BCM Search Launcher—An Integrated Interface to Molecular Biology Data Base Search and Analysis Services Available on the World Wide Web

Randall F. Smith,[1,2,4] Brent A. Wiese,[1] Mary K. Wojzynski,[1] Daniel B. Davison,[2,3] and Kim C. Worley[1]

[1]Human Genome Center, Department of Molecular and Human Genetics, and [2]Department of Cell Biology, Baylor College of Medicine, Houston, Texas 77030; [3]Department of Biochemical and Biophysical Sciences, University of Houston, Houston, Texas 77004

The BCM Search Launcher is an integrated set of World Wide Web (WWW) pages that organize molecular biology-related search and analysis services available on the WWW by function, and provide a single point of entry for related searches. The Protein Sequence Search Page, for example, provides a single sequence entry form for submitting sequences to WWW servers that offer remote access to a variety of different protein sequence search tools, including BLAST, FASTA, Smith-Waterman, BEAUTY, PROSITE, and BLOCKS searches. Other Launch pages provide access to (1) nucleic acid sequence searches, (2) multiple and pair-wise sequence alignments, (3) gene feature searches, (4) protein secondary structure prediction, and (5) miscellaneous sequence utilities (e.g., six-frame translation). The BCM Search Launcher also provides a mechanism to extend the utility of other WWW services by adding supplementary hypertext links to results returned by remote servers. For example, links to the NCBI's *Entrez* data base and to the Sequence Retrieval System (SRS) are added to search results returned by the NCBI's WWW BLAST server. These links provide easy access to auxiliary information, such as Medline abstracts, that can be extremely helpful when analyzing BLAST data base hits. For new or infrequent users of sequence data base search tools, we have preset the default search parameters to provide the most informative first-pass sequence analysis possible. We have also developed a batch client interface for Unix and Macintosh computers that allows multiple input sequences to be searched automatically as a background task, with the results returned as individual HTML documents directly to the user's system. The BCM Search Launcher and batch client are available on the WWW at URL http://gc.bcm.tmc.edu:8088/search-launcher.html.

DNA and protein sequence analysis services are now available from a large number of sources on the Internet, including Email, Gopher, and World Wide Web (WWW) servers (for review, see Boguski 1994; Harper 1994). These services are extremely useful for molecular biologists, as they allow access to the ever-expanding sequence data bases without requiring copious local data base storage, frequent data base updates, the cost of expensive and sophisticated hardware and software, and the cost and effort of continuous system maintenance.

The development of the WWW, a hypertext–multimedia communications system on the Internet, has been particularly useful in making international biological resources readily available and more easy to use (Jacobson 1994; Harper 1995; Schatz and Hardin 1994). Analysis results returned by WWW servers are easy to view (one can move forward and backward through a document using slider bars), and a hypertext document can include any number of hypertext links, allowing direct access to a variety of additional information sources on the Internet (e.g., related documents and other services).

Although comprehensive lists of molecular biology-related search and analysis resources currently available on the WWW have been compiled (e.g., Coutinho 1995; Robison 1995; Smith

[4]**Corresponding author.**
E-MAIL rsmith@bcm.tmc.edu; FAX (713) 798-5386.

6:454–462 ©1996 by Cold Spring Harbor Laboratory Press ISSN 1054-9803/96 $5.00

1995), the fact that individual WWW server sites are scattered throughout the Web hinders their effecient use. To perform more than one type of search or analysis on a particular sequence, for example, a user must traverse the Web to each individual WWW server, cut and paste one's sequence into each page, choose from an often large and incomprehensible list of search options, submit the search, wait for the results to be returned, store the results, then traverse the Web to the next site, repeating the process, and so forth.

The BCM Search Launcher addresses these limitations by providing an improved interface to molecular biology-related search and analysis services on the WWW. The most useful sequence analysis tools are collected and organized by search type within individual Search Launcher pages. All of the tools collected here are well maintained, and several have daily updates. The BCM Search Launcher pages take advantage of WWW data structures and hypertext links to provide auxiliary information from various sources [such as the National Center for Biotechnology Information's (NCBI) *Entrez* data base; Epstein et al. 1994] directly to the search results. The results can be saved and viewed later with the links to additional information intact. Using the Batch Client interface to the BCM Search Launcher, multiple searches can be performed at once, with the individual results stored in separate files on the user's system without further manipulation. These features combine to simplify access to and improve the utility of sequence analysis resources on the WWW.

RESULTS AND DISCUSSION

The format of the BCM Search Launcher pages addresses several of the drawbacks of the dispersed nature of the large number of molecular biology search services currently available on the WWW. The Search Launcher organizes searches by type and provides a single input form for related searches (protein sequence searches, nucleic acid searches, etc.; see Table 1). This arrangement allows the researcher to quickly identify the most useful tools for a specific type of search and allows multiple searches to be initiated from a single entry page. On each page a short program description is provided for each service (see Fig. 1) so that the novice user can quickly locate the appropriate analysis tools(s). Three links are also provided ([H][O][P]) that can be used to display help pages, full options pages, and default search parameters, respectively. In addition, as the sequence input field and the **Perform search** button are displayed at the top of each page, performing a search does not require scrolling down the page to paste in the sequence, as is common to many sequence search forms on the WWW. This is a very time-saving feature when performing several different searches for a given query sequence.

