

BIOINFORMATICS TOOLS FOR WHOLE GENOMES

David B. Searls

Bioinformatics Department, SmithKline Beecham Pharmaceuticals, King of Prussia, Pennsylvania 19406; e-mail: David_B_Searls@sbphrd.com

Key Words genome analysis, algorithms, databases

■ **Abstract** The advent of whole-genome data resources—not only sequence but also other genome-scale data collections such as gene expression, protein interaction, and genetic variation—is having two marked, complementary effects on the relatively new discipline of bioinformatics. First, the veritable flood of data is creating a need and demand for new tools for dealing adequately with the deluge, and, second, the unprecedented extent, diversity, and impending completeness of the data sets are creating opportunities for new approaches to discovery based on computational methods.

THE NATURE OF WHOLE-GENOME DATA SETS

At the brink of the twenty-first century, there are 24 complete genomes available in public databases (94), including 16 bacterial, 6 archaeal, and 2 eukaryotic (*Saccharomyces cerevisiae* and *Caenorhabditis elegans*) genomes. In addition there are estimated to be 82 prokaryotic and 24 eukaryotic genome-sequencing efforts under way, including, of course, that of the human genome. Already fragments of most human genes and considerable fractions of the human and many other partially completed genomic sequences are available in expressed-sequence-tag (EST) data sets. Overall, nearly 5 billion nucleotides of sequence are contained in the GenBank database (release 115.0).

It has been noted that “. . . genome sequencing risks becoming expensive molecular stamp-collecting without the tools to mine the data and fuel hypothesis-driven laboratory-based research” (123). Such tools are emerging from the nascent science of bioinformatics, an interdisciplinary field that combines molecular biology with computer science and software engineering (143). As this review illustrates, bioinformatics is crucial to transforming the torrent of raw data into biological knowledge.

To place the recent advances in bioinformatics that are described into some context, it is important to consider first the general nature and history of the data with which bioinformatics deals.

Expressed Sequence Tags

Before the 1990s, the human gene sequence data that were available were largely those derived from academic studies of individual genes and just a few extended regions. The advent of ESTs (1) was a milestone that promised that nearly the entire expressed genome would be expressed within a few years, which prompted a flurry of bioinformatics activity in both the public and private sectors (168). The dbEST database was established as a division of GenBank, which, by 1995, had overtaken the rest of the database in terms of number of records entered (21); public efforts, in turn, were dwarfed by private collections from Incyte (Palo Alto, CA) and Human Genome Sciences, Inc (Rockville, MD). dbEST has now surpassed 3 million entries, more than half human, from hundreds of different libraries (release 101599), with recent efforts being focused on comparisons of tumor libraries with normal tissues in the Cancer Gene Anatomy Project. Despite their fragmentary nature, high error rate, and other foibles, these data proved their worth even at the outset, when simple BLAST (basic local alignment search tool, National Center for Biotechnology Information, Bethesda, MD) (4) searches revealed tantalizing glimpses of novel genes in huge numbers.

The somewhat chaotic “oversampling” of genes, in fact, was the impulse for the reduced-redundancy, gene-oriented UniGene resource, consisting of EST sets that were anchored on unique 3′ sequences and further clustered by overlapping regions and origin from the same clone (138). UniGene, however, did not attempt to assemble ESTs into consistent contigs, so that sets may contain results of alternative splices or even merged paralogs, and, for that matter, there are doubtlessly unconnected sets that actually are derived from the same gene. ESTs, in fact, have proven useful in examining alternative splicing, although these investigations have also suggested the presence of many artifactual ESTs containing intronic sequences, as well as inappropriately grouped ESTs (170). EST data sets, which are “noisy” in many respects, have nevertheless proven to be very versatile resources, as discussed below. Yet their utility for novel gene identification remains their most prominent success.

Microbial Genomes

Beginning in 1995, the publication of the first complete microbial genomes ushered in another revolution in genomics and bioinformatics analysis (50). Together with the publication of the sequence data came pioneering efforts at large-scale annotation and functional protein classifications of whole genomes, generally based on straightforward homology searches that, however, were seen to leave surprisingly high percentages of open reading frames without functional assignments. The availability of the genomes for increasing numbers of diverse species has been a boon to evolutionary biologists and has also led to a cottage industry of speculation about such topics as the core set of “ancient conserved regions” [found in ~70% of microbial proteins (86)] and the minimal essential set of genes required for life

(113). This diversity has also created the opportunity to compile comprehensive classifications of homologous relationships across phylogeny, as, for example, in the Clusters of Orthologous Groups (COGs) resource (158), which has now surpassed 2000 clusters of ~27,000 proteins from 21 species. Microbial genome sequences have been of immediate use in antibiotic discovery, in which phylogenetic analyses are important in selecting protein targets that maintain sufficient similarity through bacterial clades to promise broad-spectrum antibiotics, while being well diverged from human (25). The very completeness of the genomes is also important in this regard, for example, to rule out targets for which there is a suggestion of redundancy of function in the genome.

Genomic Sequence

It is difficult to comprehend that, barely a decade ago, the beta-globin gene cluster on chromosome 11 was, at 73 kb, the most impressive stretch of contiguous human sequence available for bioinformatics analysis. In May 1999, when the fraction of the human genome available as an accurate, finished sequence first passed 10%, a 3.8-Mb sequence completely covering the region of the human major histocompatibility complex on chromosome 6 was deposited. Just 4 months later, a 14.6-Mb contig from chromosome 22 was finished (120), and indeed this is the first human chromosome to be completed, for all practical purposes. The opportunities and challenges presented by such a scale-up are self-evident.

It is likely that a significant portion of genes are not represented in dbEST, owing to their low abundance or highly specific distribution in tissues or time of expression (168). The impending completion of the human genomic sequence promises to fill this gap, as well as to “flesh out” the sequences of those genes that are only lightly touched by ESTs or for which ESTs may present a confused picture caused by sequencing errors, aberrant splicing, etc. The genomic sequence will also provide useful information on intron/exon structure, promoters and other regulatory regions, clustering of related genes, syntenic relationships with model organism genomes, and overall chromosomal organization. However, accurate detection of those genes still requires advances in bioinformatics technology, which are discussed below, and, more generally, large-scale annotation of the genomic sequence will also benefit from increased automation. In the near term as well, the nature of the sequencing process means that, for those who desire an early look, various intermediate forms of sequences will also be available, including interspersed single-pass sequences and short unordered contigs from the Genome Survey Sequence and High Throughput Genomic divisions, respectively, of GenBank.

Model Organisms

The completion of the genomes of *S. cerevisiae* in 1996 and *C. elegans* in 1998 lent new momentum to the principles of functional genomics based on comparative studies; moreover, the impending release of the *Drosophila* genome by a

consortium led by Celera Genomics (Rockville, MD) will continue this increase in momentum “with a vengeance.” Many more homologies with human and other mammalian genes have thus become evident, and through them it has also become possible to tap into additional experimental systems for the rapid elucidation of pathways, protein-protein interactions, etc. For each of the major model organisms, specialized databases and tools adapted to comparative analysis have been developed (e.g. 56, 74), which do present some significant challenges for data integration. XREFdb, for example, was created as a means to cross-reference the genetics of model organisms (beginning with yeast) to mammalian phenotypes and thus accelerate the identification of genes that are mutated in human diseases (16).

In the somewhat longer term, it can be expected that the completion of the mouse genome and other mammalian species will provide extraordinary value. The syntenic relationships within the mammalian radiations and the genetic maps available (dense in mouse and rat, moderate-resolution in many others) should contribute crucially to gene discovery, functional assessment, and evolutionary studies (121), as well as disease modeling based on phenotypes that are more likely to correlate with human phenotypes than are those of more distant taxa. In this regard, it should be noted that a mouse EST sequencing program (107), which encompasses libraries of high quality and notably some from very early stages of development, has become a particularly effective resource for comparative studies with human sequences.

Variation

Genetic maps are now well advanced and promise to become even more effective for candidate gene identification with the advent of high-density single-nucleotide polymorphism (SNP) collections for association studies. As of this writing, the National Center for Biotechnology Information (NCBI) dbSNP database (<http://www.ncbi.nlm.nih.gov/SNP/index.html>) contains ~26,000 entries, not nearly sufficient to contemplate such studies, but a consortium of private and public institutions is currently undertaking the identification of $\leq 300,000$ SNPs through redundant sequencing of reduced-representation libraries and the precise mapping of half of them (see <http://www.snp.cshl.org>). Besides the database resources and analysis for such efforts, bioinformatics is also contributing directly to SNP discovery; alignments of deeply sequenced ESTs are proving to be good sources of putative SNPs (27), as are overlaps between various genomic sequencing efforts (156).

Aside from mapping studies, so-called coding SNPs (cSNPs) are of intrinsic interest when they result in nonsynonymous changes to protein sequences. EST-derived SNPs are, of course, enriched in cSNPs, although they require careful analysis because of the error rates in ESTs. Not only do cSNPs represent a distinct minority of all SNPs, but, even among cSNPs, the nonsynonymous variety appears to be less frequent and to have lower allele frequencies as well, perhaps owing to selection (31).

Expression Data

Many technologies are now available for genome-scale analysis of patterns of gene expression, beginning with the simple expedient of counting ESTs from various libraries that contribute to different transcripts or at least clusters (e.g. 136). Gene indexes such as UniGene now routinely provide such quantitative information, and in fact the bulk of the public EST sequencing effort in recent years has been directed toward sample expression from a wide variety of tumor cell types, the so-called Cancer Gene Anatomy Project (155), which currently covers 130 libraries with ~670,000 ESTs. Other high-throughput sources of expression data include “SAGE” [serial analysis of gene expression (164)], for which the results are also now reported through UniGene, and various differential display schemes that are available on scale particularly through the private sector.

