de.embl.bk.wwwTools.GroupLeftEMBL/argos/stride/stride_info.html	Secondary structure from atomic coordinates
Jpred	http://jura.ebi.ac.uk:8888/	Consensus method for secondary structure prediction
Interpro	http://www.ebi.ac.uk/interpro/	Integrated resource of protein families, domains and sites
Pfam	http://www.cgr.ki.se/Pfam/	Alignments and hidden Markov models of protein domains
Sift	http://blocks.fhcrc.org/ ~ pauline/SIFT.html	Predicts tolerated and untolerated protein substitutions
TransFac	http://transfac.gbf.de/TRANSFAC/index.html	Transcription factor database and tool collection
SMART	http://smart.embl-heidelberg.de/	Simple modular architecture research tool
ProDom	http://prodes.toulouse.inra.fr/prodom/doc/prodom.html	Protein domain database
ScanProsite	http://www.expasy.org/tools/scanprosite/	Scans a sequence against Prosite
ProfileScan	http://hits.isb-sib.ch/cgi-bin/PFSCAN	Scans sequence against profile databases
HMMER	http://hmmer.wustl.edu/	Profile hidden Markov models for sequence analysis
SAM	http://www.cse.ucsc.edu/research/compbio/sam.html	Sequence alignment and modeling system by HMM
Blimps	http://bioweb.pasteur.fr/seqanal/motif/blimps-uk.html	BLOCKS search tool
Motif	http://motif.genome.ad.jp/	Searches for motifs in query sequence
Scop	http://scop.mrc-lmb.cam.ac.uk/scop/	Structural classification of proteins
Pratt	http://scop.mrc-lmb.cam.ac.uk/scop/	Search for conserved patterns in protein sequences
MAST	http://meme.sdsc.edu/meme/website/mast.html	Motif alignment search tool
Profam	http://mips.gsf.de/proj/protfam/	Curated protein classification and homology domains
Dali	http://www.ebi.ac	Coordinates of query compared to PDB database
123D	http://genomic.sanger.ac.uk/123D/run123D.shtml	Predicts tertiary structure
PredictProt	http://dodo.cpmc.columbia.edu/pp/submit_adv.html	Threading, secondary, solvent and transmembrane
CE	http://cl.sdsc.edu/ce/all-to-all/all-to-all-r.html	Finds structural alignments from PDB
SCWRL	http://www.fccc.edu/research/labs/dunbrack/scwrl/	Side-chain placement using a rotamer library
AA Contacts	http://promoter.ics.uci.edu/BRNN-PRED/	Residue contacts, solvent accessibility
SwissPdbView	http://www.expasy.ch/spdbv/	Deep view for manipulation and analysis of PDB structures
Kinemages	http://www.faseb.org/protein/kinemages/kinpage.html	Interactive three-dimensional protein web display
RasMol	http://www.umass.edu/microbio/rasmol/index2.htm	Molecular visualization resource
Chime	http://www.umass.edu/microbio/chime/	Interactive web plug-in for 3D molecular structures

Table 1 Continued

Site	Address	Description
COMBOSA3D	http://bioinformatics.org/combosa3d/index_b.html	Combines alignment, protein three-dimensional and Chime to color
Boxshade	http://www.ch.embnet.org/software/BOX_form.html	Pretty shading of multiple alignments

