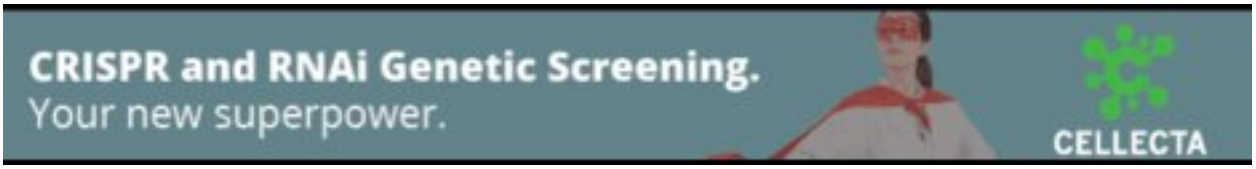
BEAUTY: an enhanced BLAST-based search tool that integrates multiple biological information resources into sequence similarity search results.

K C Worley, B A Wiese and R F Smith

Genome Res. 1995 5: 173-184
Access the most recent version at doi:10.1101/gr.5.2.173



RESEARCH

BEAUTY: An Enhanced BLAST-based Search Tool that Integrates Multiple Biological Information Resources into Sequence Similarity Search Results

Kim C. Worley,[1] Brent A. Wiese, and Randall F. Smith

Human Genome Center, Department of Molecular and Human Genetics, Department of Cell Biology, and W.M. Keck Center for Computational Biology, Baylor College of Medicine, Houston, Texas 77030

BEAUTY (BLAST enhanced alignment utility) is an enhanced version of the NCBI's BLAST data base search tool that facilitates identification of the functions of matched sequences. We have created new data bases of conserved regions and functional domains for protein sequences in NCBI's *Entrez* data base, and BEAUTY allows this information to be incorporated directly into BLAST search results. A Conserved Regions Data Base, containing the locations of conserved regions within *Entrez* protein sequences, was constructed by (1) clustering the entire data base into families, (2) aligning each family using our PIMA multiple sequence alignment program, and (3) scanning the multiple alignments to locate the conserved regions within each aligned sequence. A separate Annotated Domains Data Base was constructed by extracting the locations of all annotated domains and sites from sequences represented in the *Entrez*, PROSITE, BLOCKS, and PRINTS data bases. BEAUTY performs a BLAST search of those *Entrez* sequences with conserved regions and/or annotated domains. BEAUTY then uses the information from the Conserved Regions and Annotated Domains data bases to generate, for each matched sequence, a schematic display that allows one to directly compare the relative locations of (1) the conserved regions, (2) annotated domains and sites, and (3) the locally aligned regions matched in the BLAST search. In addition, BEAUTY search results include World-Wide Web hypertext links to a number of external data bases that provide a variety of additional types of information on the function of matched sequences. This convenient integration of protein families, conserved regions, annotated domains, alignment displays, and World-Wide Web resources greatly enhances the biological informativeness of sequence similarity searches. BEAUTY searches can be performed remotely on our system using the "BCM Search Launcher" World-Wide Web pages (URL is ⟨http://gc.bcm.tmc.edu:8088/search-launcher/launcher.html⟩).

The central goals of the human and model organism genome projects are to completely map and sequence the genes of these organisms. As work progresses, identification of the biochemical function of newly sequenced genes becomes a major challenge. Identification of gene function using traditional biochemical methods can be an extremely slow and laborious task that can take years of effort even for a single gene. Fortunately, computational methods are available that can greatly facilitate the identification of gene function. When a gene is isolated and sequenced, it can be matched against one or more of the publicly available sequence data bases, such as GenBank (Benson et al. 1993). If a similar gene of known function can be identified in such a data base search, then the function of the newly sequenced gene can be surmised by analogy. The biochemical functions of a growing number of genes, including a number of inherited human disease genes (for review, see Collins 1995), are being determined in this way.

Currently, high-speed heuristic methods, such as the hash-coding (k-tuple) algorithm employed by FASTA (Wilbur and Lipman 1983; Pearson and Lipman 1988; Pearson 1990) and the approximate word match algorithm employed by BLAST (Altschul et al. 1990, 1994), are the most commonly used sequence data base search programs. These programs produce a list of the sequence identifiers (e.g., locus names and accession numbers) and title lines of statistically significant matches followed by a display of the

[1]**Corresponding author.**
EMAIL kworley@bcm.tmc.edu; FAX (713) 798-5386.

5:173–184 ©1995 by Cold Spring Harbor Laboratory Press ISSN 1054-9803/95 $5.00

alignments of the query with each of the matched sequences. For a sequence data base search result to be informative, two criteria must be met: (1) The query sequence must have a statistically significant match to a data base sequence (a score greater than one expected by chance alone; Karlin and Altschul 1990), and (2) there must be information available about the function of the sequence matched. It is quite common, however, that the functions of matched sequences are not obvious from the search results. Often, sequence titles are uninformative (e.g., "ORF6") and one must laboriously retrieve and scan the full sequence data base reports to look for annotations that may identify the biological functions of the matched sequence. In addition, more often than not, functionally important conserved domains, such as enzyme active sites, are not noted as such in sequence data base records.

BEAUTY (BLAST enhanced alignment utility) addresses this latter aspect of the sequence identification problem, providing information about the function of the data base sequences matched in BLAST searches. BEAUTY incorporates information on sequence family membership, the location of the conserved regions, and annotated domains and sites directly into BLAST search results. These enhancements make it much easier to identify the functions of matched sequences, which is particularly important when trying to analyze the biological significance of weak data base hits.

RESULTS AND DISCUSSION

To allow easier identification of the functions of sequences matched in a data base search, we have created BEAUTY, an enhanced version of the National Center for Biotechnology Information (NCBI) BLAST search tool. BEAUTY performs a BLAST search of sequence data bases for which we have compiled functional information on each sequence. BEAUTY incorporates this information directly into BLAST search results by adding several new tables and figures to the standard BLAST output files (Table 1). First, a table is added that lists for each data base hit (1) the sequence family to which the data base sequence belongs, (2) the number of sequences within each family matched in the search, and (3) the total number of sequences in the family (Fig. 1). This table allows one to quickly assess the number of different families matched in the search as well as the number of family members matched for each sequence family identified. Matches to only a single member in a sequence family are much more likely to be indicative of a random or spurious similarity, whereas data base hits that include all or most members of a sequence family would provide more convincing evidence that the query sequence is indeed related to that sequence family.