However, the spotlight in recent years has shifted to hybridization-based techniques, using “chips” or microarray gridding, which are now being used to compile expression data at an enormous rate (26). The capacity of these techniques is such that it is possible to design experiments by using tens of thousands of targets or even entire genomes, testing RNA samples from multiple tissues, under different conditions, even over extensive time series (e.g. see 79). Not only will it be possible to detect straightforward differences in expression of individual genes, but also entirely new approaches to coordinate pattern detection will need to be applied to the masses of data now arriving (36).

Going beyond expression data, efforts in proteomics can be expected to fill in a more complete picture of post-transcriptional events and of the overall protein content of cells (e.g. see 54). Structural genomics is also receiving increased attention, which may soon increase sharply the rate of accumulation of protein structures and thus the effectiveness of approaches to functional genomics based in fold recognition (30).

DATABASES AND DATA RESOURCES

Facilities for storing, updating, accessing, querying, and otherwise manipulating the very data itself lie at the heart of the genomics enterprise. Aspects of the technology that are relevant to genome data are reviewed here, along with the more important public databases and the pervasive character of the Internet and World Wide Web.

Database Technology

The history of genome databases to some extent recapitulates the evolution of database technology. The earliest data stores were “flat files” of continuous text, formatted for browsing by scrolling or scanning with linear search routines, at a time when the volume of sequence data made this feasible. A degree of

standardization of formats, together with the ubiquitous Perl scripting language, which is so powerful for managing, searching, and parsing such files, has allowed flat files to persist and even thrive in this domain. The Sequence Retrieval System (SRS) program allows flat-file databases to be efficiently indexed, queried, and hyperlinked to each other, and it is in wide use (48). There are examples also of ad hoc database systems designed specifically for the biological sequence realm, most notably ACEDB (165). However, general-purpose relational databases, with their standardization and versatile query capability, now form the backbone of most large-scale sequence databases. At the same time, there is a trend toward object-oriented databases in the field, which actually underlie ACEDB and are now being used to create new specialized databases (85), to back-fit existing databases (60), to overlay relational databases (33), etc. The advantages of object-oriented databases, in easier and richer modeling of complex domains such as molecular biology, are somewhat offset by the lack of standardization, especially in query tools. However, these concerns are being addressed, and efforts are also being made toward the use of common ontologies or comprehensive models of the domain (14, 140), which may be necessary in the future to maintain consistency not only in object models but also in relational schemas.

The profusion of databases (and even more so that of Web resources, described below) virtually begs for some form of intelligent integration. This widely discussed need has spawned a number of initiatives (99). From a curatorial perspective, most databases are now routinely being augmented with extensive cross-links to other databases, and some are being created explicitly as indices into other databases. Some such indices are centered on individual genes, such as NCBI's LocusLink (<http://www.ncbi.nlm.nih.gov/LocusLink>) or GeneCards (<http://bioinformatics.weizmann.ac.il/cards>) (129); others are centered on pathways (81, 122). Many of these consolidation efforts are severely limited by the paradigm of browsing, which does not allow for efficient, complex, global queries hinging on the relational "join" operator, much less the subtle pattern search implicit in the notion of data mining. To address these limitations, technology is being brought to bear in the form of systems such as OPM (object protocol model) (33) and Kleisli (34), which promise to provide uniform access to multiple heterogeneous distributed databases through high-level query languages. Data warehousing, an intermediate approach that brings multiple databases under a single data model, is also being advocated in this domain (e.g. 17).

Public Databases

The large number of databases, their degree of overlap, the rapid pace of change in the field, and funding issues have combined to produce a number of recent changes in the information environment (47), including the end of some databases, such as the Genome Database at Johns Hopkins University, Baltimore, MD [subsequently resurrected at the Hospital for Sick Children in Toronto, Canada (<http://www.gdb.org>)], and the commercialization of others, such as SWISS-PROT.

The most useful axis on which to classify the databases remains the fundamental distinction between those housing primary data on the one hand and highly curated compilations on the other. The former are typified by the triad of coordinated nucleotide sequence databases, including GenBank, the DNA Databank of Japan, and the European Molecular Biology Laboratory, which depend on direct submissions from individual researchers, genome sequencing projects, and other sources and which monitor submissions but do little else in the way of curation (154). A recent trend, however, has been the increasing number of useful views on the data and comprehensive scope offered by institutions such as the NCBI, which is beginning to extend also to more active curation, as in the RefSeq collection (<http://www.ncbi.nlm.nih.gov/LocusLink/refseq.html>). Nevertheless, the consistency and quality of the author-provided annotation in the databases remain serious concerns that are just beginning to be addressed through technical approaches (46). At the other extreme, SWISS-PROT, the highly curated database of protein sequences, remains one of the most valuable resources available to the bioinformatics community, with its extensive annotation as to function, domain structure, post-translational modification, variants, etc. (<http://www.expasy.ch/sprot> or <http://www.ebi.ac.uk/sprot>). Although it is continually being improved (13), perhaps the most significant recent development in SWISS-PROT is a change in its funding status, by which it is now licensed to commercial users for a fee through a corporate entity (<http://www.genebio.com>) but remains free to academics. This typifies a new trend for highly curated databases in the field, one that is necessary to support the considerable manual effort associated with each of the increasing number of entries.

Other important databases include those associated with protein families, patterns, domains, motifs, etc, many of which are described elsewhere in this article, insofar as they are useful in functional classification of novel genes. The Protein Databank (PDB) remains the dean of protein structure databases, the management of which was transferred in the past year to a consortium called the Research Collaboratory for Structural Bioinformatics (<http://www.rcsb.org>). Databases such as the Structural Classification of Proteins (SCOP; <http://www.scop.mrc-lmb.cam.ac.uk/scop>) and CATH (so called because it classifies proteins according to class, architecture, topology or fold, and homologous family; <http://www.biochem.ucl.ac.uk/bsm/cath>) not only provide one important perspective on functional classification but also offer gold standards in the “twilight zone” of similarity and nonredundant sets at various levels of granularity. Nonredundancy or reduced redundancy, in fact, has become increasingly important in many arenas of database searching, given the exploding sizes of the databases, to make searches tractable not only in the time required but also the manageability of the outputs.

Web Resources

Beyond a doubt the most significant new trend has been the proliferation of highly specialized Web resources. These usually begin as solo efforts by enterprising

researchers who may be frustrated by the public resources that lack key data or the appropriate organizational paradigm in that researcher's area of expertise. The natural history of such enterprises generally follows one of three courses. In the first instance, it may remain a dedicated resource of limited appeal, maintained by the originator and perhaps a circle of like-minded colleagues. Second, it may fall into disuse, and indeed there are many such instances of moribund Web sites that have not been updated for months or years, while "dead links" are increasingly a hazard to Web navigation. Finally, a database that strikes a chord and becomes unusually popular may, as a result, receive significant funding for maintenance and growth, for example, in its coverage, features, or timeliness. Such resources may be associated with a particular algorithm or classification system for a broad swath of data, or they may support a sizable research community, in particular those surrounding model organisms. In some cases, such databases are so successful as to be commercialized in whole or in part, as was the Yeast Proteome Database (74).

Examples of Web resources abound, so much so that it would be fruitless to attempt a comprehensive review, particularly in light of their volatility. A sampling suffices to give the flavor. One can find many resources for particular gene families, a typical case being the G-protein-coupled receptors (GPCRs), extensively cataloged in both GCRDb [<http://www.gcrdb.uthscsa.edu> (84)] and GPCRDB [<http://www.gcrdb.uthscsa.edu> (76)]. A separate database, "GRAP" (<http://www-grap.fagmed.uit.no/GRAP/homepage.html>), houses GPCR mutants (90), whereas the subset of olfactory GPCRs is treated in the olfactory receptor database [ORDB; <http://www.ycmi.med.yale.edu/senselab/ordb> (144)]. These databases exhibit typical characteristics of the genre. What appear to be competing databases will often cross-reference each other, and most link to "standard" databases for sequence data, citations, etc. The unique aspects are often ancillary data from a local laboratory, expert classifications, stylized schematic depictions, specialized search tools or algorithms, etc. Sites vary wildly as to their dependence on simple flat representations or the provision of query capabilities against more sophisticated underlying data representations, and the organizing hierarchies may be ad rem or even idiosyncratic. The MEROPS database of peptidases (<http://www.bi.bbsrc.ac.uk/Merops/merops.htm>) organizes itself into families based on similarities in the peptidase unit that is responsible for activity, thence into clans based on a common evolutionary origin of the folds, using a file card paradigm (128). The Protein Kinase Resource (<http://www.sdsc.edu/kinases>) makes use of the "Hanks Classification," which clusters based on catalytic domains that are named and further grouped by structural and functional considerations (145). It provides extensive support for search and analysis of both sequence and structure. The ImMunoGeneTics (IMGT) database of immunoglobulins, T-cell receptors, and major histocompatibility complex molecules [<http://imgt.ciues.fr:8104> (62)] offers yet another sort of subset of proteins.

Web sites tend to offer every imaginable way of “slicing and dicing” information, seemingly ever more finely. Consider just organelle resources, beginning with the comprehensive GOBASE database (<http://megasun.bch.umontreal.ca/gobase/gobase.html>) that covers both mitochondria and chloroplasts (87). Other sites are devoted to mitochondria alone, including MITOP [<http://www.mips.biochem.mpg.de/proj/medgen/mitop> (135)] and MitBASE [<http://www.ebi.ac.uk/htbin/Mitbase/mitbase.pl> (8)]. The AMmtDB database collects multiply aligned sequences of vertebrate mitochondrial genes [bio-http://www.ba.cnr.it:8000/BioWWW/#AMMTDB (95)]. Still others deal in human mitochondrial mutations and disease [MITOMAP, at <http://infinity.gen.emory.edu/mitomap.html> (83)], or more specifically in neurological diseases involving mitochondria (<http://www.neuro.wustl.edu/neuromuscular/mitosyn.html>), as well as nuclear gene products involved in mitochondrial biogenesis and function (MitoDat, at <http://www-lecb.ncicrf.gov/mitoDat>), and still other combinations of eclectic topics (MitoPick, at <http://www-dsv.cea.fr/MitoPick/Default.html>). To summarize, a researcher with an interest in mitochondria may choose among MITOP, MitBASE, AmmtDB, MITOMAP, MitoDat, MitoPick, and doubtlessly a number of others.