One of the features that makes the BCM Search Launcher so useful for novice and occasional users is that we have preset the default parameters for the search tools. This has two beneficial effects. First, a novice user will not have to choose among options that may or may not be appropriate for a given search type. For example, a researcher with a nucleic acid sequence query will not need to know which options and data bases are appropriate for nucleic acid versus protein searches. Second, the parameters have been set to optimize their utility for typical searches. For instance, for NCBI BLASTP protein data base searches (Altschul et al. 1990; Recipon and Gish 1995), we have set XNU (Claverie and States 1993) and SEG (Wooton and Federhen 1993) prefiltering to "on" (no filtering is the default). These filters mask "low complexity" regions, such as poly-Q runs, within query sequences, eliminating high-scoring hits to these regions that can mask more informative data base hits (see Altschul et al. 1994). For the open reading frame (ORF) identification tool provided by the Virtual Genome Center, University of Minnesota School of Medicine (Scherer 1995), we provide two different sets of default parameters, one for typical eukaryotic ORFs and one for typical bacterial ORFs (with a methionine start codon). The availability of preset parameters reduces the time needed for an inexperienced or infrequent user to obtain useful search results. A user with a special search requirement can always change these default parameters using the [O] link.

Another very useful feature provided by the Search Launcher is our ability to postprocess search results to add additional WWW links to a Hyper Text Mark up Language (HTML) document before displaying it to a user. Currently, for example, we add *Entrez* and Sequence Retrieval System (SRS) links for each data base sequence matched in a NCBI BLAST search (Fig. 2). Links to the NCBI's WWW *Entrez* interface allow the corresponding Medline literature abstracts to be easily retrieved and browsed for more detailed infor-

Table 1. Current BCM Search Launcher Services (listed by page)

- Protein sequence and pattern searches
  - BLASTP / NCBI non-redundant protein DB[a]
  - TBLASTN / NCBI non-redundant dna DB [a]
  - TBLASTN / dbest DB [a]
  - BEAUTY / CRSeqAnnot DB [b]
  - BEAUTY / CRSeq DB [b]
  - BEAUTY / CROS DB [b]
  - FASTA-SWAP / EC Pattern DB [c]
  - FASTA-PAT / PIMA Pattern DB [c]
  - PPSEARCH / PROSITE Pattern DB [d]
  - BLOCKS DB Search [e]
  - FASTA / SwissProt DB [f]
  - BLASTP / SBASE DB (Email[g]) [h]
  - BLITZ / SwissProt DB (Email[g]) [i]
  - SSEARCH / SwissProt DB (Email[g]) [j]
- Nucleic acid sequence searches
  - BLASTN / NCBI non-redundant dna DB [a]
  - BLASTN / dbEST DB [a]
  - BLASTX / NCBI non-redundant protein DB[a]
  - TBLASTX / dbEST DB [a]
  - BEAUTY-X / CRSeqAnnot DB [k]
  - BEAUTY-X / CRSeq DB [k]
  - BEAUTY-X / CROS DB [k]
- Multiple sequence alignments
  - ClustalW (protein/DNA)[l]
  - MAP (protein/DNA) [m]
  - PIMA (protein sequences only) [n]
  - BLOCK MAKER (protein sequences) [o]
- Pairwise sequence alignments
  - ALIGN (Optimal global alignments) [p]
  - LALIGN (N-best local alignments) [q]
  - LFASTA (Local similarity searches) [r]
- Gene feature searches
  - GRAIL (protein coding potential) [s]
  - FGENEH (exon assembly) [t]
  - FEXH (potential coding exons) [t]
  - HEXON (potential internal exons) [t]
  - HSPL (potential splice sites) [t]
  - RNASPL (exon-exon junctions) [t]
  - CDSB (*E.coli* protein coding regions) [t]
  - HBR (human and *E.coli* sequences) [t]
  - POLYAH (3' cleavage and poly-A) [t]
  - TSSG (human PolII promotor) [t]
  - TSSW (human PolII promotor) [t]
  - ORF Identification (eukaryotic) [u]
  - ORF Identification (prokaryotic) [u]
- Sequence utilities
  - ReadSeq (converts sequence format) [v]
  - RepMask (masks DNA repeats) [w]
  - Primer Selection [x]
  - 6 Frame Translation [y]
  - Reverse Complement [z]
  - WebCutter (restriction map) [0]
- Protein secondary structure prediction
  - Coils (coiled coil regions) [1]
  - nnPredict (using a neural network) [2]
  - PSSP / SSP (segment-oriented) [3]
  - PSSP / NNSSP (nearest-neighbor) [3]
  - TMpred (transmembrane regions) [4]
  - PHDsec (profile network) (Email[g]) [5]
  - PSA (globular proteins) (Email[g]) [6]
  - SOPM (Self optimized) (Email[g]) [7]
  - SSPRED (aa exchange stats.) (Email[g]) [8]
  - Swiss-Model (crystal structure) (Email[g]) [9]