A new figure is then added for each data base sequence matched. This figure shows the locations of each of the local BLAST hits within the query sequence (Fig. 2) and allows one to quickly assess if the hits occur primarily in one or a few local regions or if the hits are scattered throughout the sequence. Multiple hits within the same region of a query sequence may indicate a functionally important domain within that region. The query sequence is also compared with the PROSITE pattern data base, and the location of any matched patterns are displayed in the same figure. This allows one to immediately correlate the locations of hits with the locations of potential functional domains identified by PROSITE motif matches.

Third, for each data base sequence matched in a search, a figure is added that shows the locations of the local BLAST hits with respect to the positions of the conserved regions and any annotated sites and domains for that data base sequence (Fig. 3). These figures allow one to quickly assess if any potential functionally important regions have been hit during the data base search. Hits within all or most of the known conserved domains within a data base sequence are much more likely to be functionally important than hits within nonconserved regions. Also, query sequences with weak matches against some or all of the conserved domains within a data base sequence are much more likely to be related than cases where only nonconserved regions are matched sequences. Matches to conserved regions can also help identify potential functionally important domains in those cases where no annotation of functional domains is provided in the data base reports. When annotated domain information is available, this figure also allows a researcher to directly correlate the locations of such domains with the positions of all local BLAST hits within the sequence. In addition, hits matching known domains and sites are readily discernible without looking up the individual data base reports for each sequence.

In addition to these enhancements, BEAUTY

Table 1. Comparison of the information presented in BLAST and BEAUTY results as returned by their respective WWW-Search Servers

NCBI blast WWW server[a]	BCM beauty WWW server[b]
List of data base sequences with statistically significant matches	List of data base sequences with statistically significant matches
Alignments of HSPs for each match	Alignments of HSPs for each match
Hypertext links to flat file sequence reports	Hypertext links to flat file sequence reports
	Information on sequence family membership, including number of family members matched in the search
	Links to multiple alignments of sequence families (clusters)
	Summary diagram of the locations of all the local hits within the query sequence
	Location of all PROSITE pattern matches within query
	Links to PROSITE data base reports for these PROSITE matches
	Diagram showing relative locations of conserved regions, annotated domains, and locally aligned regions for each data base sequence matched
	Links to BLOCKS, PROSITE, PRINTS and *Entrez* data base reports for annotations
	Links to *Entrez* (including MEDLINE abstracts and nucleotide sequence reports)
	Links to related sequences ("protein neighbors")
	Links to SRS cross-referenced data bases (including PDB, EMBL, GenBank, ENZYME, PROSITE, OMIM, etc.)

[a]WWW URL: http://www.ncbi.nlm.nih.gov/Recipon/index.html
[b]WWW URL: http://dot.imgen.bcm.tmc.edu:9331/seq-search/protein-search.html
(HSPs) High-scoring segment pairs.

search results returned by our World-Wide Web (WWW) search interface include hypertext links to a number of in-house and external on-line resources. Using these links, additional functional information on matched sequences can be assessed immediately. For example, links are provided to the NCBI's WWW *Entrez* interface, allowing MEDLINE literature abstracts referenced in sequence reports to be retreived immediately and browsed for more detailed information on matched data base sequences. Links to the SRS (Sequence Retrieval System; Etzold and Argos 1993) WWW interface allow information cross-referenced from >30 linked data bases (including EMBL, GenBank, SWISSPROT, PIR, PDB, ENZYME, PROSITE, BLOCKS, and OMIM; see URL: http://www.embl-heidelberg.de/srs/srsc) to be obtained similarly. Links to our own alignment data base allow the multiple sequence alignment to be displayed for that family to which a given data base sequence is a member (Fig. 4). Using the WWW interface, all of this linked information can be easily browsed for biological meaning without the distraction of performing keyword searches separately for each of these individual on-line resources, then storing each of the results. As a result, thoroughly analyzing BEAUTY search results can take significantly less time than analyzing a corresponding standard BLAST search.

SBASE is another data base of sequences with annotated domains that can be remotely searched using BLAST (Pongor et al. 1994; URL: http://base.icgeb.trieste.it/sbase/). The

Locus_ID	Clus_ID	Cluster_Title	Members M / T
sp|P10081|IF4A	431.26	eukaryotic initiation elegans i factor 4a eif-	3/4
"	431.29	eukaryotic translation initiation rna factor e	27/30
sp|P10630|IF42	431.29	eukaryotic translation initiation rna factor e	27/30
gi|485388	431.29	eukaryotic translation initiation rna factor e	27/30
sp|P04765|IF41	431.29	eukaryotic translation initiation rna factor e	27/30
pir||S00986|gi	431.29	eukaryotic translation initiation rna factor e	27/30
sp|P34529|YM68	12682.1	hypothetical ribonuclease 208.3 kd protein iii	1/2
sp|P25808|SPB4	7196.1	putative atp-dependent rrna rna helicase spb4	2/2
"	7196.2	probable putative atp-dependent rrna rna helic	2/3
gi|563986	393.25	p68 p68-like putative atp-dependent rna protei	1/10
"	393.28	p68 p68-like pre-mrna putative atp atp-depende	2/24
sp|Q02748|IF4A	431.26	eukaryotic initiation elegans i factor 4a eif-	3/4
"	431.29	eukaryotic translation initiation rna factor e	27/30
sp|P33906|DEAD	637.22	putative saccharomyces atp-dependent cerevisia	6/23
gi|496902	431.29	eukaryotic translation initiation rna factor e	27/30
sp|P35683|IF4A	431.29	eukaryotic translation initiation rna factor e	27/30
pir||JN0839|gi	431.29	eukaryotic translation initiation rna factor e	27/30
sp|P23304|DEAD	637.22	putative saccharomyces atp-dependent cerevisia	6/23
gi|306875	4353.4	heterogeneous nab3p rna-binding nuclear protei	1/5
sp|P21693|DBPA	637.22	putative saccharomyces atp-dependent cerevisia	6/23
gi|475219	431.29	eukaryotic translation initiation rna factor e	27/30
gi|475213	431.29	eukaryotic translation initiation rna factor e	27/30
gi|485949	431.29	eukaryotic translation initiation rna factor e	27/30
gi|485947	431.29	eukaryotic translation initiation rna factor e	27/30
gi|475438	1699.10	a489r a498r a505r a506r a528r a542r multigene	2/11
gi|485987	431.29	eukaryotic translation initiation rna factor e	27/30
pir||JC1453|gi	431.29	eukaryotic translation initiation rna factor e	27/30
gi|485945	431.29	eukaryotic translation initiation rna factor e	27/30
sp|P26802|DB73	6760.2	d-e-a-d putative atp-dependent protein rna hel	2/3
gi|476338	637.22	putative saccharomyces atp-dependent cerevisia	6/23
sp|P34640|YOQ2	7196.1	putative atp-dependent rrna rna helicase spb4	2/2
"	7196.2	probable putative atp-dependent rrna rna helic	2/3
gi|532754	3306.4	extracellular storage ribonuclease ne protein	1/5
"	3306.5	extracellular storage ribonuclease ne protein	1/6
gi|499204	6760.2	d-e-a-d putative atp-dependent protein rna hel	2/3
pir||S22579|gi	431.29	eukaryotic translation initiation rna factor e	27/30
pir||S31229|gi	637.22	putative saccharomyces atp-dependent cerevisia	6/23
pir||S15963|gi	4247.4	hypothetical ntpase probable killer nucleic pl	3/5
gi|286075	393.18	putative atp-dependent vasa rna protein=xvlg1	1/7
"	393.28	p68 p68-like pre-mrna putative atp atp-depende	2/24
gi|337424	6033	nad+ polyadp-ribose adp-ribose adp-ribosyltran	1/3
sp|P10125|UVRB	2428.5	excinuclease uvr-402 uvrb abc homologoue prote	2/6
"	2428.6	excinuclease uvr-402 uvrb abc excision homolog	2/7
pir||S10342|gi	4247.4	hypothetical ntpase probable killer nucleic pl	3/5
sp|P05470|YKP4	4247.4	hypothetical ntpase probable killer nucleic pl	3/5
gi|433120	1699.10	a489r a498r a505r a506r a528r a542r multigene	2/11