Resources exist for specific diseases [e.g. the Asthma Gene Database, at <http://cooke.gsf.de> (78)], specific chromosomes [e.g. the Integrated X Chromosome Database (IXDB), at <http://ixdb.mpimg-berlin-dahlem.mpg.de> (98)], and virtually every other imaginable cross-section of biology. There are compilations of information about all ongoing sequencing projects [GOLD, at <http://www.ebi.ac.uk/research/cgg/genomes.html> (94)], and there are a plethora of databases of mutations in specific genes or families, for example, p53 [<http://www.iarc.fr/p53> (73)], and androgen receptors [<http://www.mcgill.ca/androgendb> (63)], to name just a few. Regarding their quality, all of these resources are proffered virtually without warranty of any kind, other than what is apparent by direct examination or the reputation of the developers. Certainly there is a huge variety of formats and a great deal of redundancy. Such a cultural phenomenon as the Web cannot help but give rise to concerns about an informational Tower of Babel, at the same time that it inspires users with its innate power.

SEARCH AND ANALYSIS TOOLS

To address the cornucopia of data described above, the computational biologists’ toolbox continues to grow apace, with wide-ranging theoretical and practical advances in recent years. This review concentrates primarily on those recent advances that are particularly relevant to large-scale, whole-genome analysis, rather than the functional exegesis of one gene or protein at a time. A basic familiarity with standard tools and concepts is assumed, for example, alignments, profiles, phylogenetic trees, gene finding, etc. (See 22 for a high-level overview, 19 for a practical guide, and 41 for a comprehensive theoretical treatment.)

Similarity Search and Alignment

The venerable BLAST software (4), long since the sine qua non of sequence database searching because of its efficient word-based algorithm and well-founded statistics, has been enhanced for genome-scale investigations in several ways. A new search engine, BLAST 2.0, provides increased speed and sensitivity, as well as support for several new variants of the basic algorithm—Gapped BLAST and position-specific, iterated (PSI)-BLAST (5). Gapped BLAST, as the name implies, allows for insertions and deletions within alignments, easing biological interpretation over longer regions. [This capability is also provided by WU-BLAST2 (3).] PSI-BLAST combines two pre-existing techniques to provide a considerably more sensitive search. First, it melds an alignment of hits from an initial run of Gapped BLAST into a profile or position-specific scoring matrix, and then PSI-BLAST uses this profile to perform another search to draw more distant sequences into the alignment. This process is repeated until no further sequences are added to the set, resulting in a “family” of related sequences and a profile descriptor for that family.

PSI-BLAST has proven to be immensely and deservedly popular, although sophisticated users have found that great care may be required in adjusting parameters to avoid a “runaway” search that pulls in unrelated sequences profligately. Often it is necessary to perform a cluster analysis by using pairwise BLAST similarities to identify such outliers (43). Another useful technique for the validation of very distant similarities uncovered by PSI-BLAST is to use such hits to seed a reverse PSI-BLAST search, to see whether the original sequence appears in the resulting collection; empirically, this has been seen to discriminate structurally similar gene products from false positives produced by one-way searching (88). For pairwise searching itself, the modern generation of algorithms has been shown to be reliable at detecting almost all relationships among proteins (known to be related from structure) that exhibit >30% sequence identity, although only about half of those that exhibit 20%–30% identity (24). [It also appears that sequence similarity correlates very well with structural similarity across many families (171).] Used judiciously, PSI-BLAST and related profile techniques such as HMMER and SAM (see below) promise to plumb the twilight zone of similarity with unprecedented effectiveness, in one test finding threefold the remote homologs found by the pairwise methods (124).

Of increasing importance in the era of large-scale sequencing are algorithms that align various forms of nucleic acid sequences against proteins on the one hand and proteins or cDNA-derived sequences against genomic sequences on the other. Such algorithms, for instance, may be necessary to search effectively with single-pass sequence data, such as ESTs or Genome Survey Sequence (GenBank) data, in which a relatively high rate of sequencing error produces frameshifts that sharply reduce the sensitivity of a conventional six-frame translation search against protein databases. This problem has been addressed in some cases by enhancing

conventional search algorithms, as in the versatile FASTX family of programs (177) or “full-alignment” dynamic programming methods (64).¹

Searching with either cDNA or proteins against genomic sequences entails accounting for introns in addition to other forms of indels and possible frameshifts. Again, some of the new tools entail modification of existing algorithms, such as EST_GENOME, which does full alignment of spliced to unspliced DNA with a simple model of introns based on recognition of invariant dinucleotides in the splice sites (110), and sim4, which achieves similar results but at higher speeds based on its use of BLAST heuristics (52).

More elaborate and novel methods have also been developed for aligning genomic sequence with known proteins, allowing introns. The Procrustes software package preprocesses the genomic sequence to find likely exons by statistical measures and then searches for the best chain of exons to fit a given protein or cDNA, a technique termed spliced alignment (57, 109). The highly flexible WiseTool family of programs, especially GeneWise, allows for a very sophisticated intron model, profile search, and many useful variations, by virtue of a clever extension of dynamic programming principles (20). Both tools have found wide use, although the latter set tends to be especially computationally intensive in large-scale search applications.

The rapidly increasing volume of sequence data available and degree of coverage of genomes in itself lends advantages to similarity search. The increased density of genes in “sequence space” makes it more likely that a significant similarity will be detected to any novel gene, thus offering a clue to function. [However, it is remarkable that as many as half of open reading frames in recently sequenced species still cannot be assigned a function (82, 118).] A second-order salutary effect of this increased density is to render iterative search techniques more effective, by making it more likely that intermediate steps will be present (albeit perhaps unknowns) to allow the transitive detection of distant homologs of known function. In fact, one criticism of profile techniques has been that they tend to dilute the information that is present in any single sequence and may prove to be such a key evolutionary intermediate. Although PSI-BLAST may represent a happy medium in this regard, other approaches to this problem have been taken, described

¹Dynamic programming is a well-known technique in computer science whereby tabular methods are used to store intermediate results and obviate their recomputation. As a rule, dynamic programming produces guaranteed optimal solutions for certain otherwise combinatorial problems but may be too expensive in time and space for large inputs, leading to the search for ways to speed the process and heuristics that sacrifice as little as possible of the sensitivity or other advantages of the “full” dynamic programming solution. The archetypal example in bioinformatics is the classic Smith-Waterman algorithm, which performs full optimal local alignment of sequences by dynamic programming (146) but which is generally too slow (without specialized hardware) for large-scale database searching, compared with BLAST.

below. Finally, the diversity of sequence data is also highly advantageous. EST data are indicative of what is transcribed but are error prone and difficult to cluster reliably, whereas genomic sequences are accurate but present difficulties in gene identification; taken together these data sources are highly complementary, as is described below.

Pattern Discovery and Search

Profiles are one example of a style of search based on some representation of a family of sequences, rather than a single query sequence. Sequence patterns are often based on the presence in most families of “blocks” of ungapped, well-aligned regions separated by less conserved segments. The BLOCKS database, for example, was built on this theme, and it has recently been extended to contain conserved motifs from a wider variety of protein domain databases, as well as new facilities for blocks-vs-blocks searching (71). This database has also been used to generate profiles (or even simple consensus sequences) for blocks, which are then embedded within the more distinctive regions of individual sequences in such a way that pairwise searches can be performed that presumably afford the wide-ranging sensitivity of profiles without sacrificing the individual character of each family member (70). Superior performance has been claimed for various BLOCKS-based techniques (72). The PRINTS database, which is similar in concept but uses unweighted rather than weighted blocks, called fingerprints, has recently been enhanced with an on-line BLAST server and search software that are more efficient and have provided improved statistics for estimating the reliability of retrieved matches (9). The PROSITE database, which represents motifs with regular expression and, now, profiles, remains a very current and useful repository (75), and the ProDom collection of protein domains in SWISS-PROT is now built with an improved procedure based on PSI-BLAST (38). The SMART database is a highly curated and valuable set of signaling and, now, extracellular domains, which in its latest version has improved search tools, alerts, and output formats (139).

Perhaps the most widely used such collection is the Pfam database and its associated tools (18). Pfam’s protein domain families are based on Hidden Markov Model (HMM) representations, a kind of profile that makes use of a more sophisticated and versatile probabilistic model (44). Recent developments in Pfam include the provision of an advanced version of the associated search software (HMMER), which is more sensitive and provides expectation values, and more complete database coverage, such that over 54% of the proteins in SWISS-PROT and TREMBL now match one of the 1313 families in this database (release 3.1).

HMMs, indeed, have proven to be very versatile computational constructs in the bioinformatics domain, and they are representative of a marked trend in the field in recent years toward Bayesian statistical methods (41). HMMER is actually a suite of programs that builds profiles and searches with them, similar to the SAM system, which, on at least one test set of remote homologs, demonstrated performance

that was somewhat better than iterative methods such as PSI-BLAST [and, of course, is far superior to conventional pairwise search (124)]. HMMs can also be profitably used to build more elaborate models for other bioinformatics problems, as is described in the next section; profile HMMs have even been extended to deal with the detection of genomic sequences that code for folded RNA structures, such as tRNA genes (100, 134) and methylation guide small nucleolar RNAs in yeasts (101), by modeling the covariation at base-paired positions. (The latter citation is a felicitous example of the close coordination of bioinformatics analysis with bench confirmation, in what has been termed a “wet/dry cycle.”)

An as yet controversial question surrounding pattern search is the statistical basis for evaluation of search results, along the lines of the well-known BLAST expectation values. One approach to combining pattern search with more familiar routines is the Pattern-Hit Initiated BLAST (PHI-BLAST) program, which takes as input both a protein sequence and a pattern of interest that it contains (178). PHI-BLAST searches by using the input pattern and then uses the resulting hits as seeds for the construction of local alignments to the query sequence. These hits can then be sorted by score and evaluated statistically to produce a more meaningful analysis and greater sensitivity than BLAST search alone.