[a] http://www.ncbi.nlm.nih.gov/Recipon/bs_seq.html

[b] http://dot.imgen.bcm.tmc.edu:9331/seq-search/Options/beautyP.html

[c] http://dot.imgen.bcm.tmc.edu:9331/seq-search/Options/fastapat.html

[d] http://www.ebi.ac.uk/searches/prosite_input.html

[e] http://blocks.fhcrc.org/blocks_search.html

[f] http://www.gdb.org/Dan/gq/gq.form.html

[g] Search results returned via Email not in HTML format

[h] http://base.icgeb.trieste.it/sbase/blast.html

[i] http://www.ebi.ac.uk/searches/blitz_input.html

[j] http://genome.eerie.fr/fasta/ssearch-query.html

[k] http://dot.imgen.bcm.tmc.edu:9331/seq-search/Options/beautyN.html

[l] http://dot.imgen.bcm.tmc.edu:9331/multi-align/Options/clustalw.html

[m] http://dot.imgen.bcm.tmc.edu:9331/multi-align/Options/map.html

[n] http://dot.imgen.bcm.tmc.edu:9331/multi-align/Options/pima.html

[o] http://www.blocks.fhcrc.org/blockmkr/make_blocks.html

[p] http://genome.eerie.fr/fasta/align-query.html

[q] http://genome.eerie.fr/fasta/lalign-query.html

[r] http://genome.eerie.fr/fasta/lfasta-query.html

[s] http://avalon.epm.ornl.gov/Grail-bin/EmptyGrailForm

[t] http://dot.imgen.bcm.tmc.edu:9331/gene-finder/gf.html

[u] http://alces.med.umn.edu/rawtrans.html

[v] http://dot.imgen.bcm.tmc.edu:9331/seq-util/Options/readseq.html

[w] http://www.tigem.it/LOCAL/SEQUTILS/repeat..html

[x] http://alces.med.umn.edu/rawprimer.html

[y] http://dot.imgen.bcm.tmc.edu:9331/seq-util/Options/sixframe.html

[z] http://dot.imgen.bcm.tmc.edu:9331/seq-util/Options/revcomp.html

[0] http://firstmarket.com/firstmarket/cutter

[1] http://ulrec3.unil.ch/software/COILS_form.html

[2] http://www.cmpharm.ucsf.edu/~nomi/nnpredict.html

[3] http://dot.imgen.bcm.tmc.edu:9331/pssprediction/pssp.html

[4] http://ulrec3.unil.ch/software/TMPRED_form.html

[5] http://www.embl-heidelberg.de/predictprotein/phd_pred.html

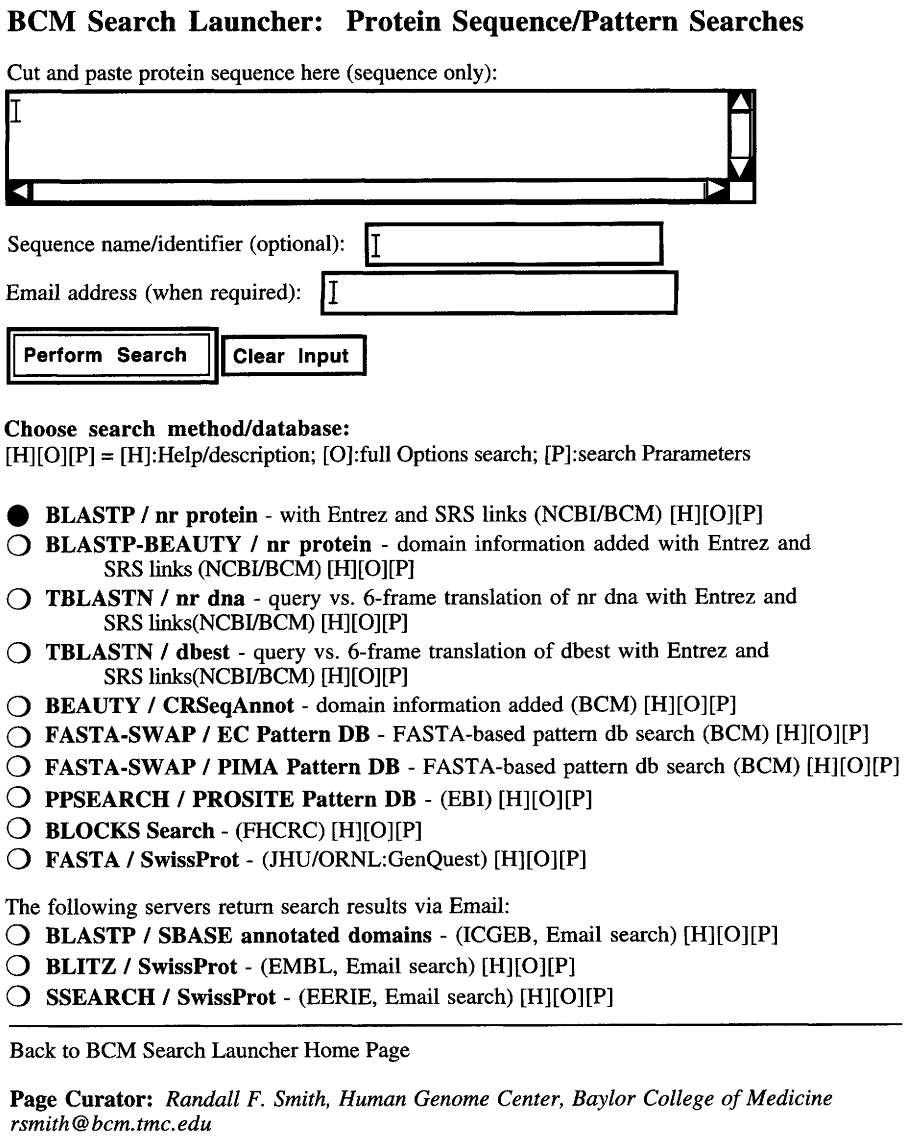
[6] http://bmerc-www.bu.edu/psa/request.htm

[7] http://www.ibcp.fr/predict.html

[8] http://www.embl-heidelberg.de/sspred/ssp_sin.html

[9] http://expasy.hcuge.ch/swissmod/PROMODSERV.html

mation on matched sequences. Links to sequences in the *Entrez* data base that are similar to a sequence matched in a data base search (*Entrez* "neighbors"; Boguski 1992; Epstein et al. 1994) make it much more easy to identify the possible function of a matched sequence in cases where the title line of a data base sequence is uninformative (e.g., a data base hit identified only as