Figure 1 BEAUTY output table showing information on sequence family membership, including the number of family members matched in the search. This example is from a BEAUTY search of the CRSeq data base using a yeast hypothetical protein (PIR locus S28368) as a query sequence. This table is added to the BLAST search results file immediately following the normal BLAST list of significant matches (not shown). Listed for each of the data base sequences hit, are the unique identfier (Clus_ID) and consensus title line (Cluster_Title) of the sequence family alignment containing the matched sequence. (*Right*) The number of family members matched in the search (M) along with the total number of members of that family (T). Clicking on a Locus_ID in this table jumps the reader to the alignment report for that hit (Fig. 3). The cluster identifiers (Clus_IDs) are hypertext linked to the cluster alignments (Fig. 4).

search results returned by the SBASE server more closely resemble standard BLAST search output and do not include the integrated conserved region/annotated region overview display nor the hypertext links to *Entrez* and other data bases provided by BEAUTY. The display of conserved regions provided in BEAUTY searches also allows the location of potential functionally important regions to be detected even for those sequences that do not have annotated sites or domains.

CONCLUSION

When using sequence data base searches to identify the function of new sequences, obtaining information on the functions of the matched data base sequences can be just as important as the degree of similarity observed in the search; even for highly significant matches, no functional identification can be made if the function of the matched sequences are not readily discernible from the search results. As we have shown here, BEAUTY greatly improves access to information about the function of matched data base sequences, and in so doing, it can significantly aid in the functional identification of a new sequence. Adaptation of these enhancements to other search tools, such as FASTA, is in progress. We predict that further work in enhancing access to functional information about sequences may have a greater practical impact on the usefulness of data base search tools than further enhancements to data base search algorithms, per se.

METHODS

Figure 5 illustrates the general scheme used to develop BEAUTY. Details of the steps taken are described below.

Conserved Regions Data Base Construction

To directly incorporate information on locations of conserved regions within a matched data base sequence into BLAST search results, we have constructed a new data base of conserved sequence regions for protein sequences in the NCBI *Entrez* release 14 data base (v. 1.7; Epstein et al. 1994).

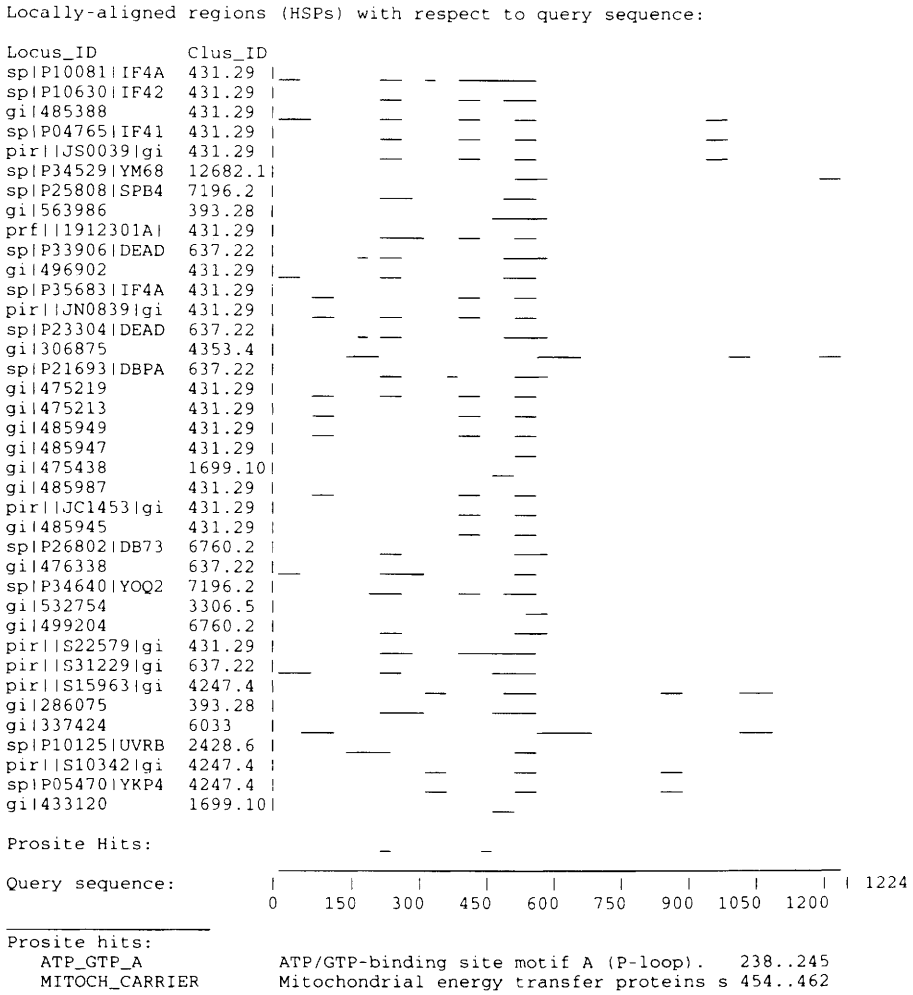


Figure 2 Summary diagram showing the locations of all of the locally aligned regions (high-scoring segment pairs) within the query sequence for each of the data base sequences identified in a search. The locations of PROSITE pattern matches within the query sequence are displayed in the last line of the list (labeled Prosite Hits). The title line for each PROSITE match is shown at the *bottom*; each line is hypertext linked to the PROSITE data base via the EMBL's SRS WWW server (URL: http://www.embl-heidelberg.de/srs/srsc). This example is from the BEAUTY/CRseq search used in Fig. 1; the output shown here has been shortened to conserve space. Two regions within the query sequence, beginning approximately at positions 230 and 510, respectively, are aligned in large number data base matches across a number of different families. The first of these regions corresponds to the position of a PROSITE pattern match to an ATP/GTP-binding site motif at position 238–245. The second PROSITE hit, a mitochondrial energy transfer protein motif, does not correspond to any of the commonly aligned regions and most likely represents a false-positive PROSITE match.