An increasingly useful class of algorithms is that which detects individual motifs common to a set of initially unaligned and perhaps only distantly related sequences. An example is the MEME system, which uses a statistical learning technique called expectation maximization; it has been used (together with its cognate search tool, MAST) in a large-scale test to find motifs among the family of steroid dehydrogenases, which were then shown to map well onto known structural features (12). Another approach to this problem is based on a different statistical technique called Gibbs sampling (96), which, like MEME, was first applied to protein motif detection, but has more recently been shown to be very well adapted to detection of common regulatory regions in DNA sequences (see below). Gibbs sampling combined with HMMs has also been used in iterative search and multiple alignment applications by the PROBE software package, as in the recent characterization of the AAA⁺ class of chaperonelike ATPases (119). Yet another style of solution for motif search is represented by the combinatorial TEIRESIAS algorithm (131).

Gene Finding

A classic problem in bioinformatics is that of identifying genes in novel genomic sequence data. This complex and varied topic is one of the most heavily reviewed areas of the field (e.g. see 29, 35, 111), and this article only briefly recapitulates the history while commenting on current challenges.

The approaches to this problem may be roughly divided into homology-based and *ab initio* methods and, in recent years, hybrids of the two. *Ab initio* methods, which are based on general properties and characteristics of protein-encoding genes, began some 2 decades ago with simple statistical measures of the coding

potential of exonic vs intronic and intergenic sequence, using a wide variety of word frequency and abstruse signal-processing metrics, tallied in moving windows across putative open reading frames. This approach reached its zenith with the first release of the famous GRAIL program, which combined many such lines of evidence as input to a neural net (162). The limitations of purely statistical measures of coding potential, which ignored useful biological knowledge of gene structure, were addressed by what may be termed syntactic or model-based methods that also took account of signals such as those at splice junctions, as well as constraints associated with reading frame (e.g. see 40). At the same time, accounting for intron/exon structure led to the combinatorial problem of assembling the optimal consistent gene structure from many potential exons, which was largely solved by dynamic programming approaches (58, 148).

The model-based approach to *ab initio* gene finding has culminated with the application of HMMs, which, because of their state-based architecture, are well suited to representing both cyclic transitions between exons and introns and the statistical and periodic properties within each such state, as well as boundary states and the profiles characteristic of biological signals. Moreover, HMMs have associated with them well-known dynamic programming algorithms not only for recognition of patterns but also for learning those patterns, within a well-founded Bayesian framework. These advantages have led to a proliferation of HMM-based gene finders, including Genie (130), HMM-gene (91), and GeneMark.hmm (102). Other current trends in gene finding include several efforts with improved discriminant-analysis techniques, including the FGENE family (149) and MZEF (175).

Although it is dangerous to suggest that any particular program is superior in this crowded and dynamic field, there is at least an evanescent consensus among many users that the model-based GENSCAN software currently has an overall edge (28). However, it would be a mistake to believe that, *ab initio*, the gene finding problem in novel genomic sequences is solved. The accuracy of programs such as GENSCAN, which is excellent on individual genes, drops markedly when genes are embedded in a much larger context of continuous genomic sequences (R Guigo, P Agarwal, J Abril, M Burset, J Fickett, personal communication). Even when exons in this much larger context are predicted accurately, there remains the problem of segmentation, that is, correctly calling the beginnings and ends of adjacent genes.

The best hope for genomic sequence data probably resides in the other major approach to gene finding, based on homology to known genes. Searching novel DNA sequences for coding regions similar to any known protein was the explicit goal of the classic BLASTX program (61). In fact, it can be seen that the problem of gene finding merges with that of database search, particularly as the set of known genes expands to make the detection of new ones steadily more reliable. In the same study cited above, programs such as BLASTX, GeneWise, and Procrustes were much more robust to embedding of genes in a lengthy genomic context, although, as might be expected, the accuracy dropped as models were built with more distant homologs. ESTs are also an excellent resource for delineation even

of uncharacterized genes (11, 173), and they are particularly useful at calling 3' ends.

However, as has been noted, EST collections are uneven, touching many more genes than they actually cover in terms of coding sequence and also containing artifactual untranslated sequences, whereas known proteins and full-length cDNAs still inform only a portion of novel open reading frames. Thus, *ab initio* techniques can be expected to continue to play an important role in gene finding. It has been obvious to workers in this field for a number of years that a hybrid system, which combined the best of *ab initio* and homology-based approaches, would be ideal. Indeed, a number of promising attempts have been made along these lines (e.g. 93), although it must be said that a completely effective, technically clean integration of all of the available information and technology has yet to be fielded.

Gene Expression

As has been noted, recent advances in the production of genome-scale expression data have created tremendous opportunities and challenges in the elucidation of individual gene modulation, patterns of coordinate expression among sets of genes (by which they can be mathematically clustered into putative regulons), and ultimately genetic networks. This progress has stimulated the exploration of a number of algorithms (36) and the implementation of several combined clustering and visualization tools, including CLUSTER/TREEVIEW (45), which has achieved widespread use on yeast data, and GENECLUSTER (157), which is based on a learning algorithm called a self-organizing map. For expression studies over the course of the cell cycle, Fourier analysis has also been applied to detection of periodic genes (151).

Of particular interest in this review is the use of such data together with genomic sequences in the detection of novel regulatory elements. A number of authors (see 176 for an early review) have now made use of the complete yeast genome and multiple expression studies to first cluster putatively coregulated genes and then examine their immediate upstream regions for common *cis*-regulatory elements. In a recent example (159), cell-cycle-periodic gene clusters from yeast were used for a blind and systematic upstream sequence search, with the program AlignACE, to find common motifs. AlignACE (133), GibbsDNA (176), and several other algorithms used in such studies are based on the Gibbs sampling method described previously. These motifs were then retested against all clusters, and in many cases a remarkably strong specificity was observed for the original cluster, providing good presumptive evidence for a biological role.

With the arrival of the human genomic sequence, one might hope that a similar approach would immediately bear fruit, but the less compact nature of mammalian regulatory schemata presents serious challenges. Yet help is on the way, in the form of additional mammalian genomes that can be aligned with the human genome to find noncoding regions that are highly conserved, that is, appear to be under selective pressure similar to coding regions. This technique, termed phylogenetic

footprinting, has been in use for some time² but has recently been embodied in algorithms that are effective over very long genomic sequences (66). Phylogenetic footprinting alone has been used successfully to identify putative coding regions between human and mouse genomes, narrowing the search sufficiently that expensive bench confirmation can be undertaken or simply lending confidence to computational predictions based on similarity to known transcription elements (42). Such predictions, based on searching by consensus sequence or position weight matrices (essentially profiles) derived from databases such as TRANSFAC (68), are otherwise too numerous and nonspecific when derived over extensive genomic stretches. Although progress has been made in increasing their specificity by examining aggregate patterns of putative hits (166), phylogenetic footprinting has added value to such studies and promises to be one of the most important benefits of model organism genome sequencing.

Genome Annotation

It is now commonly averred that fully automated sequence annotation is a stark necessity, given the rate of accumulation of genomic sequences, with the goal of minimizing if not eliminating manual intervention. What is usually meant by “annotation” ranges from the nearly real-time decoration of an emerging genomic data stream with gene calls and other biological features (which we term on-line annotation), to the post hoc analysis of completed genomes with both individual and aggregate analyses of open reading frames, including aggressive attempts at functional assignments, classification systems, and even versatile query and display tools. In all cases, the exigency of value-added annotation must be counterbalanced by repeated warnings about the risk of erroneous annotation being propagated through databases in destructive ways (23).

On-line annotation, which has also been termed framework annotation (11), is particularly associated with ongoing large-genome sequencing efforts in which data are released incrementally and a need is recognized for continuous gene calling, for example, for quality control or timely discovery of important new genes (92). Indeed, private sector sequencing efforts are accompanied by

²Actually, it may be argued that the underlying concept of phylogenetic footprinting goes back at least to World War II, when the eminent mathematician Abraham Wald analyzed data on patterns of bullet holes in combat aircraft returning from missions. The military noticed that certain surfaces of the planes had significantly fewer hits and others more, per unit area. They proposed to add extra armor where more hits were observed. Wald pointed out that in all likelihood the density of hits was uniform and that fewer hits were observed in some areas because the planes hit there were not returning. Thus, he argued, attention should be paid instead to places where the planes were (apparently) hit less often (105). Phylogenetic footprinting substitutes the notions of mutations for bullets and Darwinian selection for the fortunes of war.

high-throughput, computation-intensive annotation “pipelines” to afford a competitive advantage. Particular concerns about on-line annotation relate to schemes for coordinated updating of results when both the target sequence and query databases are growing and changing in character over time, and graphical user interfaces that aid in the analysis and perhaps editing of possibly unfinished sequences and even the design of confirmatory experiments.

As an example of the sort of tools required, BLAST has been adapted for large-scale analysis of genomic sequence data by the provision of a network client called PowerBLAST (174), which breaks up a lengthy query sequence into overlapping segments, searches by using Gapped BLAST, and then reassembles the results. PowerBLAST also offers several features that are common to and illustrative of a number of efforts at automated genome annotation. These include various options for masking repetitive elements and low-complexity subsequences, a perennial problem also addressed by software such as RepeatMasker (<ftp://genome.washington.edu/RM/RepeatMasker.html>; A Smit, P Green, unpublished data), XNU (37), and SEG (172). [In addition to controlling for low-entropy regions, in searching for very distant homologies it may be important to screen out transmembrane regions (150) and/or coiled coils (103).] It can also focus its search based on taxonomic information, for comparative genomics, and it offers a graphical display with annotations superimposed on sequences. Currently PowerBLAST can analyze and annotate a 100-kb query in about an hour on the NCBI BLAST server.