BCM Search Launcher: Protein Sequence/Pattern Searches

Cut and paste protein sequence here (sequence only):

Sequence name/identifier (optional):

Email address (when required):

Choose search method/database:
[H][O][P] = [H]:Help/description; [O]:full Options search; [P]:search Prarameters

- ● **BLASTP / nr protein** - with Entrez and SRS links (NCBI/BCM) [H][O][P]
- ○ **BLASTP-BEAUTY / nr protein** - domain information added with Entrez and SRS links (NCBI/BCM) [H][O][P]
- ○ **TBLASTN / nr dna** - query vs. 6-frame translation of nr dna with Entrez and SRS links(NCBI/BCM) [H][O][P]
- ○ **TBLASTN / dbest** - query vs. 6-frame translation of dbest with Entrez and SRS links(NCBI/BCM) [H][O][P]
- ○ **BEAUTY / CRSeqAnnot** - domain information added (BCM) [H][O][P]
- ○ **FASTA-SWAP / EC Pattern DB** - FASTA-based pattern db search (BCM) [H][O][P]
- ○ **FASTA-SWAP / PIMA Pattern DB** - FASTA-based pattern db search (BCM) [H][O][P]
- ○ **PPSEARCH / PROSITE Pattern DB** - (EBI) [H][O][P]
- ○ **BLOCKS Search** - (FHCRC) [H][O][P]
- ○ **FASTA / SwissProt** - (JHU/ORNL:GenQuest) [H][O][P]

The following servers return search results via Email:

- ○ **BLASTP / SBASE annotated domains** - (ICGEB, Email search) [H][O][P]
- ○ **BLITZ / SwissProt** - (EMBL, Email search) [H][O][P]
- ○ **SSEARCH / SwissProt** - (EERIE, Email search) [H][O][P]

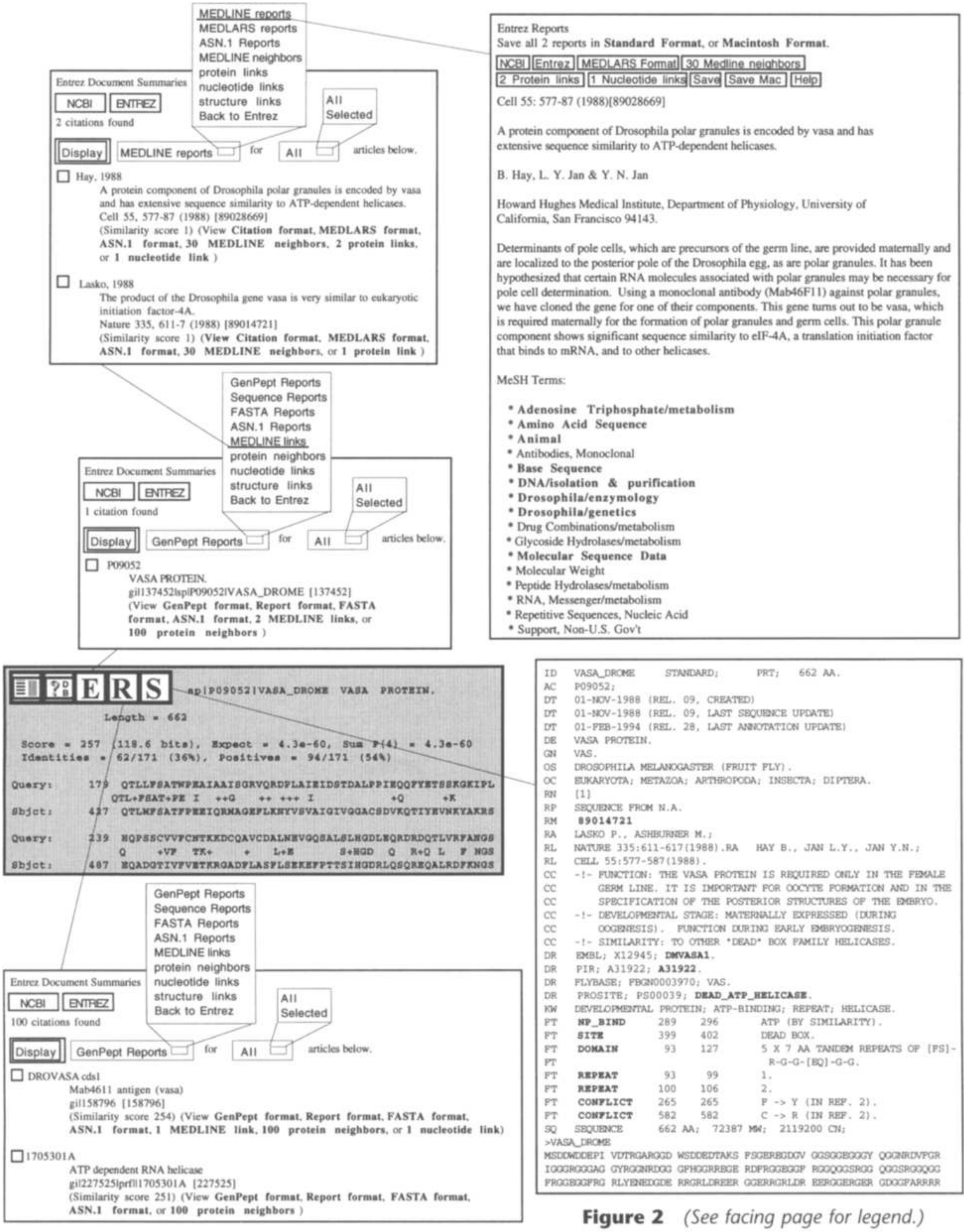
Back to BCM Search Launcher Home Page

Page Curator: *Randall F. Smith, Human Genome Center, Baylor College of Medicine rsmith@bcm.tmc.edu*

