Downloaded from genome.cshlp.org on June 21, 2026 . Published by Cold Spring Harbor Laboratory Press

Perspectives: sequence data base searching in the era of large-scale genomic sequencing.

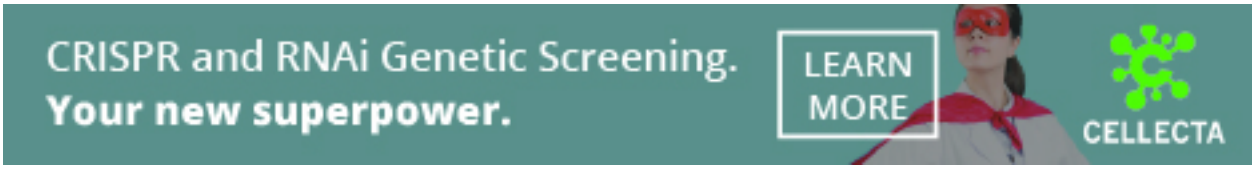
R F Smith

Genome Res. 1996 6: 653-660
Access the most recent version at doi:10.1101/gr.6.8.653

References This article cites 33 articles, 6 of which can be accessed free at:
http://genome.cshlp.org/content/6/8/653.full.html#ref-list-1

License

Email Alerting Service Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or **click here.**



To subscribe to *Genome Research* go to:
https://genome.cshlp.org/subscriptions

Copyright © Cold Spring Harbor Laboratory Press

Downloaded from genome.cshlp.org on June 21, 2026 . Published by Cold Spring Harbor Laboratory Press

REVIEW

Perspectives: Sequence Data Base Searching in the Era of Large-scale Genomic Sequencing

Randall F. Smith[1]

Human Genome Center, Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, Texas 77030 USA

Large-scale sequencing of human and model organism genomes will have a profound impact on our ability to use sequence data base searching to predict the biochemical functions of sequences of interest. Despite the great value of more sequences in the data bases, a huge increase in data base size will also have adverse effects on data base searches. Upcoming problems will include (1) greatly increased search times, (2) an increase in background noise of high-scoring but biologically irrelevant matches, (3) inaccurate coding region prediction, leading to problems in protein data base searching, and (4) limited first-pass sequence annotation, making it difficult to determine the biological relevance of data base hits. Improved data base annotation tools and construction of smaller data bases of representative and highly-annotated sequences for first-pass analyses will be essential to deal with the impending flood of new genomic sequence.

Before 1995, the size of GenBank doubled about every 21 months. During 1995, however, the data base nearly doubled in size in one year alone, growing from 230 Mb in December 1994 to 426 Mb in December 1995. This growth is primarily the result of major sequencing initiatives that have expanded during the last year, including the Merck–Washington University (St. Louis, MO) expressed sequence tag (EST) project (which alone has added nearly 100 Mb of human cDNA sequence in 1995) and the expansion of several whole-genome sequencing efforts. The latter includes the recently completed sequencing of several bacterial genomes (including the 1.8-Mb *Haemophilus influenzae* and the 0.58-Mb *Mycoplasma genitalium* genomes), the completion of the 14-Mb yeast (*Saccharomyces cerevisiae*) genome, rapid progress on the sequencing of the 100-Mb genome of *Caenorhabditis elegans*, and expanding work on the sequence of the 165-Mb *Drosophila* genome (see Gibbs 1995). The proposed plan to sequence completely the 3000-Mb human genome during the next 5–7 years (Collins and Galas 1993) alone will produce a six-fold increase in the current size of GenBank.

[1]Present address: SmithKline Beecham Pharmaceuticals, R&D Bioinformatics Research Group, King of Prussia, Pennsylvania 19406. E-MAIL Randall_F_Smith@sbphrd.com; FAX (610) 270-5580.

Benefits of Increased Gene Diversity

The rapid expansion of large-scale genomic sequencing efforts will have both a positive and a negative impact on our ability to use sequence data base searches to predict the biochemical functions of sequences of interests. On the positive side, large-scale genomic sequencing will identify new gene families and increase the size and diversity of families identified previously. With larger numbers of diverse homologs in the data bases (both within and between species), it will be much easier to identify new family members when conducting standard sequence data base searches.