Construction of a Nonredundant Sequence Data Set

As the first step in constructing the Conserved Regions Data Base, a nonredundant set of sequences from the *Entrez* protein sequence data base was compiled. The *Entrez* data base (Epstein et al. 1994) is a redundant set of sequences generated from a union of a number of protein sequence data bases including GenPept, PIR (George et al. 1994), SWISS-PROT (Bairoch and Boeckmann 1992), PRF, and the NCBI Backbone. To reduce the redundancy in this set, we took advantage of the precomputed protein "neighbor" information available from the NCBI. For the *Entrez* data base, the NCBI matches each sequence against the entire data base using BLAST (Altschul et al. 1990, 1994) with the BLOSUM62 matrix and the expectation value set to 0.2. Both the query and the data base filtered for low-entropy regions using XNU (with parameters -n 4 -m 3; Claverie and States 1993) and SEG (with parameters 12 1.8 2.0; Wootton and Federhen 1993). All matches with BLAST *P* values (probabilities) ≤ 0.001 are saved for use in the *Entrez* data base search and retrieval system to identify all sequences related to any other sequence in the data base. The *Entrez* retrieval system uses these scores to automatically generate links between similar sequences. The NCBI has kindly provided us with the complete neighbor information for each *Entrez* release. To create a nonredundant sequence data set, all sequences with neighbor links having BLAST *P* values $\leq 10^{-100}$ were compared for exact string or substring matches using the perl *index* function (practical extraction and report language; Wall and Schwartz 1990). If one sequence was found to be an exact substring of another, the shorter sequence was excluded. If two sequences were identical and of the same length, then the sequence occurring first in the following order was kept: SWISS-PROT > PIR > GenPept > NCBI Backbone > PRF). This process excluded 73,365 redundant sequences from further consideration. Classes of sequences found in previous work to produce aberrant clustering and/or alignment (e.g., mutant, artificial, chimeric, and patent sequences) were identified by searching title lines for matches to a set of key words. This excluded an additional 9969 sequences.

Sequence Clustering

The remaining 123,220 *Entrez* protein sequences were clustered into families of related sequences. In previous studies