Figure 1 BCM Search Launcher: Protein Sequence Search WWW Page (URL http://dot.imgen.bcm.tmc.edu:9331/seq-search/protein-search.html). This page organizes WWW-accessible sequence and pattern data base search tools that can be used to analyze a user's protein sequence. At the top of the page are text-entry boxes into which a user can enter a protein sequence query, an optional sequence name/identifier, and Email return address (when required). The search to be performed is selected by choosing one of the selection buttons (*left*). A BLASTP search has been selected in this example (the highlighted button). For each of the search tools listed, there are links to Help or description pages ([H]), links to the full-Options search pages ([O]), and links to pages listing the parameters used in the search ([P]). The search is launched using the **Perform search** button. The three programs listed at the bottom return search results via Email rather than as a HTML page displayed on the browser.

ORF 186 says little about the function of that gene; browsing the list of related sequences and noting from their title lines that most are members of a known enzyme family, for example, is quite informative). Links to the SRS WWW interface [Etzold and Argos 1993; see Uniform Resource Locator (URL) http://www.embl-heidelberg.de/srs/srsc] provides information obtained from >30 cross-referenced data bases (e.g., EMBL, GenBank, SWISS-PROT, PIR, PDB, PROSITE, OMIM) to be accessed easily for matched sequence. Using the BCM Search Launcher WWW interface, all of this information is now linked directly into the search results and can be browsed immediately without the distraction of performing separate keyword searches for each matched sequence with each of these on-line resources. As a result, NCBI BLAST data base searches submitted via the Search Launcher can take significantly less time to analyze for both the novice and expert.

For those users who wish to perform a particular search on a number of sequence at a time, we have developed a Unix and Macintosh batch interface to the BCM Search Launcher WWW page. The batch client will (1) automatically read multiple sequences from one or more files (no need to paste the sequences one at a time), (2) run the searches in the background (so that one can log off if running on a public terminal), and (3) store the HTML result files directly in a user's Mac or Unix directory (no need to do inividual saves). The Results files can then be browsed at any later date using a standard WWW viewer such as NCSA Mosaic or Netscape. In all other respects



Entrez Document Summaries

2 citations found

□ Hay, 1988
A protein component of Drosophila polar granules is encoded by vasa and has extensive sequence similarity to ATP-dependent helicases.
Cell 55, 577-87 (1988) [89028669]
(Similarity score 1) (View **Citation format**, **MEDLARS format**, **ASN.1 format**, **30 MEDLINE neighbors**, **2 protein links**, or **1 nucleotide link**)

□ Lasko, 1988
The product of the Drosophila gene vasa is very similar to eukaryotic initiation factor-4A.
Nature 335, 611-7 (1988) [89014721]
(Similarity score 1) (**View Citation format**, **MEDLARS format**, **ASN.1 format**, **30 MEDLINE neighbors**, or **1 protein link**)

Entrez Reports
Save all 2 reports in **Standard Format**, or **Macintosh Format**.

Cell 55: 577-87 (1988)[89028669]

A protein component of Drosophila polar granules is encoded by vasa and has extensive sequence similarity to ATP-dependent helicases.

B. Hay, L. Y. Jan & Y. N. Jan

Howard Hughes Medical Institute, Department of Physiology, University of California, San Francisco 94143.

Determinants of pole cells, which are precursors of the germ line, are provided maternally and are localized to the posterior pole of the Drosophila egg, as are polar granules. It has been hypothesized that certain RNA molecules associated with polar granules may be necessary for pole cell determination. Using a monoclonal antibody (Mab46F11) against polar granules, we have cloned the gene for one of their components. This gene turns out to be vasa, which is required maternally for the formation of polar granules and germ cells. This polar granule component shows significant sequence similarity to eIF-4A, a translation initiation factor that binds to mRNA, and to other helicases.

MeSH Terms:

- **Adenosine Triphosphate/metabolism**
- **Amino Acid Sequence**
- **Animal**
- Antibodies, Monoclonal
- **Base Sequence**
- **DNA/isolation & purification**
- **Drosophila/enzymology**
- **Drosophila/genetics**
- Drug Combinations/metabolism
- Glycoside Hydrolases/metabolism
- **Molecular Sequence Data**
- Molecular Weight
- Peptide Hydrolases/metabolism
- RNA, Messenger/metabolism
- Repetitive Sequences, Nucleic Acid
- Support, Non-U.S. Gov't