An increase in the size and diversity of gene families will also facilitate the identification of diagnostic sequence motifs (fingerprints) for gene and protein functional domains. Conserved sites and regions identified in multiple sequence alignments of family members, for example, can be used to construct new sequence patterns, such as those currently identified in the PROSITE, BLOCKS, and PRINTS data bases (Attwood et al. 1996; Bairoch et al. 1996; Pietrokovski et al. 1996). Such patterns can be used in turn to enhance the identification of functional domains in distantly related homologs (e.g., Henikoff and Henikoff 1994; Koonin et al. 1994; Tatusov et al. 1994; Smith and King 1995).

The greatest increase in gene diversity will come from sequencing the genomes of phylogenetically diverse organisms from different king-

Downloaded from genome.cshlp.org on June 21, 2026 . Published by Cold Spring Harbor Laboratory Press

doms or phyla. Therefore, it is encouraging that the sequencing of the genomes of a wide variety of organisms (e.g., archaebacteria, eubacteria, yeast, *C. elegans*, *Drosophila*, and humans) is already under way.

The Problem of Increasing Search Times

The most obvious negative impact of a huge increase in data base size will be to dramatically lengthen data base search times, because for all the current popular sequence data base search tools (e.g., BLAST, FASTA), the time it takes to conduct a full search is directly proportional to the sum of the lengths of all of the sequences in the data base. Fortunately, raw computer speeds have also been doubling every 1–2 years and, by all accounts, this trend is expected to continue for the foreseeable future. This alone should maintain search times at roughly the present rate. In addition, advances in parallel computing, including combining networked workstations into virtual parallel machines, should increase search speeds greatly. For example, a number of groups, including those at the Oak Ridge National Laboratories, the Lawrence Berkeley Laboratory, and the Pittsburgh Supercomputing Center, have been using the PVM (Parallel Virtual Machine) software package (Geist et al. 1994) to run data base search programs in parallel on collections of networked workstations (Shah et al. 1994; Zorn et al. 1994; Ropelewski et al. 1995).

Search Speed vs. Thoroughness Tradeoffs

Development of inherently faster data base search tools will also most likely continue. However, it is typical of general search problems in computer science that there is a trade-off between search speed and search thoroughness, so that further increases in speed come at the expense of a loss in either search sensitivity (the ability to detect true positive matches) or search specificity (the ability to reject false-positive matches). This is seen, for example, in the three most commonly used sequence data base search methods (BLAST, FASTA, and Smith–Waterman algorithm; Smith and Waterman 1981; Pearson 1989; Altschul et al. 1990) where in terms of speed, BLAST > FASTA >> Smith–Waterman, whereas for search performance, Smith–Waterman > FASTA > BLAST (Pearson 1995). Given this tradeoff, further efforts to use parallel computing to improve the search speeds of most sensitive search methods currently available should be a high priority in future work.

Distributing Search Services

Unfortunately, we have recently entered into a transition phase where the recent rapid increase in data base size has impacted search performance for the average researcher. More and more researchers have been turning to centralized resources to conduct sequence data base searches [e.g., the National Center for Biotechnology Information's (NCBI) BLAST Network Server; Benson et al. 1996]. As a consequence, the performance of these servers has been worsening. In addition, the recent rapid increase in worldwide Internet network traffic has also had a negative impact on performance, especially between continents. One remedial approach to the combined problems of overloaded networks and centralized servers would be to distribute the search loads across several geographically distant servers. To do this, methods to automatically distribute and update sequence data bases over the Internet need to be put into place to support any site wishing to establish a public (or private) search server. The Genome Sequence Database (GSDB), for example, currently supports a data base replication system for installing and automatically maintaining its relational data base at remote sites (GSDB "satellities;" see http://www.ncgr.org/gsdb/gsdb.html). Other data bases need to follow suit.

As large-scale human and model-organism genomic sequencing progresses, however, individual researchers may be sequencing fewer and fewer genes themselves, and therefore thus fewer and fewer data base searches would need to be conducted by individuals. In the medium-to-long term, relatively more searches will be conducted "in-house" by the large-scale sequence producers or by centralized groups responsible for annotating new genomic sequence. These groups presumably will have much better computational resources than the average researcher. Hopefully this trend, in combination with increased use of parallel computing to improve search speeds, will take us past the short-term bottleneck that we have been experiencing recently in decreasing search performance.