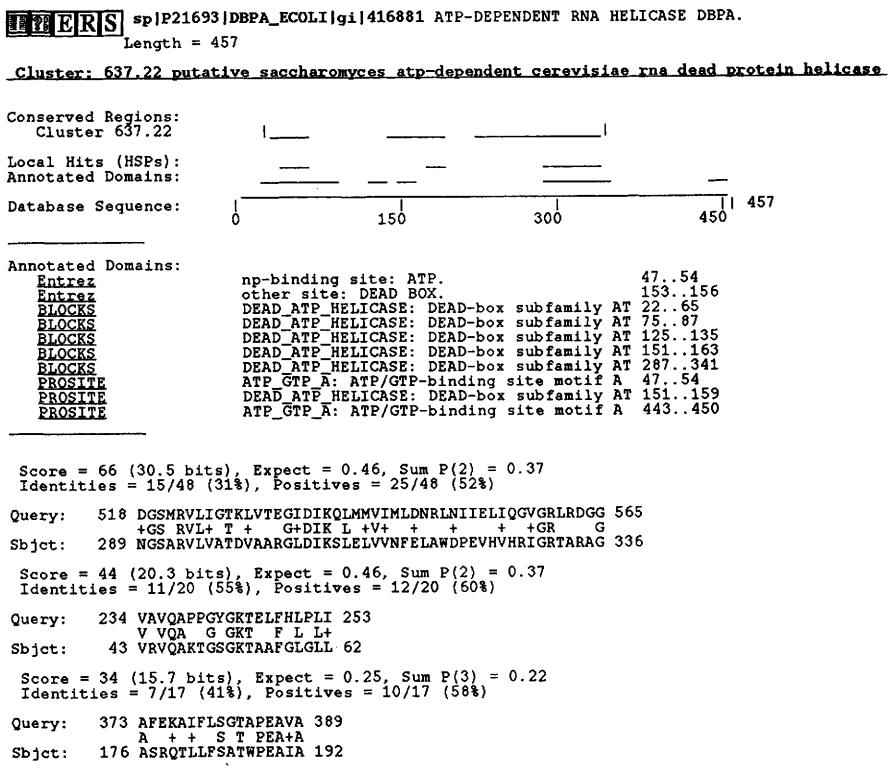


Figure 3 Alignment summary display and associated hypertext links generated for each matched data base sequence. (*Top*) To the *left* of the title line of the matched data base sequence are five hypertext-linked buttons. The first two buttons are generated by the NCBI's blast output parser program (Recipon 1995). Clicking on the first button jumps the user to the initial BLAST summary list of matched sequences. The second button retrieves a flat-file data base sequence report for the matched sequence. The three other buttons (E, R, and S) are generated by our WWW server. These buttons retrieve links to the *Entrez* data base (E), related sequences (R; *Entrez* protein neighbors), and the SRS data base (S) (see text). In the example given here (from the same BEAUTY search report used in Figs. 1 and 2), the S button retrieves an SRS report with hypertext links to cross-referenced information in the MEDLINE, EMBL, PIR, and PROSITE data bases. The consensus title line of the sequence family (Cluster_ID) containing the matched sequence, with a hypertext link to the mulitple alignment for that family (Fig. 4), is displayed below the buttons. The alignment summary diagram shows the relative positions of the conserved regions, the locally aligned regions (HSPs), and any annotated domains in the data base sequence. The vertical lines (|) bounding the conserved regions indicate the start and end positions of that region of the alignment scanned for conserved domains. These bounds are determined by the length of the shortest sequence in that cluster and thus are usually shorter than the sequence hit (as in this example). A comparison of the locations of the local hits (HSPs) relative to the locations of the conserved regions is therefore directly comparable only within these bounds. Below the schematic is a list of the source (*Entrez*, PROSITE, BLOCKS, etc.), title lines, and positions of each of the annotated domains and sites. Each of the annotation title lines is hypertext linked to the source data base record from which the annotation was derived. (*Bottom*) The standard BLAST alignments of the local hits between the query and the data base sequences are shown. In this example, the first locally aligned region (at position 43–62) corresponds to a conserved region in the matched data base sequence, as well as to a nucleotide-binding site annotated in *Entrez* (47–54) and matched by a PROSITE nucleotide-binding site pattern (47–54). This region also corresponds to the BLOCKS domain (22–65) common to all DEAD box subfamily members. In the query sequence, this region contains the same PROSITE pattern match (Fig. 2). Thus, this region of the query probably contains a nucleotide-binding site similar to that found in DEAD-box-containing ATP-dependent RNA helicases. The second locally aligned region (176–192) falls within a conserved region in this sequence family but does not correspond to a region that has been annotated in the data bases. The third locally aligned region (289–336) falls within the latter half of the carboxy-terminal conserved region, a region that corresponds closely to the last annotated BLOCKS domain (287–341) found in DEAD-box-containing helicases. The DEAD box itself (the *Entrez* annotated site at position 153–156), however, and two other annotated BLOCKS domains conserved in this subfamily of helicases (75–87, 125–135) are not found within the query sequence. This suggests that the query sequence may be a related helicase belonging to a different (and currently uncharacterized) family within the helicase superfamily (see the PROSITE helicase family description at URL: http://www.embl-heidelberg.de/srs/srsc?[prositedoc-id:PDOC00039]). This example demonstrates the usefulness of having easily accessible functional information from a variety of data base sources directly integrated into data base search results.

Cluster 637.22 Members

Consensus Title: putative saccharomyces atp-dependent cerevisiae rna dead protein helicase t26g10.1 ybr142w

External links: [D]=DB report; [E]=Entrez links; [R]=Related seqs; [S]=SRS report

[D][E][R][S] **129376:** splP26196lP54_HUMANlgil129376 -- **PUTATIVE ATP-DEPENDENT RNA HELICASE P54.**
[D][E][R][_] **145727:** gil145727 -- **deaD**
[D][E][R][S] **416881:** splP21693lDBPA_ECOLIlgil416881 -- **ATP-DEPENDENT RNA HELICASE DBPA.**
[D][E][R][S] **416915:** splP32892lDRS1_YEASTlgil416915 -- **PUTATIVE ATP-DEPENDENT RNA HELICASE DRS1.**
[D][E][R][S] **481209:** pirllS38329lgil481209 -- **gene Dbp45A protein - fruit fly (Drosophila melanogaster)**
[D][E][R][_] **509403:** gil509403 -- **BAT1**
[D][E][R][S] **539572:** pirllA47743lgil539572 -- **DEAD box protein RB - human**

Multiple Alignment of Cluster 637.22

```
...
            796         810 811         825 826         840 841         855
    1 637.22 XXXXXgggqXLJXdB SXXRG*DJXXiXXnX ;XXXiXXXXXXiHPX RPXnPXyXXGXe --  623
    2 145727 DGRLD----ILIATD VAARGLDVERISLVI NYDIPMDSESYVHRI GRTGRAGRAGRALLF  365
    3 416881 NGSAR----VLVATD VAARGLDIKSLELVI NFELAWDPEVHVHRI GRTARAGNSGLAISF  344
    4 416915 NLEVP----YLICTD LASRGLDIPRIETVI NYDMPFDYEIYLHRV GRTARAGREGRSVIF  552

    5 481209 SNQIR----TLIATD VAARGLDIPSVELVM NHMLPRTPREYIHRT GRTARAGRKGMSISI  354
    6 509403 QQFKDFQRGILVATN LFGRGMDIERVNIAF NYDMPEDSDTYLHRV ARAGRFGTKGLAITF  389
    7 129376 HDFRNGLCRNLVCTD LFTRGIDIQAVNVVI NFDFPKLAETYLHRI GRSGRFGHLGLAINL  437
    8 539572 ERFKKGDVRFLICTD VAARGIDIHGVPYVI NVTLPDEKQNYVHRI GRVGRAERMGLAISL  612

...
```

Figure 4 Cluster Multiple Alignment. The mutltiple alignment for a given cluster is displayed by clicking on one of the Clus_ID hypertext links in the BEAUTY search output (see Figs. 1–3). (*Top*) The list of the locus names and title lines of the cluster members. Links to the sequence report, *Entrez* data base, related *Entrez* sequences, and SRS reports are provided for each sequence. (*Bottom*) The PIMA multiple alignment of the sequences in the cluster. In the alignment, the sequences are labeled using the NCBI's GenInfo (gi) identifier. The first line of the alignment is the PIMA covering pattern, representing the conserved positions in the alignment (see Methods). This cluster has 23 members, and the example shown here has been shortened significantly to conserve space.

(Smith and Smith 1990, 1992) we clustered the PIR (George et al. 1994) or SWISS-PROT (Bairoch and Boeckmann 1992) data base into sequence families by performing all pairwise comparisons between all sequences in the data base using a high-speed sequence comparison program, such as BLAST (Altschul et al. 1990), then clustering the scores into family sets using a maximal linkage algorithm (Sneath and Sokal 1973). Here, we used the NCBI's precomputed protein neighbor information, described above, to cluster *Entrez* sequences into related sets. Maximal linkage was used to cluster into families the BLAST neighbor link scores from the 123,220 protein sequences using a BLAST score cutoff of $P = 10^{-3}$. This generated 12,669 families of 2 or more sequences per set (comprised of 97,521 sequences) with 25,699 other sequences (singletons) not clustered into any family.

Multiply Aligning Each Family Using PIMA

To identify the conserved regions within the sequences of each family, each sequence set was multiply aligned using PIMA, our pattern-induced multiple-sequence alignment program (Smith and Smith 1992). PIMA constructs a single regular expression-like "covering" pattern for each family during the multiple alignment (Smith and Smith 1990, 1992). The patterns are generated using a predefined set of amino acid classes (I. Ladunga, B. Wiese, and R.F. Smith, in prep.). The patterns represent the smallest amino acid class that includes (covers) the set of amino acids observed at each aligned column in the alignment (Smith and Smith 1990). Gapped positions in an alignment are converted into gap characters ("g") in patterns and each gap character can function as 0 or 1 amino acid of any type during subsequent pattern alignment.