Dedicated on-line annotation systems also typically depend upon *ab initio* and hybrid gene-finding tools (163), as well as ancillary feature detectors such as tRNA finders (100), and these systems may provide underlying database support for projects (11). The most long standing and well proven of such systems is ACEDB (R Durbin, J Thierry-Mieg, unpublished data), originally developed to support the *C. elegans* project but now quite far flung and used in many guises (153, 165). More recent efforts include Genotator (67), GAIA (10), and MAGPIE (55), which are not so widely deployed as ACEDB, although MAGPIE has been used for several bacterial sequencing projects. The Genome Channel (112) is an ambitious consortium effort to annotate the human genomic sequence as it appears and to provide a public Web resource with the results. Closer to the source, the major genome-sequencing centers are also undertaking on-line annotation projects, for example the enSEMBL systems being created jointly by the Sanger Centre and the European Bioinformatics Institute (<http://ensembl.ebi.ac.uk>).

Examples of post hoc annotation systems include GeneQuiz (6) and MIPS (108), which deploy analytical tools comprehensively against sequences of complete genomes and provide systematic functional classifications of putative protein sequences found therein. In fact for any set of open reading frames, such tools may attempt to provide functional predictions encompassing description, function, catalytic activity, cofactors, pathway, subcellular location, quaternary structure, similarity to other proteins, active sites, and so on (51). Such large-scale annotations are now extending even into the structural world (65).

Other Tools

In addition to the categories above, other classes of tools have also made advances recently, which are not reviewed extensively here. Multiple alignment programs are now available in a variety of forms (e.g. progressive, iterative, local vs global, etc) with various advantages and disadvantages (reviewed in 15 and 161); their use is somewhat a matter of taste, particularly as regards their graphical user interfaces and connectivity with other utilities. Similarly, phylogenetic reconstruction remains an active research topic (reviewed in 117), with a variety of approaches based on parsimony, likelihood, distance, etc, and no small degree of controversy but with a few widely used practical packages such as PHYLIP (<http://www.evolution.genetics.washington.edu/phylip.html>) (49) and PAUP (<http://www.lms.si.edu/PAUP>). What is uncontroversial is the vastly increasing importance of phylogenetic analysis, together with multiple alignment, in the interpretation of function in an evolutionary context, as more and more genomes become available.

INTERFACES AND VISUALIZATION TOOLS

One other consequence of the volume and variety of whole-genome data is the necessity for tools that provide for a visual appreciation of the data. Sequence data have long since become unmanageable as text, and visual approaches to bioinformatics have developed along three lines: (a) graphical user interfaces to sequence management and analysis packages, which provide convenient desktop metaphors for viewing data and interacting with it; (b) scientific visualization techniques that seek to portray the gray masses of data with an imaginative use of form, color, dimension, etc, so as to best take advantage of the human cognitive apparatus in picking out features and patterns; and (c) visual programming, the pictorial specification of algorithms in specialized, high-level, and perhaps domain-specific computer languages. Use has been made of all three approaches in bioinformatics (142), although only the first two are reviewed briefly here.

Graphical User Interfaces

Literally dozens of comprehensive graphical interfaces have been created for genomic data, few of which have found any significant following. One that has is the ubiquitous ACEDB, which acquired much of its original clientele by virtue of an impressive, biologically intuitive visual interface, and this is now being enhanced with a Java interface called JADE (<http://www.stein.cshl.org/jade>) (152). In general, efforts to create comprehensive desktop toolkits with sophisticated user interfaces have passed over to the private sector, to be embodied in products by companies such as DNASTAR (Madison, WI; <http://www.dnastar.com>), NetGenics (Cleveland, OH; <http://www.netgenics.com>), InforMax (North Bethesda, MD; <http://www.informaxinc.com>), DoubleTwist (Oakland, CA; <http://www.doubletwist.com>), and Genomica (Boulder, CO; <http://www.genomica.com>), among an

increasing number. To be sure, a number of the systems that have already been described for genome annotation, etc, have significant investments in graphical interfaces, but, partly in reaction to the proliferation of heavily integrated monolithic systems, there has also been a movement toward “lightweight” interface components in this domain (141). This philosophy stresses the development of highly reconfigurable and reusable software modules (sometimes called “widgets”) that can be assembled into “plug-and-play” systems in building new interfaces readily, an example being the bioWidgets set (<http://www.cbil.upenn.edu/bioWidgets>). This and other systems have been implemented in Java (69), have found their way into products (<http://www.neomorphic.com>), and, more important, have contributed to a movement towards industry standards of interoperability, working through the Life Sciences Research Group of the CORBA-oriented Object Management Group (<http://www.lsr.ebi.ac.uk>). Examples of reusable components would be multiple alignment viewers such as CINEMA (125) and any of a number of similar widgets in the above-cited tool sets.

Scientific Visualization

Many of the user interfaces described above make heavy use of visualization principles in depicting massive data sets. Molecular biology actually has a rich tradition of inventive visualization techniques, for example in the use of dot plots for sequence comparison (as in phylogenetic footprinting), various graphical schemes for depicting folded RNA structures, helical wheels in demonstrating amphipathic helices, and many other examples from structural biology (142). Some representations have become so prevalent as to achieve a status as new icons of molecular biology, such as the “sequence logo” representation of the information content in consensus sequences, which continue to be refined (137) and applied in new contexts. Many of the newer tools are closely linked with software components (132) and with state-of-the-art methods for navigation and data transformation, with intriguing names like hyperbolic tree viewers, semantic zooming, and magic lenses (127). Often they are targeted to specific challenges of the domain such as depiction of pathways (81), maps (89), and especially expression data by using viewers described above as well as commercial visualization tools that are being adapted to the purpose, for example Spotfire (2). Just as highly automated annotation “pipelines” are required to deal computationally with the unprecedented volume of sequence data, so are well-thought-out visualization systems needed to deal cognitively with it.

INSIGHTS FROM ANALYSIS OF WHOLE GENOMES

Beyond the platform biotechnologies that actually produce the data, bioinformatics is truly the enabler of genome-scale biology. This review has not even touched on its crucial role in the acquisition of the raw data, in laboratory information management systems, for example, or the very sophisticated physical mapping and contig assembly algorithms that are primarily the concern of specialized centers.

Nor has it covered the infrastructural issues of data management and support for high-performance computing that are proving challenging to many institutional information-technology organizations in this arena. Having concentrated instead on the tools by which scientific value can be derived from whole-genome data, this review concludes by examining some cases in which such data offer unique opportunities for imaginative computational analysis.

The advantages afforded by large-scale genomics lie in the numbers of genes, the numbers of genomes, and the completeness of the genomes, plus the opportunity to integrate multiple additional data sources that deal with variation, expression, structure, etc. Many examples have been given already of the benefits of filling in sequence space with more genes, as well as an extended account of how, with powerful effect, expression data can be integrated with phylogenetic footprinting of diverse genomes and algorithms for motif alignment and regulatory-region analysis. The completeness of genomes allows for an extended logic following from a closed universe, for example permitting a confident analysis of variation and evolution of the citric acid cycle in 19 complete genomes (77). This study was able to make generalizations about the presence and absence of the cycle and its various shunts, branches, inputs, and outputs and even the role of incomplete cycles in various anabolic processes, based on essentially complete knowledge of the gene content of the various organisms. Even where mysteries remain, there is a strong basis to seek out possibly overlooked open reading frames or to reason about possibilities of homologous and nonhomologous gene displacement (different genes coding for proteins that perform the same function), where, for incomplete genomes, such speculation would be hopelessly overextended.

By itself, the sheer number of genomes becoming available opens up new vistas. Not only can clusters of orthologous genes be examined together (158), but it is increasingly possible to characterize proteins simply by their presence or absence through many taxa, producing a vector called a phylogenetic profile that can itself be predictive of function (126). Phylogenetic reconstruction can also be based quantitatively on the commonality of total gene content, rather than comparisons of single genes, which can be obfuscated by horizontal gene transfer, unrecognized paralogy (i.e. genes in a single species sharing a common ancestor), and highly variable rates of evolution (147).

Examining the arrangement of genes and other elements within genomes adds yet another dimension to functional analysis. Particularly in microbial genomes, it is possible to detect clustering of coordinately expressed genes in operons [although the significance of this is a subject of debate (97)], and, in higher organisms, patterns of segment conservation offer helpful clues to gene identities and, more important, genome evolution (116). Evolutionary insights also follow from the ability to examine complete fossil records of genome duplications [e.g. the finding of roughly equal rates of functional divergence and gene loss after duplication (115)] and horizontal gene transfer [e.g. the emerging picture of frequent transfers of operational or housekeeping genes, in contrast to informational genes related to transcription, translation, etc (80)]. Even some aspects of traditional genetics and genetic disorders are now seen as best understood in the context of whole-genome

architectures, for example, when such duplications provide opportunities for DNA rearrangements that lead to disease (104).

The association of genes in genomes can be extended even to the association of protein domains. It has long been observed that proteins that interact or participate in a common pathway may be distinct in one organism but are often found fused in some other organism. A comprehensive search of many genomes found a surprising number of such protein pairs (6809 in *Escherichia coli* and 45,502 in yeast, before further computational filtering), many of which were confirmed as functionally related (106), opening up a new approach to functional prediction based on the availability of many genomes. Increasingly, it is being recognized that genomic approaches significantly complement the traditional methods of formal genetics, biochemistry, and cell biology in the elucidation of physiological and even developmental pathways (114).

Genomic sequencing has been a boon to structural biology, even in advance of concerted efforts to accumulate three-dimensional structures on a similar scale. The completeness of genomes allows, for the first time, precise assessments of the degree of coverage by and in-built biases of existing collections of structures (59), as well as interesting aggregate analyses of the distributions of fold families (53, 169). The availability of complete genomes has also stimulated the development of processes for their systematic structural annotation (reviewed in 160), which have particularly benefited from the PSI-BLAST algorithm. Indeed, protocols that use PSI-BLAST search and back-validation combined with secondary structural and other analyses are beginning to rival threading algorithms in fold recognition tasks and promise to help fill out the protein universe almost on pace with genomic sequencing (88).