Downloaded from genome.cshlp.org on June 21, 2026 . Published by Cold Spring Harbor Laboratory Press

The Problem of Increasing Background Noise: The Larger the Data Base, the Higher the Score Required for a Statistically Significant Match

According to the theory of score probabilities used in the BLAST program, the minimal match score required for a specified level of statistical significance is proportional to the natural logarithm of the data base size (Altschul et al. 1990; Karlin et al. 1990; Altschul 1990). Therefore, a 10-fold increase in data base size would result in a 2.3-fold increase in the minimal score required for a statistically significant match. However, the degree of similarity between homologous sequences, as reflected in match scores, is not a function of data base size. Therefore, as the size of the data base increases, a large number of random matches may actually get higher scores than distantly related sequences, obscuring biologically meaningful similarities.

Reducing Data Base Redundancy

Given this problem, any approach that would reduce apparent data base size would be beneficial. Several nonredundant sequence data bases have already been developed (e.g., the OWL protein sequence data base, the NCBI's nonredundant protein and DNA data bases, The Institute of Genome Research's Human Transcript data base; Gish 1992; Bleasby et al. 1994; Adams et al. 1995), where identical or nearly identical sequences are identified and only one example is used. An even further reduction in redundancy could be achieved by using only a single representative from each distinct sub-family within a gene or protein family. For example, the sequence similarity ("neighbor") information provided in the NCBI's Entrez data base (Benson et al. 1996) can be used to cluster sequences into families and sub-families (see, e.g., Worley et al. 1995). Representative sequences can then be picked for inclusion into low-redundancy data bases for use with standard sequence data base search programs.

Another very effective method of reducing data base size is to represent each family as a single multiple alignment or as sequence patterns, motifs, or profiles derived from multiple alignments, as described above. Specialized search programs can then be used to compare query sequences against such alignment and pattern data bases (e.g., BLOCK, PROFILE, Markov model, and PROSITE search tools; Gribskov et al. 1988; Fuchs 194; Krogh et al. 1994; Henikoff et al. 1995).

First-pass Searches of Smaller, Highly Annotated Data Bases

Sorting through data base search results, especially where large numbers of only very weak matches are observed, can be a daunting task given even the current size of the data bases. The increased size, complexity, and background noise associated with a 10- or 100-fold increase in data base size would make this task exponentially more complicated. One approach that would greatly simplify sequence searching would be to conduct first-pass analyses using a representative data base where the functions and features of each sequence in the data base is known and clearly labeled. If there are no strong matches to any of the representative sequences, then a search of a more complete data base can be conducted.

Constructing highly annotated data sets has been an ongoing major effort of a number of groups, including the PIR-International and SWISS-PROT protein sequence data bases (Bairoch and Apweiler 1996; George et al. 1996; also see other group reports in the recent data base issue of *Nucleic Acids Res.*, Volume 24, number 1, 1996). The PIR data base, for example, includes a large subset, where each sequence has been annotated and classified into a protein superfamily (George et al. 1996). For the December 1995 release of the PIR data base, 35,000 entries (of 82,000 total) have been classified into 3700 superfamilies. When searching such data sets, knowing the function of each matched sequence greatly facilitates the analysis of similarity search results. Generating smaller representative data bases from such highly annotated data sets that are optimized for both protein and DNA sequence searches should be another high priority in future work.

The Problem of Coding Region Identification

Advantages of Protein Sequence Data Base Searching

As large-scale genomic sequencing scales up, the accuracy of protein-coding region prediction within newly sequenced genomic regions will be an important issue for sequence data base searching. The reason is that protein vs. protein se-