If, during multiple alignment, the information content of a pattern constructed by the alignment of two subfamilies drops below a minimum threshold of 10-amino-acid equivalents (Smith and Smith 1990), then the alignment process is terminated. The multiple alignments of all subfamilies identified at this stage are then collected. This process generated a total of 13,368 multiple alignments from the original set of 12,669 families.

Scanning Alignments for Fragments

In PIMA alignments, the final ("root") covering pattern represents those sequence elements common to all members of the aligned set (Smith and Smith 1990, 1992). Because a root pattern can be no longer than the shortest sequence in the set, the presence of one or more sequence

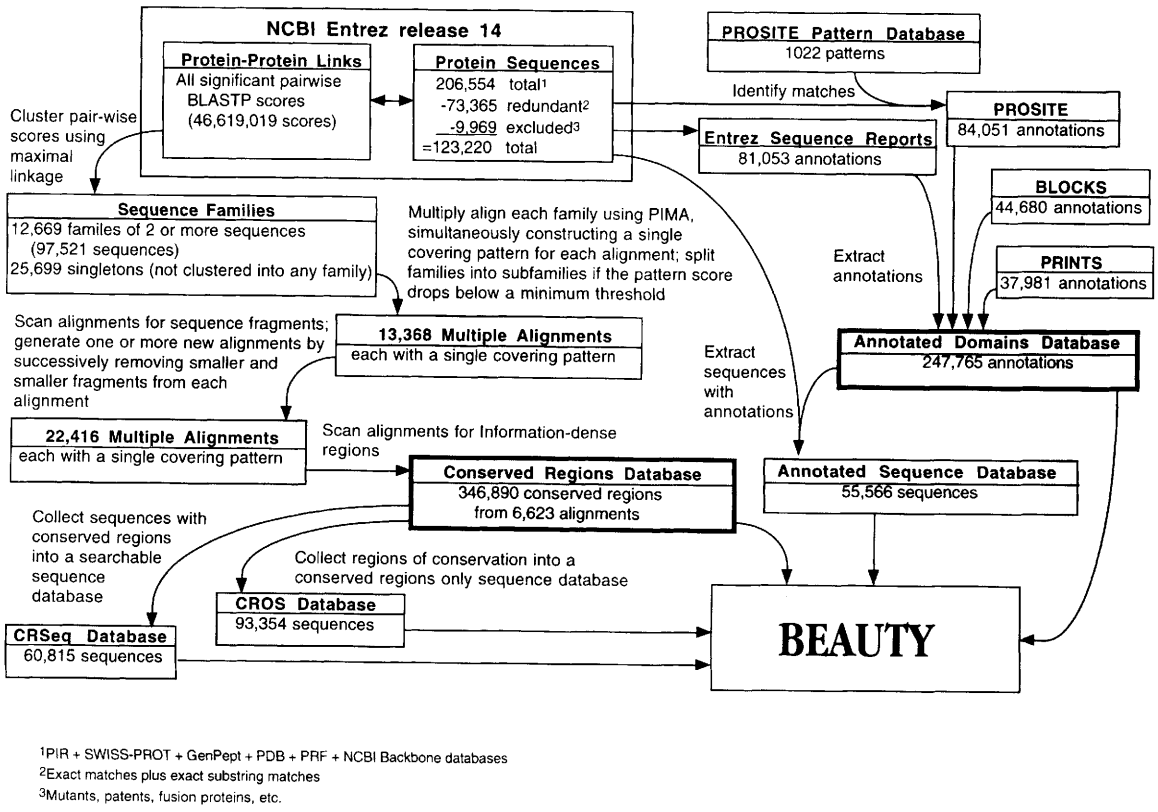


Figure 5 Overview of BEAUTY data base construction. The methods used to develop the data bases used by BEAUTY are summarized. The methods used to develop the Conserved Regions Data Base and the Annotated Domains Data Base are shown at *left* and *right*, respectively. BEAUTY directly searches any one of three sequence data bases (CRSeq, CROS, and the Annotated Sequence data bases) and incorporates information from the Conserved Regions and Annotated Domains data bases directly into the search results.

fragments in an alignment can significantly reduce the information content of the final patterns. However, we do not want to totally eliminate fragments from the families because a significant amount of information on sequence variability would be lost if that were done. Also, the definition of what constitutes a sequence fragment is often unclear. In some families there is a wide range of sequence lengths varying evenly from the largest to the smallest, so that any objective definition of a fragment would be totally arbitrary. Also, although some sequence fragments are labeled "fragment" or "partial peptide" in their data base entries, these sequences represent only a fraction of the fragments that we have identified in the *Entrez* data base.

To include sequence fragments in multiple alignments while at the same time preventing them from reducing alignment lengths, we have implemented the following procedure: The PIMA multiple-alignment method is based on a progressive pairwise alignment approach (Smith and Smith 1990, 1992). The order of pairwise alignments is based on the branching order of a dendrogram produced by a maximal linkage clustering algorithm. During alignment construction, an initial covering pattern is constructed from the alignment of the two most similar sequences in the dendrogram. Patterns are similarly constructed for each node in the dendrogram by moving down the tree, aligning at each node the patterns/sequences connected by the next most similar node. If, at any node, one of the two aligned patterns/sequences is a specified percentage shorter than the longer of the two inputs (default: 20%), then the longer pattern/sequence is saved to a separate file. The alignment/pattern construction process is then continued with the common (now shorter) pattern, as normal. Each saved alignment file is then added to the alignment data base. In this manner, sequence information from the longer sequences as well as from the fragments can be represented in the alignment data base.

We have also implemented the following modification of this method to retain and utilize information from sequence families with subfamily members containing heterologous domains (e.g., a subfamily with domains A and C grouped with a subfamily containing domains B and C, as may occur with different receptor protein kinase subfamilies sharing a common catalytic domain). If, at any node during pattern construction, the *pattern* resulting

from the alignment of the two input patterns/sequences is a specified percentage shorter than the *shortest* of the two inputs (default: 20%), then *both* input pattern/sequence files are saved, and the process continued. Thus, in the example above, information from heterologous domains A and B will be saved (in files with AC and BC patterns) and represented in the pattern data base as well as the final pattern resulting from all of the alignments of the homologous C domains.

Scanning the initial 13,368 PIMA alignments for fragments and heterologous domains generated an additional 10,762 alignments. The alignments were then scanned for those containing only identical sequences, and such sets were removed from the alignment data base (these alignments primarily contained very short identical sequences with BLAST probability (*P*) values greater than the 10^{-100} score cutoff used during the initial screen for redundant sequences). This eliminated 1714 alignments, leaving a final set of 22,416 alignments. Each of the alignments was then given a unique cluster identifier (Clus_ID); alignments generated by removing fragments were given a unique cluster identifier extension.