Table 2 Selected graduate programs in bioinformatics

University	URL	Short description
Boston University	http://bioinfo.bu.edu/	Offers an MSc and a PhD in bioinformatics. The program has 30 faculties and a number of centers and departments involved in the program
George Mason University	http://www.krasnow.gmu.edu/bioinformatics/bioinfresrch.htm	GMU offers a PhD in bioinformatics and computational biology
University of Manchester School of Biological Sciences	http://bioinf.man.ac.uk/	This college offers an MSc in bioinformatics and computational molecular biology
Pasteur Institute	http://www.pasteur.fr/formation/infobio/infobio-en.html	The bioinformatics courses at the Pasteur Institute are organized with the collaboration of the two largest Parisian scientific universities: Paris VI 'Pierre et Marie Curie' and Paris VII 'Denis Diderot'
University of Pennsylvania	http://www.cbil.upenn.edu/UPCB/	This program admits students into standard PhD programs of either the Computer and Information Science department, or a biological department, but then provides additional training in computational biology
Rutgers University	http://cmb.rutgers.edu/	This joint effort of Rutgers and the University of Medicine and Dentistry of New Jersey offers a PhD program
S-star	http://s-star.org/main.html	Online bioinformatics program
The WM Keck Center for Computational Biology in Houston, Texas	http://www.keckcenter.org/training.cfm	The center offers studies in computational biology through three partner institutions: Baylor College of Medicine, Rice University and the University of Houston
University of Toronto	http://p-b.med.utoronto.ca/	Multidepartment and affiliated research institutes' academic program
Iowa State University	http://www.bcb.iastate.edu/	An interdisciplinary PhD program in bioinformatics and computational biology (BCB)
University of the Sciences in Philadelphia	http://tonga.usip.edu/zauhar/bioinformatics_program.html	MSc program in bioinformatics
University of Bielefeld, Germany	http://www.cebitec.uni-bielefeld.de/GradSchool/	International Graduate School in bioinformatics and genome research
Chalmers University of Technology, Gothenburg, Sweden	http://www.md.chalmers.se/Stat/Bioinfo/Master/	International Master's programme in bioinformatics

or retrained computer scientists. This dilemma will most likely disappear soon, as a cohort of scientists who have been trained as professional bioinformaticians leave different universities with newly established degrees in bioinformatics and have a background in these traditionally disparate disciplines.

Most of the universities that want to keep up to date with modern trends have established some programs in bioinformatics (see **Table 2**). Balancing a background in biology with a background in computer science is not easy. As it is unlikely that we can train bioinformaticians with an exhaustive background in both biology and computer science, we will continue to produce two types of bioinformaticians: biology oriented and more computer science oriented. The first group will be better suited for the biological data analysis, the latter for more traditional computer science tasks, for example algorithms development, database design and software engineering. As bioinformatics analysis increasingly depends on statistical methods, statistics should be a significant component of bioinformatics education. The Canadian Genetic Diseases Network recently published a white paper: *Bioinformatics Curriculum Recommendations for Undergraduate, Graduate, and Professional Programs*, which recognizes the importance of the two tracks in bioinformatics education (see Web Links).

Conclusions

The success of the HGP would not be possible without the parallel development of bioinformatics techniques. All the stages of the HGP, from cloning and sequencing to disease gene discovery, heavily depend on computer aid. Many methods and tools were developed. However, there are even more challenges awaiting. Despite tremendous effort, eukaryotic gene prediction methods are far from being perfect. It is still relatively difficult to navigate in the vast amount of information predicted by the HGP. New tools to make these results conveniently available to all (including hard wet bench) biologists are required. For the moment, the 'gene-centric' approach dominates, but to fully understand how an organism works, we need the holistic approach to understand all of the interactions between the different components. System biology is a promising direction toward this goal. Bioinformatics will continue to evolve and will continue to revolutionize biology.

See also

Bioinformatics: Technical Aspects
Genetic Databases
Genetic Databases: Mining
Microarray Bioinformatics

Further Reading

Branden C-I and Tooze J (1999) *Introduction to Protein Structure*. New York: Garland Publishing.

Mount DW (2001) *Bioinformatics: Sequence and Genome Analysis*. Plainview, NY: Cold Spring Harbor Laboratory Press.

Web Links

Bioinformatics Curriculum. This document provides recommendations for undergraduate, graduate and professional programs in bioinformatics developed by a group of Canadian scientists www.bioinformatics.ca/docs/whitepaper.pdf