Downloaded from genome.cshlp.org on June 21, 2026 . Published by Cold Spring Harbor Laboratory Press

quence searches are much more sensitive in detecting distantly related sequences than DNA vs. DNA sequence searches (Doolittle et al. 1986; Pearson 1996). This increased sensitivity is attributable to two factors: (1) the triplet nucleotide code is degenerate; therefore, the amino acid sequences of distantly related sequences are more similar than their respective DNA sequences, and (2) the use of amino acid scoring matrices such as the PAM and BLOSUM matrices (Dayhoff et al. 1978; Henikoff and Henikoff 1992) allows matches between structurally and functionally conserved amino acids to have positive scores rather than being treated as negative scoring mismatches (Pearson 1996). Therefore, when one is analyzing a new gene sequence, if the coding region is known (e.g., from a cDNA sequence), then the most informative search would be to use the translated amino acid sequence as the query against a protein sequence data base. If the coding region of a sequence of interest is not known (e.g., genomic DNA, a 5′- or 3′-derived EST that may or may not contain coding sequence), then programs such a BLASTX can be used to search a six-frame translation of the query sequence against a protein sequence data base (Gish and States 1993; Altschul et al. 1994).

Updating the Protein Sequence Data Bases from Genomic Sequences

Given the importance of protein data base searching for gene function identification, it is a critical issue as to how the protein sequence data bases will be updated given large-scale genomic sequence submissions. Currently, most sequences that enter the protein sequence data bases are translations of coding regions annotated by researchers submitting individual DNA sequences. Unfortunately, there are no current guidelines regarding how putative coding regions in DNA sequences submitted by genomic sequencing projects are to be annotated. Most large-scale sequencing groups will likely annotate genomic sequences with coding regions and gene structures predicted using computational approaches, as is currently being done by the *C. elegans* sequencing groups (Berks 1995). However, the accuracy of gene structure prediction programs is currently modest at best. Given inaccuracies in the predicted locations of donor and acceptor splice sites, the frequencies of (1) completely missing exons, (2) wrong exons (predicted exons not overlapping actual exons), and (3) predicted exons that partially overlap actual exons result in predicted protein sequences that have on average only 52%–62% similarity to the true amino acid sequences of vertebrate genes, depending on the program (Burset and Guigo 1996). The wholesale inclusion of predicted sequences into the protein data bases would therefore litter these data bases with inaccurately predicted coding sequences, decreasing the effectiveness of protein sequence data base searching.

Updating the protein sequence data bases with the translations of all open reading frames (ORFs) greater than some minimal length would prevent missed coding regions from being lost from the protein data bases. However, increased background noise as a result of the presence of large numbers of true noncoding ORFs, coupled with the inability to detect matches that cross splice junctions, would limit the value of this approach. Another approach would be to forego searching protein sequence data bases altogether and use programs such as TBLASTN and TFASTA. These programs can be used to search a protein sequence query against a complete six-frame translation of a DNA sequence data base (see Altschul et al. 1994). However, similar to ORF searches, matches across splice junctions would be missed and background noise would be increased significantly because the effective data base size is six times the translated length of the DNA data base being searched. Using the 340-Mb NCBI nonredundant DNA data base, for example, $(6 \times 340\ \text{Mb}/3) = 680\ \text{Mb}$, 12 times the size of the 57-Mb NCBI nonredundant protein sequence data base. For this reason, one should perform TBLASTN or TFASTA searches only if a standard BLASTP or FASTA search of a protein sequence data base yields no informative matches.

Because all of the above approaches for updating protein sequence data bases using genomic sequence data have serious drawbacks, a concerted effort to improve computational approaches for predicting gene structure should be a high priority for future work, and/or large-scale sequencing of full-length cDNAs from a variety of cell and tissue types and developmental stages should be initiated.

One additional advantage of generating high-quality protein sequence data would be that the protein data bases would not grow as quickly relative to the DNA sequence data bases. For example, whereas sequencing the 3000-Mb human genome will increase the size of GenBank sixfold

Downloaded from genome.cshlp.org on June 21, 2026 . Published by Cold Spring Harbor Laboratory Press

over its current size (500 Mb), the protein coding portion of the 80,000–100,000 genes present in the human genome will contribute roughly only 100–200 Mb of sequence (at 1–2 kb of protein sequence per gene). This is only about two- to fourfold the size of the NCBI's nonredundant protein sequence data base (57 Mb from 202,000 sequence entries). Also, because the model organisms being sequenced (e.g., bacteria, *C. elegans*, and *Drosophila*) only have 4,000–15,000 genes apiece, these sequencing projects will not increase the total number of protein sequences in the data bases significantly.