Identifying Conserved Regions Within Aligned Sequences

The 22,416 alignments were then scanned for information-dense regions using a program we have developed previously (Smith and Smith 1992). This program extracts all local regions of length n or longer within a multiple alignment that have an information density (ID) above a set threshold, T. To generate a list of conserved regions for each sequence in a multiple alignment, T was set to 1.5 times the average ID of the entire alignment. Alignments having no regions above this threshold (e.g., those composed entirely of nearly identical sequences) will therefore not contribute to this data set.

The 346,890 conserved regions generated with this procedure were used to construct a Conserved Regions Data Base. This data base contains four tables: Cluster, Cluster-Locus, Locus, and Domain. For each cluster identifier (Clus_ID), the Cluster table lists the number of sequences in the cluster and the consensus Cluster_Title. The consensus Cluster_Title is made up of the ten most common words used in each of the sequence title lines, in the rank order of most common occurrence. The Cluster-Locus table links each Clus_ID to each of the locus names of the sequences in that cluster. The Locus table lists the number of domains in the sequence for each locus name. Finally, the Domain table lists, for each sequence, the domain information (domain number, domain start position, and domain length) for each conserved region in the sequence.

Sequence Data Base Construction: The CROS and CRSeq Data Bases

Sequences represented in the Conserved Regions Data Base were used to create two new sequence data bases for use in BEAUTY searches (Table 2). The CROS (conserved regions only sequence) data base consists of the sequences from each alignment for which all nonconserved regions within each sequence have been converted to X's. This eliminates spurious matches to nonconserved regions during data base searching, reducing potential false-positive matches to nonrelated sequences. The 93,354 entries in the CROS data base were derived from 60,815 sequences (a sequence can contribute more than one entry in the CROS data base because sequences can be included in more than one alignment for those families containing one or more sequence fragments).

The CRSeq (conserved regions sequence) data base contains 60,815 full-length sequences (i.e., with both conserved and nonconserved regions represented) comprising all sequences that have at least one conserved region. Using the CRSeq data base, local BLAST hits are not restricted solely to the conserved regions but can occur anywhere within the data base sequence. Matches to a data base sequence in the CRSeq data base can thus have a potentially higher score than a match to the corresponding sequence in the CROS data base. Comparison of query sequences to both data bases is therefore recommended.

Annotated Domains Data Base Construction

A data base of annotated domains/sites was created by combining annotations from four sources, described below (Fig. 5). The following information was stored for each annotation: (1) the sequence's "gi" number (the GenInfo ID; NCBI's unique identifier for sequences across multiple data bases), (2) the source of the annotation (PROSITE, BLOCKS, etc.), (3) the annotation type (domain, site, etc.,), if applicable, (4) a comment/title line for the annotation, and (5) the starting position and length of the annotated region with respect to the first residue of the sequence. The four annotation sources were:

Table 2. Domain information presented in BEAUTY searches

	BEAUTY-searchable sequence data base		
	CRSeq	**CROS**	**Annotation**
Number of sequence entries	78,934	130,158	55,566
Conserved regions	yes	yes	no
Annotated domains	yes	yes	yes

1. Entrez *Sequence Reports*

Feature table entries within the sequence reports of each sequence in the *Entrez* data base (release 14; Epstein et al. 1994) were scanned for any annotations describing known domains and sites within each sequence. The gi number (GenInfo ID) was used as the unique identifier for each sequence. The feature table entries "region" or "site" provided the annotation type, whereas the "comment" and the location ("start" and "length") of the domain were taken directly from the corresponding feature table fields. This generated 81,053 annotations, originating primarily from PIR (George et al. 1994) and SWISS-PROT (Bairoch and Boeckmann 1992) sequence reports.

2. PROSITE

Each sequence in the *Entrez* data base was compared with the patterns in PROSITE (release 12.1; Bairoch 1992; containing 1022 patterns) using the Prosite program (version 2.1.4; K. Hartmuth and M.D. Zorn, unpub.) with a minimum pattern length of four. The start position and length of each match were stored. The comment field was taken from the "ID" (pattern name) and "DE" (pattern description) labeled lines in PROSITE data base. These matches yielded 84,051 annotations.

3. BLOCKS

All of the conserved sequence blocks were extracted from sequences represented in the BLOCKS data base (version 8.0; Henikoff and Henikoff 1991, 1992, 1993). Sequence locus names were converted to gi's using the accession search capability of *Entrez* (Epstein et al. 1994). The comment field was extracted from the ID and DE lines of BLOCKS data base. The length was taken from the corresponding "BL" line that contains the block length. This generated 44,680 BLOCKS annotations.

4. PRINTS

All domains designated as "final motifs" were extracted from sequences represented in the PRINTS protein fingerprint data base (version 6.0; Attwood and Beck 1994; Attwood et al. 1994). The "fd" line of the PRINTS data base contains the name of the protein motif, the sequence accession name for the sequence that contains the motif, and the start position of the motif in the sequence. For each fd line, the accession name was converted to a gi using the accession search capability of *Entrez* (Epstein et al. 1994) and the start position was extracted. The comment field was taken from of the "fc" (motif code) and "ft" (motif title) labeled lines listed for each motif, and the length was taken from the corresponding "fl" (motif length) line. This yielded 37,981 PRINTS annotations.

Constructing an Annotated Sequence Data Base

All of the nonredundant *Entrez* sequences containing at least 1 of the 247,765 annotated domains or sites were used to create a separate Annotated Sequence Data Base. This data base currently contains 55,566 sequences (Table 2; Fig. 5).

BEAUTY Program Implementation

BEAUTY is an enhanced version of the NCBI's BLASTP program (Altschul et al. 1990) created by adding three new routines to the BLASTP "C" code (v. 1.4.7; last modified 10/16/94). The first function adds a table of sequence family information to the BLAST output (Fig. 1). This function (print_domain_headers) requires the BLAST hit list pointer, the P value cutoff, and the query name. The second function draws a figure showing the locations of the locally aligned regions (HSPs, high-scoring segment pairs) with respect to the query sequence (Fig. 2). This function (draw_hits) requires the hit list pointer, the query sequence length, and the *P* value cutoff. The third function draws, for each data base sequence hit, a figure showing the relative positions of all conserved and annotated domains as well as the locally aligned regions in the data base sequence matched (Fig. 3). This function (draw_pictures) requires the hit list pointer, the query name, and the sequence length. All three functions are written in the C programming language, to be compatible with the original BLAST program. The three functions access the Conserved Regions and Annotated Domains data bases via calls to gdbm functions (GNU database management C library). The gdbm functions directly access binary-formatted hash files generated from the appropriate data base tables.