Thus, striking advantages arise from genomics and its midwife, bioinformatics, in both quantitative and qualitative ways. When we can contemplate further integrations with the other emerging “omics”—proteomics, physiomics, and so on—the challenges and potential benefits can only inspire awe.

ACKNOWLEDGMENTS

The author is heavily indebted to Nora Odendahl, James Fickett, Randall Smith, Pankaj Agarwal, Jason Miller, and other colleagues in the SmithKline Beecham Bioinformatics Department, for their expert guidance and comments.

Visit the Annual Reviews home page at www.AnnualReviews.org

LITERATURE CITED

1. Adams MD, Kerlavage AR, Fields C, Venter JC. 1993. 3,400 new expressed sequence tags identify diversity of transcripts in human brain. *Nat. Genet.* 4(3):256–67
2. Ahlberg C. 1999. Visual exploration of HTS databases: bridging the gap between chemistry and biology. *Drug Discov. Today* 4(8):370–76

3. Altschul SF, Gish W. 1996. Local alignment statistics. *Methods Enzymol.* 266:460–80
4. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J. Mol. Biol.* 215(3):403–10
5. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, et al. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25(17):3389–402
6. Andrade MA, Brown NP, Leroy C, Hoersch S, de Daruvar A, et al. 1999. Automated genome sequence analysis and annotation. *Bioinformatics* 15(5):391–412
7. Deleted in proof
8. Attimonelli M, Altamura N, Benne R, Boyen C, Brennicke A, et al. 1999. MitBASE: a comprehensive and integrated mitochondrial DNA database. *Nucleic Acids Res.* 27(1):128–33
9. Attwood TK, Flower DR, Lewis AP, Mabey JE, Morgan SR, et al. 1999. PRINTS prepares for the new millennium. *Nucleic Acids Res.* 27(1):220–25
10. Bailey LC Jr, Fischer S, Schug J, Crabtree J, Gibson M, Overton GC. 1998. GAIA: framework annotation of genomic sequence. *Genome Res.* 8(3):234–50
11. Bailey LC Jr, Searls DB, Overton GC. 1998. Analysis of EST-driven gene annotation in human genomic sequence. *Genome Res.* 8(4):362–76
12. Bailey TL, Baker ME, Elkan CP. 1997. An artificial intelligence approach to motif discovery in protein sequences: application to steroid dehydrogenases. *J. Steroid Biochem. Mol. Biol.* 62(1):29–44
13. Bairoch A, Apweiler R. 1999. The SWISS-PROT protein sequence data bank and its supplement TrEMBL in 1999. *Nucleic Acids Res.* 27(1):49–54
14. Baker PG, Goble CA, Bechhofer S, Paton NW, Stevens R, Brass A. 1999. An ontology for bioinformatics applications. *Bioinformatics* 15(6):510–20
15. Barton GJ. 1998. Protein sequence alignment techniques. *Acta Crystallogr. D. Biol. Crystallogr.* 54(1):1139–46
16. Bassett DE Jr, Boguski MS, Spencer F, Reeves R, Goebel M, Hieter P. 1995. Comparative genomics, genome cross-referencing and XREFdb. *Trends Genet.* 11(9):372–73
17. Bassett DE Jr, Eisen MB, Boguski MS. 1999. Gene expression informatics—it's all in your mine. *Nat. Genet.* 21(1):51–55 (Suppl.)
18. Bateman A, Birney E, Durbin R, Eddy SR, Finn RD, Sonnhammer EL. 1999. Pfam 3.1: 1313 multiple alignments and profile HMMs match the majority of proteins. *Nucleic Acids Res.* 27(1):260–62
19. Baxevanis A, Ouellette BFF, eds. 1998. *Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins.* New York: Wiley. 370 pp.
20. Birney E, Thompson JD, Gibson TJ. 1996. PairWise and SearchWise: finding the optimal alignment in a simultaneous comparison of a protein profile against all DNA translation frames. *Nucleic Acids Res.* 24(14):2730–39
21. Boguski MS. 1995. The turning point in genome research. *Trends Biochem. Sci.* 20(8):295–96
22. Brenner S, Lewitter F, eds. 1998. *Trends Guide to Bioinformatics.* London: Elsevier
23. Brenner SE. 1999. Errors in genome annotation. *Trends Genet.* 15(4):132–33
24. Brenner SE, Chothia C, Hubbard TJP. 1998. Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships. *Proc. Natl. Acad. Sci. USA* 95(11):6073–78
25. Brown JR, Warren PV. 1998. Antibiotic discovery: Is it all in the genes? *Drug Discov. Today* 3(12):564–66
26. Brown PO, Botstein D. 1999. Exploring the new world of the genome with

- DNA microarrays. *Nat. Genet.* 21(1):33–37 (Suppl.)
27. Buetow KH, Edmonson MN, Cassidy AB. 1999. Reliable identification of large numbers of candidate SNPs from public EST data. *Nat. Genet.* 21(3):323–25
 28. Burge C, Karlin S. 1997. Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* 268(1):78–94
 29. Burge CB, Karlin S. 1998. Finding the genes in genomic DNA. *Curr. Opin. Struct. Biol.* 8(3):346–54
 30. Burley SK, Almo SC, Bonanno JB, Capel M, Chance MR, et al. 1999. Structural genomics: beyond the human genome project. *Nat. Genet.* 23(2):151–57
 31. Cargill M, Altshuler D, Ireland J, Sklar P, Ardlie K, et al. 1999. Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nat. Genet.* 22(3):231–38
 32. Deleted in proof
 33. Chen IM, Kosky AS, Markowitz VM, Szeto E, Topaloglou T. 1998. Advanced query mechanisms for biological databases. *Intell. Syst. Mol. Biol.* 6:43–51
 34. Chung SY, Wong L. 1999. Kleisli: a new tool for data integration in biology. *Trends Biotechnol.* 17(9):351–55
 35. Claverie JM. 1997. Computational methods for the identification of genes in vertebrate genomic sequences. *Hum. Mol. Genet.* 6(10):1735–44
 36. Claverie JM. 1999. Computational methods for the identification of differential and coordinated gene expression. *Hum. Mol. Genet.* 8(10):1821–32
 37. Claverie JM, States DJ. 1993. Information enhancement methods for large scale sequence analysis. *Comput. Chem.* 17:191–201
 38. Corpet F, Gouzy J, Kahn D. 1999. Recent improvements of the ProDom database of protein domain families. *Nucleic Acids Res.* 27(1):263–67
 39. Deleted in proof
 40. Dong S, Searls DB. 1994. Gene structure prediction by linguistic methods. *Genomics* 23:540–51
 41. Durbin R, Eddy R, Krogh A, Mitchison G. 1998. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. New York: Cambridge Univ. Press
 42. Duret L, Bucher P. 1997. Searching for regulatory elements in human noncoding sequences. *Curr. Opin. Struct. Biol.* 7(3):399–406
 43. Eddy S. 1998. Multiple-alignment and -sequence searches. See Ref. 22, pp. 15–17
 44. Eddy SR. 1998. Profile hidden Markov models. *Bioinformatics* 14(9):755–63
 45. Eisen MB, Spellman PT, Brown PO, Botstein D. 1998. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA* 95(25):14863–68
 46. Eisenhaber F, Bork P. 1999. Evaluation of human-readable annotation in biomolecular sequence databases with biological rule libraries. *Bioinformatics* 15(7):528–35
 47. Ellis LB, Kalumbi D. 1999. Financing a future for public biological data. *Bioinformatics* 15(9):717–22
 48. Etzold T, Ulyanov A, Argos P. 1966. SRS: information retrieval system for molecular biology data banks. *Methods Enzymol.* 266:114–28
 49. Felsenstein J. 1996. Inferring phylogenies from protein sequences by parsimony, distance, and likelihood methods. *Methods Enzymol.* 266:418–27
 50. Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, et al. 1995. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* 269:496–512
 51. Fleischmann W, Moller S, Gateau A, Apweiler R. 1999. A novel method for automatic functional annotation of proteins. *Bioinformatics* 15(3):228–33
 52. Florea L, Hartzell G, Zhang Z, Rubin GM, Miller W. 1998. A computer program for