Entrez Document Summaries

1 citation found

□ P09052
VASA PROTEIN.
gi|137452|sp|P09052|VASA_DROME [137452]
(View **GenPept format**, **Report format**, **FASTA format**, **ASN.1 format**, **2 MEDLINE links**, or **100 protein neighbors**)

```
sp|P09052|VASA_DROME  VASA  PROTEIN.

          Length = 662

 Score = 257 (118.6 bits), Expect = 4.3e-60, Sum P(4) = 4.3e-60
 Identities = 62/171 (36%), Positives = 94/171 (54%)

Query:     179  QTLLFSATWPEAIAAISGRVQRDPLAIEIDSTDALPPIEQQFYETSSKGKIPL
                QTL+FSAT+PE I   ++Q    ++ +++ I             +Q      +K
Sbjct:     427  QTLMFSATFPEEIQRMAGEFLKNYVSVAIGIVGGACSDVKQTIYEVNKYAKRS

Query:     239  EQPSSCVVFCNTKKDCQAVCDALNEVGQSALSLHGDLEQRDRDQTLVRFANGS
                 Q     +VF   TK+      +      L+E          S+HGD   Q   R+Q L    F NGS
Sbjct:     487  EQADGTIVFVETKRGADFLASFLSEKEFPTTSIHGDRLQSQREQALRDFKNGS
```

```
ID   VASA_DROME     STANDARD;      PRT;   662 AA.
AC   P09052;
DT   01-NOV-1988 (REL. 09, CREATED)
DT   01-NOV-1988 (REL. 09, LAST SEQUENCE UPDATE)
DT   01-FEB-1994 (REL. 28, LAST ANNOTATION UPDATE)
DE   VASA PROTEIN.
GN   VAS.
OS   DROSOPHILA MELANOGASTER (FRUIT FLY).
OC   EUKARYOTA; METAZOA; ARTHROPODA; INSECTA; DIPTERA.
RN   [1]
RP   SEQUENCE FROM N.A.
RM    89014721
RA   LASKO P., ASHBURNER M.;
RL   NATURE 335:611-617(1988).RA   HAY B., JAN L.Y., JAN Y.N.;
RL   CELL 55:577-587(1988).
CC   -!- FUNCTION: THE VASA PROTEIN IS REQUIRED ONLY IN THE FEMALE
CC       GERM LINE. IT IS IMPORTANT FOR OOCYTE FORMATION AND IN THE
CC       SPECIFICATION OF THE POSTERIOR STRUCTURES OF THE EMBRYO.
CC   -!- DEVELOPMENTAL STAGE: MATERNALLY EXPRESSED (DURING
CC       OOGENESIS).  FUNCTION DURING EARLY EMBRYOGENESIS.
CC   -!- SIMILARITY: TO OTHER *DEAD* BOX FAMILY HELICASES.
DR    EMBL; X12945; DMVASA1.
DR    PIR; A31922; A31922.
DR    FLYBASE; FBGN0003970; VAS.
DR    PROSITE; PS00039; DEAD_ATP_HELICASE.
KW   DEVELOPMENTAL PROTEIN; ATP-BINDING; REPEAT; HELICASE.
FT    NP_BIND      289    296        ATP (BY SIMILARITY).
FT    SITE         399    402        DEAD BOX.
FT    DOMAIN        93    127        5 X 7 AA TANDEM REPEATS OF [FS]-
FT                                    R-G-G-[EQ]-G-G.
FT    REPEAT        93     99        1.
FT    REPEAT       100    106        2.
FT    CONFLICT     265    265        F -> Y (IN REF. 2).
FT    CONFLICT     582    582        C -> R (IN REF. 2).
SQ    SEQUENCE     662 AA;  72387 MW;  2119200 CN;
>VASA_DROME
MSDDWDDEPI VDTRGARGGD WSDDEDTAKS FSGERBGDGV GGSGGBGGGY QGGNRDVFGR
IGGGRGGGAG GYRGGNRDGG GPHGGRREGE RDFRGGEGGF RGGQGGSRGG QGGSRGGQGG
FRGGEGGFRG RLYENEDGDE RRGRLDREER GGERRGRLDR EERGGERGER GDGGFARRRR
```

Entrez Document Summaries

100 citations found

□ DROVASA cds1
Mab4611 antigen (vasa)
gi|158796 [158796]
(Similarity score 254) (View **GenPept format**, **Report format**, **FASTA format**, **ASN.1 format**, **1 MEDLINE link**, **100 protein neighbors**, or **1 nucleotide link**)

□ 1705301A
ATP dependent RNA helicase
gi|227525|prf||1705301A [227525]
(Similarity score 251) (View **GenPept format**, **Report format**, **FASTA format**, **ASN.1 format**, or **100 protein neighbors**)

Figure 2 *(See facing page for legend.)*

the search results will be the same as if the search were performed directly from the BCM Search Launcher WWW Page. The batch client thus provides all of the advantages of the Search Launcher with the convenience of batch submission and background operation, greatly simplifying and expediting the search process.

Our batch client also offers several advantages over Email-based batch clients for sequence data base searching, such as the MAILFASTA (deBoer 1993; see Harper 1994) and MSU (Fuchs 1994) programs. Since the Search Launcher is WWW-based, search results are returned as HTML documents with embedded hypertext links to additional information resourses, rather than as simple text files, as returned by Email servers. Second, search results are deposited directly into one's directory, and the results are saved with the name of the query and the search in the name of the results. This obviates the time-consuming and occasionally confusing step of sorting and saving individual results out of one's Email system. Finally, the list of available search and analysis methods is always current, as it is obtained automatically from the BCM Server each time the batch client is run.

In summary, the BCM Search Launcher and batch client offer an enhanced, integrated, easy-to-use, and time-saving interface to the large number of extremely useful molecular biology data base search and analysis services available on the WWW. Direct access to external analysis tools via the WWW is revolutionizing the way in which researchers can analyze their data, and the Search Launcher, by organizing and improving the access to these tools, should prove extremely valuable to genome researchers worldwide.