The relatively slower growth of the protein sequence data bases would therefore further ease the burden on increasing search times with increasing data base sizes. These estimates assume, however, that only one protein sequence per gene will be deposited in the data bases, and it is an open question as to the number of different low-level alternatively expressed products that might exist for an average gene in various cell and tissue types and at different times in development. Therefore, sometime in the future the sequences of alternatively expressed genes could significantly increase the number of sequences in the protein data bases.

The Problem of Sequence Annotation

Large-scale sequencing will require large-scale sequence annotation, if the results of genomic sequencing efforts are to be of maximal value to the community. Currently, researchers sequencing individual genes of interest spend a great deal of time and effort annotating their sequences with various features (e.g., mRNA and protein-coding regions, promoter regions, binding sites, and repeats) before submission to the public data bases. For large-scale, very high-throughput sequencing, however, first-pass sequence annotation will need to be mostly automated. Software systems for automated sequence annotation are currently being developed (e.g., BASS, GeneQuiz; Scharf et al. 1994; Adams et al. 1995; Casari et al. 1995). However, it is unlikely that automated systems will, anytime in the foreseeable future, provide the level of accuracy, detail, and biological content and insight matched by researchers annotating their own sequences of interest.

Transitive Annotations

Another potential problem with automated annotation is that it may exacerbate what Steve Ladunga in my group has called the problem of "transitive annotation." This problem is illustrated as follows. A computer (or human) finds that a new sequence, *A*, is similar to a data base sequence, *B*, and assigns *A* the function of *B*. However, *B* may have had its function assigned from a previous automated search, where *B* was found to be similar to sequence *C*, and so on. Transitively assigning function to a series of closely related sequences would probably not be a serious issue. However, transitive annotation could produce misleading or totally incorrect assignments in cases where one or more pairs of sequences in a long series are related only weakly. Always using a single reference data base for sequence searching, where the features and functions of entries have been experimentally verified in as many cases as possible, would be one approach to this problem. Certainly, it should always be made clear in data base entries if a particular annotation is derived from experimental evidence or inferred from sequence similarity results.

Allowing Third Parties to Annotate and Correct Data Base Entries

One approach for improving sequence annotation is to build tools that allow researchers who are working with previously submitted sequences (third parties) to easily annotate sequences of interest. Currently, it is the general policy of the sequence data bases that only the original submitter of a sequence is allowed to modify a data base entry. Such a policy is untenable given even our current state, where a large fraction of all sequence generated in the world is produced by large-scale sequencing projects. Researchers who are carefully studying the biology of particular genes must be able to annotate sequence on which they are working, regardless of whether or not they generated the original sequence themselves. For this reason, the GSDB is currently developing tools to allow and facilitate third-party sequence annotation of data base entries (Cinkosky et al. 1995; Keen et al. 1996). There are several thorny issues involved in allowing third-party annotation (e.g., Should annotations be viewed as a type of publication in order to provide an incentive for researchers to put time and effort into annotating sequences? Should annotations be peer-reviewed? Should anybody "off-the-street" be allowed to annotate sequences?).

Downloaded from genome.cshlp.org on June 21, 2026 . Published by Cold Spring Harbor Laboratory Press

These issues need to be debated carefully before such third-party annotation systems are put into use.

Need for Improved Access to Information of Sequence Features and Functions

For sequence data base searching, the accuracy and detail of sequence annotations are particularly important for the following reasons. Identifying the function of a newly sequenced gene by data base searching requires that two criteria be met: (1) Matches must be statistically significant (less probable than chance alone), and (2) the functions of the data base sequences matched in a search must be known—if none of the data base sequences matched in a search has a known function, then nothing can be inferred about the function of the query sequence. Even if the functions of matched sequences are known, inferring the function of a query sequence from the matches is often a difficult task. This is because the functions of data base sequences are often not readily identifiable from the information provided in data base search results. The only information on sequence features and functions provided by the most commonly used data base search tools is the sequence's title, and the title alone (e.g., "p20 protein") is often uninformative as to sequence's function. Therefore, not only is sequence annotation important for data base searching, but so is access to this information, as well as to any other types of information available about sequences that may originate from other sources. For this reason, any method that improves access to information on the functions of sequences matched in a data base search would facilitate the analysis of data base search results.