- aligning a cDNA sequence with a genomic DNA sequence. *Genome Res.* 8(9):967–74
53. Frishman D, Mewes HW. 1999. Genome-based structural biology. *Prog. Biophys. Mol. Biol.* 72(1):1–17
 54. Futcher B, Latter GI, Monardo P, McLaughlin CS, Garrels JI. 1999. A sampling of the yeast proteome. *Mol. Cell Biol.* 19(11):7357–68
 55. Gaasterland T, Sensen CW. 1996. MAGPIE: automated genome interpretation. *Trends Genet.* 12(2):76–78
 56. Gelbart WM, Crosby M, Matthews B, Rindone WP, Chillemi J, et al. 1997. FlyBase: a Drosophila database: the FlyBase consortium. *Nucleic Acids Res.* 25(1):63–66
 57. Gelfand MS, Mironov AA, Pevzner PA. 1996. Gene recognition via spliced sequence alignment. *Proc. Natl. Acad. Sci. USA* 93(17):9061–66
 58. Gelfand MS, Roytberg MA. 1993. A dynamic programming approach for predicting the exon-intron structure. *Biosystems* 30:173–82
 59. Gerstein M. 1998. How representative are the known structures of the proteins in a complete genome? A comprehensive structural census. *Fold Des.* 3(6):497–512
 60. Ghosh D. 1999. Object oriented Transcription Factors Database (ooTFD). *Nucleic Acids Res.* 27(1):315–17
 61. Gish W, States DJ. 1993. Identification of protein coding regions by database similarity search. *Nat. Genet.* 3(3):266–72
 62. Lefranc M-P, Giudicelli V, Ginestous C, Bodmer J, Müller W, et al. 1999. IMGT, the international ImmunoGeneTics database. *Nucleic Acids Res.* 27(1):209–12
 63. Gottlieb B, Beitel LK, Lumbroso R, Pinsky L, Trifiro M. 1999. Update of the androgen receptor gene mutations database. *Hum. Mutat.* 14(2):103–14
 64. Guan X, Uberbacher EC. 1996. Alignments of DNA and protein sequences containing frameshift errors. *Comput. Appl. Biosci.* 12(1):31–40
 65. Guex N, Diemand A, Peitsch MC. 1999. Protein modelling for all. *Trends Biochem. Sci.* 24(9):364–67
 66. Hardison RC, Oeltjen J, Miller W. 1997. Long human-mouse sequence alignments reveal novel regulatory elements: a reason to sequence the mouse genome. *Genome Res.* 7(10):959–66
 67. Harris NL. 1997. Genotator: a workbench for sequence annotation. *Genome Res.* 7(7):754–62
 68. Heinemeyer T, Chen X, Karas H, Kel AE, Kel OV, et al. 1999. Expanding the TRANSFAC database towards an expert system of regulatory molecular mechanisms. *Nucleic Acids Res.* 27(1):318–22
 69. Helt GA, Lewis S, Loraine AE, Rubin GM. 1998. BioViews: Java-based tools for genomic data visualization. *Genome Res.* 8(3):291–305
 70. Henikoff S, Henikoff JG. 1997. Embedding strategies for effective use of information from multiple sequence alignments. *Protein Sci.* 6(3):698–705
 71. Henikoff S, Henikoff JG, Pietrokovski S. 1999. Blocks⁺: a non-redundant database of protein alignment blocks derived from multiple compilations. *Bioinformatics* 15(6):471–79
 72. Henikoff S, Pietrokovski S, Henikoff JG. 1998. Superior performance in protein homology detection with the Blocks Database servers. *Nucleic Acids Res.* 26(1):309–12
 73. Hernandez-Boussard T, Rodriguez-Tome P, Montesano R, Hainaut P. 1999. IARC p⁵³ mutation database: a relational database to compile and analyze p⁵³ mutations in human tumors and cell lines. *Hum. Mutat.* 14(1):1–8
 74. Hodges PE, McKee AH, Davis BP, Payne WE, Garrels JI. 1999. The Yeast Proteome Database (YPD): a model for the organization and presentation of genome-wide functional data. *Nucleic Acids Res.* 27:69–73

75. Hofmann K, Bucher P, Falquet L, Bairoch A. 1999. The PROSITE database, its status in 1999. *Nucleic Acids Res.* 27(1):215–19
76. Horn F, Weare J, Beukers MW, Hörsch S, Bairoch A, et al. 1998. GPCRDB: an information system for G protein-coupled receptors. *Nucleic Acids Res.* 26(1):277–81
77. Huynen MA, Dandekar T, Bork P. 1999. Variation and evolution of the citric-acid cycle: a genomic perspective. *Trends Microbiol.* 7(7):281–91
78. Immervoll T, Wjst M. 1999. Current status of the Asthma and Allergy Database. *Nucleic Acids Res.* 27(1):213–14
79. Iyer VR, Eisen MB, Ross DT, Schuler G, Moore T, et al. 1999. The transcriptional program in the response of human fibroblasts to serum. *Science* 283:83–87
80. Jain R, Rivera MC, Lake JA. 1999. Horizontal gene transfer among genomes: the complexity hypothesis. *Proc. Natl. Acad. Sci. USA* 96 (7):3801–6
81. Karp PD, Krummenacker M, Paley S, Wagg J. 1999. Integrated pathway-genome databases and their role in drug discovery. *Trends Biotechnol.* 17(7):275–81
82. Kawarabayasi Y, Hino Y, Horikawa H, Yamazaki S, Haikawa Y, et al. 1999. Complete genome sequence of an aerobic hyper-thermophilic crenarchaeon, *Aeropyrum pernix* K1. *DNA Res.* 6(2):83–101, 145–52
83. Kogelnik, AM, Lott MT, Brown MD, Navathe SB, Wallace DC. 1998. MIT-OMAP: a human mitochondrial genome database—1998 update. *Nucleic Acids Research Res.* 26(1):112–15
84. Kolakowski LF Jr. 1994. GCRDb: a G-protein-coupled receptor database. *Recept. Channels* 2(1):1–7
85. Kolpakov FA, Ananko EA, Kolesov GB, Kolchanov NA. 1998. GeneNet: a gene network database and its automated visualization. *Bioinformatics* 14(6):529–37
86. Koonin EV, Tatusov RL, Galperin MY. 1998. Beyond complete genomes: from sequence to structure and function. *Curr. Opin. Struct. Biol.* 8(3):355–63
87. Korab-Laskowska M, Rioux P, Brossard N, Littlejohn TG, Gray MW, et al. 1998. The Organelle Genome Database Project (GOBASE). *Nucleic Acids Res.* 26(1):138–44
88. Koretke KK, Russell RB, Copley RR, Lupas AN. 1999. Fold recognition using sequence and secondary structure information. *Proteins Struct. Funct. Genet.* 3(Suppl.):141–48
89. Kraemer ET, Ferrin TE. 1998. Molecules to maps: tools for visualization and interaction in support of computational biology. *Bioinformatics* 14(9):764–71
90. Kristiansen K, Dahl SG, Edvardsen Ø. 1996. A database of mutants and effects of site-directed mutagenesis experiments on G-protein coupled receptors. *Proteins: Struct. Funct. Genet.* 26:81–94
91. Krogh A. 1997. Two methods for improving performance of an HMM and their application for gene finding. *Intell. Syst. Mol. Biol.* 5:179–86
92. Kuehl PM, Weisemann JM, Touchman JW, Green ED, Boguski MS. 1999. An effective approach for analyzing “prefinished” genomic sequence data. *Genome Res.* 9(2):189–94
93. Kulp D, Haussler D, Reese MG, Eeckman FH. 1997. Integrating database homology in a probabilistic gene structure model. *Pac. Symp. Biocomput.* 2:232–44
94. Kyrpides NC. 1999. Genomes OnLine Database (GOLD 1.0): a monitor of complete and ongoing genome projects worldwide. *Bioinformatics* 15(9):773–74
95. Lanave C, Attimonelli M, De Robertis M, Licciulli F, Liuni S, et al. 1999. Update of AMmtDB: a database of multi-aligned metazoa mitochondrial DNA sequences. *Nucleic Acids Res.* 27(1):134–37
96. Lawrence CE, Altschul SF, Boguski MS, Liu JS, Neuwald AF, Wooten JC. 1993. Detecting subtle sequence signals: a Gibbs

- sampling strategy for multiple alignment. *Science* 262:208–14
97. Lawrence JG. 1997. Selfish operons and speciation by gene transfer. *Trends Microbiol.* 5(9):355–59
98. Leser U, Roest Crollius H, Lehrach H, Sudbrak R. 1999. IXDB, an X chromosome integrated database (update). *Nucleic Acids Res.* 27(1):123–27
99. Letovsky SI, ed. 1999. *Bioinformatics: Databases and Systems*. Amsterdam: Kluwer. 304 pp.
100. Lowe TM, Eddy SR. 1997. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* 25(5):955–64
101. Lowe TM, Eddy SR. 1999. A computational screen for methylation guide snoRNAs in yeast. *Science* 283:1168–71
102. Lukashin AV, Borodovsky M. 1998. GeneMark.hmm: new solutions for gene finding. *Nucleic Acids Res.* 26(4):1107–15
103. Lupas A. 1997. Predicting coiled-coil regions in proteins. *Curr. Opin. Struct. Biol.* 7(3):388–93
104. Lupski JR. 1998. Genomic disorders: structural features of the genome can lead to DNA rearrangements and human disease traits. *Trends Genet.* 14(10):417–22
105. Mangel M, Samaniego FJ. 1984. Abraham Wald's work on aircraft survivability. *J. Am. Statist. Assoc.* 79:259–70
106. Marcotte EM, Pellegrini M, Ng H-L, Rice DW, Yeates TO, Eisenberg D. 1999. Detecting protein function and protein-protein interactions from genome sequences. *Science* 285:751–53
107. Marra M, Hillier L, Kucaba T, Allen M, Barstead R, et al. 1999. An encyclopedia of mouse genes. *Nat. Genet.* 21(2):191–94
108. Mewes HW, Heumann K, Kaps A, Mayer K, Pfeiffer F, et al. 1999. MIPS: a database for genomes and protein sequences. *Nucleic Acids Res.* 27(1):44–48
109. Mironov AA, Roytberg MA, Pevzner PA, Gelfand MS. 1998. Performance-guarantee gene predictions via spliced alignment. *Genomics* 51(3):332–39
110. Mott R. 1997. EST_GENOME: a program to align spliced DNA sequences to unspliced genomic DNA. *Comput. Appl. Biosci.* 13(4):477–78
111. Mural RJ. 1999. Current status of computational gene finding: a perspective. *Methods Enzymol.* 303:77–83
112. Mural RJ, Parang M, Shah M, Snoddy J, Uberbacher EC. 1999. The Genome Channel: a browser to a uniform first-pass annotation of genomic DNA. *Trends Genet.* 15(1):38–39
113. Mushegian AR, Koonin EV. 1996. A minimal gene set for cellular life derived by comparison of complete bacterial genomes. *Proc. Natl. Acad. Sci. USA* 93 (19):10268–73
114. Nadeau JH, Dunn PJ. 1998. Genomic strategies for defining and dissecting developmental and physiological pathways. *Curr. Opin. Genet. Dev.* 8(3):311–15
115. Nadeau JH, Sankoff D. 1997. Comparable rates of gene loss and functional divergence after genome duplications early in vertebrate evolution. *Genetics* 147(3):1259–66
116. Nadeau JH, Sankoff D. 1998. Counting on comparative maps. *Trends Genet.* 14(12):495–501
117. Nei M. 1996. Phylogenetic analysis in molecular evolutionary genetics. *Annu. Rev. Genet.* 30:371–403
118. Nelson KE, Clayton RA, Gill SR, Gwinn ML, Dodson RJ, et al. 1999. Evidence for lateral gene transfer between Archaea and bacteria from genome sequence of *Thermotoga maritima*. *Nature* 399:323–29
119. Neuwald AF, Aravind L, Spouge JL, Koonin EV. 1999. AAA⁺: a class of chaperone-like ATPases associated with the assembly, operation, and disassembly of protein complexes. *Genome Res.* 9(1):27–43