METHODS

Underlying Technology

Reviews of the development of the WWW have been published recently (Schatz and Hardin 1995; McKinney et al. 1995), and the use of the Internet for sequence data base searching has also been discussed (Boguski 1994; Harper 1994; Brenner 1995a,b). Briefly, the Internet allows electronic communication between remote sites and promotes distributed information space, whereas the WWW allows communications in a hypermedia format without restriction to physical location.

The impetus for the BCM Search Launcher was the desire to be able to fill in a single HTML form on our Web page, launch a search at a remote site, return the results for further processing by our server, and present the results to the user. The development of these searching capabilities requires a mechanism to (1) choose a search server, (2) accept an appropriate query input from the user, for example, a sequence to be searched with specified search

Figure 2 Links to supplementary information added to NCBI BLAST search results by the BCM Search Launcher server. The shaded box shows an example of a BLASTP data base hit. The five icons to the left of the name of the matched data base sequence (sp|P095052|VASA_DROME VASA PROTEIN) are hypertext-linked buttons to related information. The first two buttons are provided by the NCBI as part of the standard search results returned by the NCBI BLAST WWW server (Recipon and Gish 1995); the first returns the reader to the summary table of BLAST hits, the second is a link that retrieves to the flat-file GenBank sequence report for the matched sequence. The last three buttons (**E**, **R**, **S**) are added by the Search Launcher server. (1) Clicking on the **E** button retrieves the NCBI's *Entrez* Document Summary for the matched sequence (immediately above the shaded box). This form includes a pull-down menu (shown expanded) that allows the sequence report to be displayed in various formats (e.g., FASTA, ASN.1). Additional links to related information can also be accessed, e.g., links to Medline abstracts for those literature references included in a sequence report (Medline links), and links to the nucleotide sequences reports for those DNA sequence(s) from which a matched protein sequence was derived (nucleotide links). Selecting the Medline links option and clicking the **Display** button retrieves an *Entrez* Document Summary page that lists the Medline reports linked to the matched sequence (*top left*). The titles of the articles on this page begin to give an indication of the function of the VASA sequence. Selecting the Medline reports option from this page displays the full Medline abstract for a selected report (*top right*). This report provides a great deal of functional information concerning the matched VASA protein sequence. (2) Clicking on the **R** button in the BLAST output (shaded box) returns a list of the sequences related to the matched sequence (protein neighbors) in the *Entrez* data base (*bottom left*). In this case, there are 100 related sequences, only 2 of which are shown. Scanning the list of the title lines of related sequences is often the fastest way to ascertain the function of a matched sequence with an uniformative name. In this case, the VASA sequence is highly similar to an ATP-dependent RNA helicase ($P = 10^{-251}$). (3) Clicking on the **S** button in the BLASTP output retrieves the SRS report for the matched sequence (*bottom right*). This report includes additional links (shown in bold) to a large number of cross-referenced data bases. In this case, the matched sequence is known to match a PROSITE pattern (DEAD_ATP-_HELICASE), and the corresponding link allows one to browse the PROSITE description for the DEAD-box-containing ATP-dependent RNA helicase family.

parameters, (3) transfer the query to the remote server to initiate the search, (4) return the results to our server for further processing, and (5) present the results plus any local additions to the user as an HTML document.

The Common Gateway Interface (CGI) is currently the standard method for interfacing applications (e.g., a data base search tool) with a WWW server (see URL http://hoohoo.ncsa.uiuc.edu/cgi). Unlike an HTML document, which is a static text file, CGI programs are executed in real time and can output dynamic information. CGI programs executed by a Web server can transmit query information entered into a WWW form to an external program, receive the results back from the program, and display the results to a WWW browser. For CGI programming we chose to use the Perl programming language (Wall and Schwartz 1990) for its powerful and flexible text manipulation capabilities. Also, Perl is an interpreted language that does not need to be compiled to run, enhancing rapid code development.

Fill-out forms like those used in the BCM Search Launcher WWW pages require (1) a Form tag in the HTML document to specify the server to which the query will be submitted, and (2) the type of function used to submit the form (GET or POST, see URL http:hoohoo.ncsa.uiuc.edu/forms.html). All of the sequence queries are entered into the forms of the BCM Search Launcher WWW pages using a TEXTAREA form element. The name and size of text areas can be designated; the scroll bars are automatically included. INPUT form elements are used for the **Sequence name/identifier** and **Email address** entry lines, as well as the **Perform search** and **Clear input** buttons. RADIO buttons are used to select the type of search.

GET and POST are the two types of methods available to submit fill-out forms to a query server. GET is the older of the two methods and is the default method. GET causes the fill-out form contents to be appended to the URL (the standardized WWW addressing system; see http://www.w3.org/hypertext/WWW/Addressing/Addressing.html). The POST method sends the contents of a fill-out form to the server as a data file, rather than as part of the URL. POST is the recommended method and is currently being used in our ongoing WWW development efforts. GET/POST functions are available as part of the "libwww-perl" package of programs and libraries for providing a simple and consistent programming interface to the WWW (see URL: http://www.ics.uci.edu/pub/websoft/libwww-perl). This library provides a Perl-based command line interface to the WWW, which allows one to specify a URL and one or more query parameter values, returning the search results back to the program rather than displaying it back to the user. This latter feature is useful, as it allows additional hypertext links to be added to a results file before displaying it to a user (see below).