Fortunately, the emergence of the World-Wide Web (WWW) has provided a very useful approach for facilitating access to information about sequences. A large number of sequence data base search tools are now accessible through the Web, allowing additional information about sequences to be incorporated directly into data base search results. For example, search results returned by the NCBI's BLAST WWW Network Server (http://www.ncbi.nlm.nih.gov/BLAST), include imbedded hypertext links that provide direct access to GenBank reports for matched sequences. BLAST search results returned by the Baylor College of Medicine's WWW BEAUTY server also includes information on the locations of all annotated sites and domains within matching sequences, as well as hypertext links to Medline literature abstracts for matched sequences (Worley et al. 1995; http://gc.bcm.tmc.edu:8088/search-launcher/launcher.html). Progress on extending links to other useful resources for information on sequences should also be made. For example, sequences from large-scale genomic sequencing of model organisms are also being maintained in specialized species-specific data bases such as ACeDB (Durbin and Thierry Mieg 1991; see http://gdbwww.gdb.org/gdb/docs/genomic_links.html and http://probe.nalusda.gov:8000/acedocs/allace.html for lists of model organism data bases). Sequences in these data bases are being actively annotated and curated by individuals and groups working on these model systems, and new sequence-related information that is being added to these data bases could and should be made available to those conducting general sequence data base searches.

The use of the WWW for integrating and visualizing information on sequences matched in a data base search is just a first step in building the types of analysis tools that will be required in order to cope with the huge amount of data on sequences, their features, and functions that will be generated by large-scale genomic sequencing. Clearly, to maximize the information that will be discovered on the biology of sequences produced by all of the various genome programs, enhanced data base search tools need to be developed that will integrate and graphically display as much information as possible about the features and functions of matched genomic and expressed sequences, including, for example, promoter sites, exon/intron locations, alternate splice sites, functional domains and active sites, and genetic and physical map locations.

Conclusions

Changes to the current methods of sequence data base searching must be made to adequately address the impact of large-scale genomic sequencing. The need for increased speed of sequence data base searches to compensate for large increases in data base size will be addressed through computational improvements, primarily in the use parallel hardware and distributed search servers. However, it will be more important than ever to maximize the biological information available in the sequence data bases while at the same time minimizing data base size and complexity for

Downloaded from genome.cshlp.org on June 21, 2026 . Published by Cold Spring Harbor Laboratory Press

first-pass analyses, if the informativeness of sequence data base searching is to be maintained or improved during this new era of large-scale genomic sequencing.

ACKNOWLEDGMENTS

I thank Dan Davison, Jim Fickett, Karen Kabnick, Richard Gibbs, Bill Pearson, and my research group—Istvan Ladunga, Brent Wiese, and Kim Worley, for their helpful comments. Work in my group is supported by grants from the National Center for Human Genome Research, National Institutes of Health (1R01-HG00973-01 and P30-HG00210).

REFERENCES

Adams, M.D., A.R. Kerlavage, R.D. Fleischmann, R.A. Fuldner, C.J. Bult, N.H. Lee, E.F. Kirkness, K.G. Weinstock, J.D. Gocayne, O. White, et al. 1995. Initial assessment of human gene diversity and expression patterns based upon 83 million nucleotides of cDNA sequence. *Nature* (Suppl.) **377:** 3–173.

Altschul, S.F., W. Gish, W. Miller, E.W. Myers, and D.J. Lipman. 1990. Basic local alignment search tool. *J. Mol. Biol.* **215:** 403–410.

Altschul, S.F., M.S. Boguski, W. Gish, and J.C. Wooton. 1994. Issues in searching molecular sequence data bases. *Nature Genet.* **6:** 119–129.

Attwood, T.K., M.E. Beck, A.J. Bleasby, K. Degtyarenko, and D.J. Parry Smith. 1996. Progress with the PRINTS protein fingerprint data base. *Nucleic Acids Res.* **24:** 182–188.

Bairoch, A. and R. Apweiler. 1996. The SWISS-PROT protein sequence data bank and its new supplement TREMBL. *Nucleic Acids Res.* **24:** 21–25.