120. Normile D, Pennisi E. 1999. Team wrapping up sequence of first human chromosome. *Science* 283:2038–39
121. O'Brien SJ, Menotti-Raymond M, Murphy WJ, Nash WG, Wienberg J, et al. 1999. The promise of comparative genomics in mammals. *Science* 286:458–81
122. Ogata H, Goto S, Sato K, Fujibuchi W, Bono H, Kanehisa M. 1999. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* 27(1):29–34
123. Pallen MJ. 1999. Microbial genomes. *Mol. Microbiol.* 32(5):907–12
124. Park J, Karplus K, Barrett C, Hughey R, Haussler D, et al. 1998. Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods. *J. Mol. Biol.* 284(4):1201–10
125. Parry-Smith DJ, Payne AW, Michie AD, Attwood TK. 1998. CINEMA—a novel colour interactive editor for multiple alignments. *Gene* 221(1):GC57–GC63
126. Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO. 1999. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc. Natl. Acad. Sci. USA* 96(8):4285–88
127. Pook S, Vaysseix G, Barillot E. 1998. Zomit: biological data visualization and browsing. *Bioinformatics* 14(9):807–14
128. Rawlings ND, Barrett AJ. 1999. MEROPS: the peptidase database. *Nucleic Acids Res.* 27(1):325–31
129. Rebhan M, Chalifa-Caspi V, Prilusky J, Lancet D. 1998. GeneCards: a novel functional genomics compendium with automated data mining and query reformulation support. *Bioinformatics* 14(8):656–64
130. Reese MG, Eeckman FH, Kulp D, Haussler D. 1997. Improved splice site detection in Genie. *J. Comput. Biol.* 4(3):311–23
131. Rigoutsos I, Floratos A. 1998. Combinatorial pattern discovery in biological sequences: the TEIRESIAS algorithm. *Bioinformatics* 14(1):55–67
132. Robinson AJ, Flores TP. 1997. Novel techniques for visualising biological information. *Intell. Syst. Mol. Biol.* 5:241–49
133. Roth FP, Hughes JD, Estep PW, Church GM. 1998. Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. *Nat. Biotechnol.* 16(10):939–45
134. Sakakibara Y, Brown M, Hughey R, Mian IS, Sjolander K, et al. 1994. Stochastic context-free grammars for tRNA modeling. *Nucleic Acids Res.* 22(23):5112–20
135. Scharfe C, Zaccaria P, Hoernagel K, Jaksch M, Klopstock T, et al. 1999. MITOP: database for mitochondria-related proteins, genes and diseases. *Nucleic Acids Res.* 27(1):153–55
136. Schmitt AO, Specht T, Beckmann G, Dahl E, Pilarsky CP, et al. 1999. Exhaustive mining of EST libraries for genes differentially expressed in normal and tumour tissues. *Nucleic Acids Res.* 27(21):4251–60
137. Schneider TD. 1997. Information content of individual genetic sequences. *J. Theor. Biol.* 189(4):427–41
138. Schuler GD. 1997. Pieces of the puzzle: expressed sequence tags and the catalog of human genes. *J. Mol. Med.* 75(10):694–98
139. Schultz J, Milpetz F, Bork P, Ponting CP. 1998. SMART, a simple modular architecture research tool: identification of signaling domains. *Proc. Natl. Acad. Sci. USA* 95(11):5857–64
140. Schulze-Kremer S. 1998. Ontologies for molecular biology. *Pac. Symp. Biocomput.* 3:695–706
141. Searls DB. 1995. bioTk: componentry for genome informatics graphical user interfaces. *Gene* 163(2):GC1–GC16
142. Searls DB. 1997. Visualizing the genome. In *Theoretical and Computational*

- Methods in Genome Research*, ed. S. Suhai, pp.185–204. New York: Plenum
143. Searls DB. 1998. Grand challenges in computational biology. In *Computational Methods in Molecular Biology*, pp. 3–10. London: Elsevier
 144. Skoufos E, Healy MD, Singer MS, Nadkarni PM, Miller PL, Shepherd GS. 1999. Olfactory Receptor Database: a database of the largest eukaryotic gene family. *Nucleic Acids Res.* 1:343–45
 145. Smith CM, Shindyalov IN, Veretnik S, Gribskov M, Taylor SS, et al. 1997. The protein kinase resource. *Trends Biochem. Sci.* 22(11):444–46
 146. Smith TF, Waterman MS. 1981. Identification of common molecular sequences. *J. Mol. Biol.* 147(1):195–97
 147. Snel B, Bork P, Huynen MA. 1999. Genome phylogeny based on gene content. *Nat. Genet.* 21(1):108–10
 148. Snyder EE, Stormo GD. 1993. Identification of coding regions in genomic DNA sequences: an application of dynamic programming and neural networks. *Nucleic Acids Res.* 21:607–13
 149. Solovyev V, Salamov A. 1997. The GeneFinder computer tools for analysis of human and model organisms genome sequences. *Intell. Syst. Mol. Biol.* 5:294–302
 150. Sonnhammer EL, von Heijne G, Krogh A. 1998. A hidden Markov model for predicting transmembrane helices in protein sequences. *Intell. Syst. Mol. Biol.* 6:175–82
 151. Spellman PT, Sherlock G, Zhang MQ, Iyer VR, Anders K, et al. 1998. Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol. Biol. Cell.* 9(12):3273–97
 152. Stein LD, Cartinhour S, Thierry-Mieg D, Thierry-Mieg J. 1998. JADE: an approach for interconnecting bioinformatics databases. *Gene* 209(1–2):39–43
 153. Stein LD, Thierry-Mieg J. 1998. Scriptable access to the *Caenorhabditis elegans* genome sequence and other ACEDB databases. *Genome Res.* 8(12):1308–15
 154. Stoesser G, Tuli MA, Lopez R, Sterk P. 1999. The EMBL Nucleotide Sequence Database. *Nucleic Acids Res.* 27(1):18–24
 155. Strausberg RL, Dahl CA, Klausner RD. 1997. New opportunities for uncovering the molecular basis of cancer. *Nat. Genet.* 15(Suppl.):415–16
 156. Taillon-Miller P, Gu Z, Li Q, Hillier L, Kwok PY. 1998. Overlapping genomic sequences: a treasure trove of single-nucleotide polymorphisms. *Genome Res.* 8(7):748–54
 157. Tamayo P, Slonim D, Mesirov J, Zhu Q, Kitareewan S, et al. 1999. Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc. Natl. Acad. Sci. USA* 96(6):2907–12
 158. Tatusov RL, Koonin EV, Lipman DJ. 1997. A genomic perspective on protein families. *Science* 278:631–37
 159. Tavazoie S, Hughes JD, Campbell MJ, Cho RJ, Church GM. 1999. Systematic determination of genetic network architecture. *Nat. Genet.* 22(3):281–85
 160. Teichmann SA, Chothia C, Gerstein M. 1999. Advances in structural genomics. *Curr. Opin. Struct. Biol.* 9(3):390–99
 161. Thompson JD, Plewniak F, Poch O. 1999. A comprehensive comparison of multiple sequence alignment programs. *Nucleic Acids Res.* 27(13):2682–90
 162. Uberbacher EC, Mural RJ. 1991. Locating protein-coding regions in human DNA sequences by a multiple sensor-neural network approach. *Proc. Natl. Acad. Sci. USA* 88:11261–65
 163. Uberbacher EC, Xu Y, Shah MB, Oلمان V, Parang M, Mural RJ. 1998. An editing environment for DNA sequence analysis and annotation. *Pac. Symp. Biocomput.* 3:217–27 (Abstr.)

164. Velculescu VE, Zhang L, Vogelstein B, Kinzler KW. 1995. Serial analysis of gene expression. *Science* 270:484–87
165. Walsh S, Anderson M, Cartinhour SW. 1998. ACEDB: a database for genome information. *Methods Biochem. Anal.* 39:299–318
166. Wasserman WW, Fickett JW. 1998. Identification of regulatory regions which confer muscle-specific gene expression. *J. Mol. Biol.* 278(1):167–81
167. Deleted in proof
168. Williamson AR. 1999. The Merck Gene Index project. *Drug Discov. Today* 4(3):115–22
169. Wolf YI, Brenner SE, Bash PA, Koonin EV. 1999. Distribution of protein folds in the three superkingdoms of life. *Genome Res.* 9(1):17–26
170. Wolfsberg TG, Landsman D. 1997. A comparison of expressed sequence tags (ESTs) to human genomic sequences. *Nucleic Acids Res.* 25(8):1626–32
171. Wood TC, Pearson WR. 1999. Evolution of protein sequences and structures. *J. Mol. Biol.* 291(4):977–95
172. Wootton JC, Federhen S. 1993. Statistics of local complexity in amino acid sequences and sequence databases. *Comput. Chem.* 17:149–63
173. Xu Y, Uberbacher EC. 1997. Automated gene identification in large-scale genomic sequences. *J. Comput. Biol.* 4(3):325–38
174. Zhang J, Madden TL. 1997. PowerBLAST: a new network BLAST application for interactive or automated sequence analysis and annotation. *Genome Res.* 7(6):649–56
175. Zhang MQ. 1997. Identification of protein coding regions in the human genome by quadratic discriminant analysis. *Proc. Natl. Acad. Sci. USA* 94(2):565–68
176. Zhang MQ. 1998. Large-scale gene expression data analysis: a new challenge to computational biologists. *Genome Res.* 9(8):681–88
177. Zhang Z, Pearson WR, Miller W. 1997. Aligning a DNA sequence with a protein sequence. *J. Comput. Biol.* 4(3):339–49
178. Zhang Z, Schaffer AA, Miller W, Madden TL, Lipman DJ, et al. 1998. Protein sequence similarity searches using patterns as seeds. *Nucleic Acids Res.* 26(17):3986–90