Updates and Availability

The CROS, CRSeq, and Annotated Domains data bases will be updated to correspond to the latest version of *Entrez* with every other *Entrez* release (approximately three times a year). The data base construction is largely automated, and adjustments will be made as necessary to accommodate format changes in *Entrez* and BLAST. Questions regarding the availability of the BEAUTY source code and the CROS, CRSeq, and Annotated Domains data bases should be addressed to R.F.S.

WWW Access

BEAUTY searches are available to the community via the "BCM Search Launcher" WWW pages (URL: http://gc.bcm.tmc.edu:8088/search-launcher/launcher.html; R.F. Smith, B.A. Wiese, M.K. Wojzynski, and K.C. Worley, in prep.). Users can input a query protein sequence into a WWW input form that launches a BEAUTY search on our local server. After the search is complete, the BEAUTY search results are passed through the NCBI's BLAST output filter (an awk script; Herve Recipon, 1995) to generate a HyperText Markup Language (HTML) version of the output. This filter also adds an embedded hypertext link for each data base sequence matched that allows a user to immediately view the data base sequence report for that sequence.

The output is then passed to a multifunctional perl (Wall and Schwartz 1990) script we have developed. This script first matches the query sequence against the PROSITE pattern data base (Bairoch 1992) using the Prosite

program described above. This generates a graphic that is incorporated into BEAUTY search output, showing the locations of any PROSITE pattern matches within the query sequence (Fig. 2). The script next adds several additional hypertext links for each data base sequence matched in a BEAUTY search (see Fig. 3). The "E" link retrieves a set of *Entrez* data base links to additional sequence-related information, such as the MEDLINE abstracts and GenBank reports. The "R" link retrieves a set of links to protein sequences related to the matched data base sequence (using the *Entrez* protein neighbor information). The "S" link retrieves the SRS (Etzold and Argos 1993) report for the sequence matched. SRS sequence reports include links to related information obtained from >30 multiple cross-referenced and linked data bases. The final output file is then displayed to the user as an HTML document that can be saved and viewed later with the links intact.

ACKNOWLEDGMENTS

We thank Drs. Dan Davison, Karen Kabnick, and Istvan Ladunga for critical reading of the manuscript. We also thank the NCBI for providing the *Entrez* neighbor link information used in these studies. This work was supported by a postdoctoral fellowship (1T15LM07093-02 awarded to K.C.W.) from the National Library of Medicine, National Institutes of Health; grants to the Baylor Human Genome Center (P30-HG00210) and R.F.S. (1R01-HG00973-01) from the National Center for Human Genome Research, National Institutes of Health; and the W.M. Keck Center for Computational Biology.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

References

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. Wootton. 1994. Issues in searching molecular sequence databases. *Nature Genet.* **6:** 119—129.

Attwood, T.K. and M.E. Beck. 1994. PRINTS—A protein motif fingerprint database. *Protein Eng.* **7:** 841–848.

Attwood, T.K., M.E. Beck, A.J. Bleasby, and D.J. Parry-Smith. 1994. PRINTS–A database of protein motif fingerprints. *Nucleic Acids Res.* **22:** 3590–3596.

Bairoch, A. 1992. PROSITE: A dictionary of sites and patterns in proteins. *Nucleic Acids Res.* **20:** 2013–2018.

Bairoch, A. and B. Boeckmann. 1992. The SWISS-PROT protein sequence database. *Nucleic Acids Res.* **20:** 2019–2022.

Benson, D., D.J. Lipman, and J. Ostell. 1993. Genbank. *Nucleic Acids Res.* **21:** 2963–2965.

Claverie, J.M. and D. States. 1993. Information enhancement methods for large scale sequence analysis. *Computers Chem.* **17:** 191–202.

Collins, F.S. 1995. Positional cloning moves from perditional to traditional. *Nature Genet.* **9**: 347–350.

Epstein, J.A., J.A. Kans, and G.D. Schuler. 1994. "WWW *Entrez*: A hypertext retrieval tool for molecular biology." Electronic Proceedings of the Second World Wide Web Conference '94: Mosaic and the Web.

Etzold, T. and P. Argos. 1993. SRS an indexing and retrieval tool for flat file data libraries. *Comput. Appl. Biosci.* **9:** 49–57.

George, D.G., W.C. Barker, H.W. Mewes, F. Pfeiffer, and A. Tsugita. 1994. The PIR-international protein sequence database. *Nucleic Acids Res.* **22:** 3569–3573.

Henikoff, S. and J.G. Henikoff. 1991. Automated assembly of protein blocks for database searching. *Nucleic Acids Res.* **19:** 6565–6572.

———. 1992. Amino acid substitution matrices form protein blocks. *Proc. Natl. Acad. Sci.* **89:** 10915–10919.

———. 1993. Performance evaluation of amino acid substitution matrices. *Proteins* **17:** 49–61.

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

Pearson, W.R. 1990. Rapid and sensitive sequence comparisons with FASTP and FASTA. *Methods Enzymol.* **183:** 63–98.

Pearson, W.R. and D.J. Lipman. 1988. Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci.* **85:** 2444–2448.

Pongor, S., Z. Hatsagi, K. Degtyarenko, P. Fabian, V. Skerl, H. Hegyi, J. Murvai, and V. Bevilacqua. 1994. The SBASE protein domain library, release 3.0: A collection of annotated protein sequence segments. *Nucleic Acids Res.* **22:** 3610–3615.

Recipon, H. 1995. (URL: ftp://ncbi.nlm.nih.gov/pub/recipon/BLAST_Notebook) National Center for Biotechnology Information, National Library of Medicine.

Smith, R.F. and T.F. Smith. 1990. Automatic generation of primary sequence patterns from sets of related protein sequences. *Proc. Natl. Acad. Sci.* **87:** 118–122.

———. 1992. Pattern-induced multi-sequence alignment (PIMA) algorithm employing secondary structure-

dependent gap penalties for use in comparative protein modeling. *Protein Eng.* **5:** 35–41.

Sneath, P.H. and R.R. Sokal. 1973. *Numerical taxonomy.* Freeman, San Francisco, CA.

Wall, L. and R.L. Schwartz. 1990. *Programming perl.* O'Reilly and Associates, Inc., Sebastopol, CA.

Wilbur, W.J. and D.J. Lipman. 1983. Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci.* **80:** 726–730.

Wooton, J.C. and S. Federhen. 1993. Statistics of local complexity in amino acid sequences and sequence databases. *Computers Chem.* **17:** 149–164.

Received June 19, 1995; accepted in revised form August 15, 1995.