Bairoch, A., P. Bucher, and K. Hofmann. 1996. The PROSITE data base, its status in 1995. *Nucleic Acids Res.* **24:** 189–196.

Benson, D.A., M. Boguski, D.J. Lipman, and J. Ostell. 1996. GenBank. *Nucleic Acids Res.* **24:** 1–5.

Berks, M. 1995. The *C. elegans* genome sequencing project. *Genome Res.* **5:** 99–104.

Bleasby, A.J., D. Akrigg, and T.K. Attwood. 1994. OWL—A non-redundant composite protein sequence data base. *Nucleic Acids Res.* **22:** 3547–3577.

Burset, M. and R. Guigo. 1996. Evaluation of gene structure prediction programs. *Genomics* (in press).

Casari, G., M. Andrade, P. Bork, J. Boyle, A. Daruvar, C. Ouzounis, R. Schneider, J. Tamames, A. Valencia, and C. Sander. 1995. Challenging times for bioinformatics. *Nature* **376:** 647–648.

Cinkosky, M.J., J.W. Fickett, and G.M. Keen. 1995. A new design for the Genome Sequence Data Base. *IEEE Eng. Med. Biol.* **14:** 725–729.

Collins, F. and D. Galas. 1993. A new five-year plan for the U.S. Human Genome Project. *Science* **262:** 43–46.

Dayhoff, M., R.M. Schwartz, and B.C. Orcutt. 1978. A model of evolutionary change in proteins. In *Atlas of protein sequence and structure* (ed. M. Dayhoff), Vol. 5, Suppl. 3, pp. 345–352. National Biomedical Research Foundation, Silver Spring, MD.

Doolittle, R.F., D.F. Feng, M.S. Johnson, and M.A. McClure. 1986. Relationships of human protein sequences to those of other organisms. *Cold Spring Harbor Symp. Quant. Biol.* **51:** 447–455.

Durbin, R. and J. Thierry Mieg. 1991. A. C. elegans data base. Documentation, code, and data available via anonymous ftp; see ACEDB FAQ at http://probe.nalusda.gov:8000/acedbfaq.html.

Fuchs, R. 1994. Predicting protein function: A versatile tool for the Apple Macintosh. *Comput. Appl. Biosci.* **10:** 171–178. See also http://www.ebi.ac.uk/searches/prosite_input.html.

Geist, A., A. Beguelin, J. Dongarra, W. Jiang, R. Mancheck, and V. Sunderam. 1994. PVM: Parallel virtual machine—A user's guide and tutorial for networked parallel computing. MIT Press, Cambridge, MA. Also available online at http://www.netlib.org/pvm3/book/pvm-book.html.

George, D.G., W.C. Barker, H.W. Mewes, F. Pfeiffer, and A. Tsugita. 1996. The PIR-International protein sequence data base. *Nucleic Acids Res.* **24:** 17–20.

Gibbs, R.A. 1995. Pressing ahead with human genome sequencing. *Nature Genet.* **11:** 121–125.

Gish, W. 1992. The nrdb program. National Center for Biotechnology Information. See ftp://ncbi.nlm.nih.gov/pub/nrdb/README.

Gish, W. and D.J. States. 1993. Identification of protein coding regions by data base similarity search. *Nature Genet.* **3:** 266–272.

Gribskov, M., M. Homyak, J. Edenfield, and D. Eisenberg. 1988. Profile scanning for three-dimensional structural patterns in protein sequences. *Comput. Appl. Biosci.* **4:** 61–66.

Henikoff, S. and J.G. Henikoff. 1992. Amino acid substitutions matrices from protein blocks. *Proc. Natl. Acad. Sci.* **89:** 10915–10919.

———. 1994. Protein family classification based on searching a data base of blocks. *Genomics* **19:** 7–107.

Downloaded from genome.cshlp.org on June 21, 2026 . Published by Cold Spring Harbor Laboratory Press

Henikoff, S., J.G. Henikoff, W.J. Alford, and S. Pietrokovski. 1995. Automated construction and graphical presentation of protein blocks from unaligned sequences. *Gene* **163:** GC17–26. See also http://blocks.fhcrc.org/.

Karlin, S. and S.F. Altschul. 1990. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc. Natl. Acad. Sci.* **87:** 2264–2268.

Karlin S., A. Dembo, and T. Kawabata. 1990. Statistical composition of high-scoring segments from molecular sequences. *Ann. Stat.* **18:** 571–581.

Keen, G., J. Burton, D. Crowley, E. Dickenson, A. Espinosa-Lujan, E. Franks, C. Harger, M. Manning, S. March, M. McLeod, et al. 1996. The Genome Sequence Database (GSDB): Meeting the challenge of genomic sequencing. *Nucleic Acids Res.* **24:** 13–16.

Koonin, E.V., A.R. Mushegian, R.L. Tatusov, S.F. Altschul, S.H. Bryant, P. Bork, and A. Valencia. 1994. Eukaryotic translation elongation factor 1γ contains a glutathione transferase domain—Study of a diverse, ancient protein superfamily using motif search and structural modeling. *Protein Sci.* **3:** 2045–2054.

Krogh, A., M. Brown, I.S. Mian, K. Sjolander, and D. Haussler. 1994. Hidden Markov models in computational biology. *J. Mol. Biol.* **235:** 1501–1531.

Pearson, W.R. 1989. Rapid and sensitive sequence comparisons with FASTP and FASTA. *Methods Enzymol.* **183:** 63–98.

———. 1995. Comparison of methods for searching protein sequence data bases. *Protein Sci.* **4:** 1145–1160.

———. 1996. Effective protein sequence comparison. *Methods Enzymol.* **266** (in press).

Pietrokovski, S., J.G. Henikoff, and S. Henikoff. 1996. The Blocks data base—A system for protein classification. *Nucleic Acids Res.* **24:** 197–200.

Ropelewski, A.J., H.B. Nichols Jr., and S.H. Fish. 1995. Msearch—Multiple data base search and alignment code. Pittsburgh Supercomputing Center. URL: http://www.psc.edu/~ropelews/msearch.html.

Scharf, M., R. Schneider, G. Casari, P. Bork, A. Valencia, C. Ouzounis, and C. Sander. 1994. GeneQuiz: A workbench for sequence analysis. In *Proceedings of the second international conference on intelligent systems for molecular biology* (ed. R. Altman, D. Brutlag, P. Karp., R. Lathrop, and D. Searls), pp. 348–353. AAAI Press, Menlo Park, CA.

Shah, M.B., X. Guan, J.R. Einstein, S. Matis, Y. Xu, R.J. Mural, and E.C. Uberbacher. 1994. User's guide to GRAIL and GENQUEST (Sequence Analysis, Gene Assembly and Sequence Comparison Systems) E-mail Servers and XGRAIL (Version 1.2) and XGENQUEST (Version 1.1) Client-Server Systems. Oak Ridge National Laboratories. URL: http://avalon.epm.ornl.gov/manuals/grail-genquest.9407.html.

Smith, R.F. and K.Y. King. 1995. Identifcation of a eukaryotic-like protein kinase gene in Archaebacteria. *Protein Sci.* **4:** 126–129.

Smith, T.F. and M.S. Waterman. 1981. Identification of common molecular subsequences. *J. Mol. Biol.* **13:** 195–197.

Tatusov, R.L., S.F. Altschul, and E.V. Koonin. 1994. Detection of conserved segments in proteins: Iterative scanning of sequence data bases with alignment blocks. *Proc. Natl. Acad. Sci.* **91:** 12091–12095.

Worley, K.C., B.A. Wiese, and R.F. Smith. 1995. BEAUTY: An enhanced BLAST-based search tool that integrates multiple biological information resources into sequence similarity search results. *Genome Res.* **5:** 73–184.

Zorn, M.D., J.F. Macfarlane, R. Armstrong, M.H. Cooper, and N.C. Weaver. 1994. BioPOET: Large scale sequence analysis on workstation farms. DOE Contractor-Grantee Workshop IV, November 13–17, 1994, Sante Fe, New Mexico. (CONF-9411116) p. 116. Available online at http://www.er.doe.gov/production/oher/genome/santafe/imformatics/zorn2.html.