The BCM Search Launcher system currently consists of >50 HTML documents and 20 Perl scripts. HTML pages include the launch pages (protein search page, nucleic acid search page, etc.), help, options, and parameter pages for each of the accessible services, and general system help pages (e.g., a "Frequently Asked Questions" page). Each of the search tools included in the Search Launcher has one or more Perl scripts devoted to formating and sending the appropriate query, and some search tools also have additional scripts for formatting and interpreting the results.

The BCM Search Launcher Server currently runs on a four-processor Sun SPARCstation 10-514 at Baylor College of Medicine, using the Sun Solaris 2.3 operating system and the NCSA 1.4.2 HTTPD WWW server software. Currently >225 sites worldwide access the Search Launcher server each day on average. To meet the demands of continued growth in this service, we are in the process of migrating the Search Launcher to a larger server provided by the University of Houston's Gene-Server facility. This change will provide increased performance and will be transparent to users of the Search Launcher.

Organization

The BCM Search Launcher WWW pages have been specifically organized to improve the access to sequence analysis tools on the WWW. On the BCM Search Launcher home page (URL: http://gc.bcm.tmc.edu:8088/search-launcher/launcher.html) the available services are first organized by type (see Table 1). A single Launch page is then provided for each type of search. Each Launch page includes input fields for entering the sequence to be searched/analyzed, an optional name/identifier for the sequence, and an Email return address for those services that return results to the user by Email (Fig. 1). Each available analysis tool is listed next to a radio button that is used to select the desired service. A short description of each service is provided, followed by links to three help pages. The [H] hypertext link accesses the help pages provided by the remote server for that service (if one is available). The [O] link accesses the native WWW search form provided by the remote server for that tool, allowing a user to change the optional parameters for a given tool. The [P] link displays the default parameters/options that we have preset for each analysis (see below).

Use

Use of the BCM Search Launcher requires a computer with direct Internet access and a forms-based WWW browser (e.g., NCSA Mosaic, Netscape). To perform a particular analysis, a user pastes a sequence into the sequence input field (using the windowing interface's **Paste** function), enters an optional sequence name/identifier, selects the radio button for the desired analysis tool, and pushes the **Perform search** button to launch the search (see Fig. 1). When complete, the results are returned to the user as an HTML document in the standard WWW browser window. The user can then browse the results and save the file as an HTML document to a local file system. Other searches from the same launch page can then be performed in a similar manner (the query sequence will remain in the sequence input field between searches, unless specifically cleared). Some of the servers return search results as Email rather than HTML documents, and these searches require a user to input their Email address as well.

Choice of Default Parameters

By default, each of the tools on the Search Launcher WWW pages is submitted to the specified remote server using a set of search parameters that we have selected (as noted above, the search parameters can be displayed using the [P] link attached to the name of each tool; see Fig. 1). These search parameters are not necessarily the same as the default parameters used by a particular analysis server; for each case, we have chosen parameter values that give optimal results for a typical search based our experience with that tool. Advanced users can change the default parameters for special cases using the full-options form available using the [O] links on each launch form.

Postprocessing Search Results to Add Links to Other Useful WWW Resources

Because the Search Launcher server receives results files from a remote server before they are displayed to a user, these files can be postprocessed locally to add supplementary WWW hypertext links to analysis results. Currently this is performed for all BLAST search results (BLASTN, BLASTP, BLASTX, TBLASTN, and TBLATX searches) submitted to the NCBI's WWW server (Recipon and Gish 1995). When NCBI BLAST results are returned to the Search Launcher server after search completion, they are passed to a locally executed Perl script that adds three hypertext links for each data base sequence matched in the search (Fig. 2). 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 the set of links to *Entrez* sequences related to the matched data base sequence. The **S** link retrieves the SRS (Etzold and Argos 1993) sequence report for the matched data base sequence. The final output file is then displayed to the user as a standard HTML document.

Choice of Tools and Servers

Data base search and analysis tools available on the WWW were selected for inclusion on the BCM Search Launcher pages based on several criteria: (1) Is the tool useful for molecular biologists? (2) Is the tool available from a reliable server? (3) Will the service be frequently updated and maintained? (4) Will the tool provide added utility, beyond that provided by other tools already included? Public domain tools without current WWW interfaces have also been evaluated for inclusion, and several such programs have been set up to run locally on our server, for example, the MAP multiple sequence alignment program (Huang 1994) and READSEQ, a sequence format conversion program (Gilbert 1990). To keep the tool list current, USENET newsgroups (e.g., bionet.announce, bionet.software, bionet.software.www) and updated tool compilations (e.g., Pedro's BioMolecular Research Tools; Coutinho 1995) are routinely scanned for announcements of new tools that might be incorporated into the Search Launcher.

Addition of Locally Developed Tools

In addition to the search and analysis tools available elsewhere, we have included several locally developed tools unique to this server. These tools include several new sequence data base search tools (BEAUTY and BEAUTY-X, Worley et al. 1995; and FASTA-SWAP and FASTA-PAT; Ladunga et al. 1996) as well as the PIMA multiple sequence alignment program (Smith and Smith 1992). In addition, we have added two DNA sequence utility programs (six-frame translation and DNA reverse complement) that we have written based on Perl code adapted from the DNA Workbench package (Tisdall 1995).

Batch Client Interface for Unix and Macintosh Systems

Because several of the searches available via the BCM Search Launcher can take some time to complete, we developed a batch client interface to our Search Launcher server to allow background searches to be run in an automated fashion. The batch client is written in Perl and currently runs on both Unix and Macintosh personal computers (a Microsoft–Windows version will be made available once the Perl package is fully ported to this system). The batch client will (1) query the BCM Search Launcher server for the current list of analysis services available, (2) query the user for a search/analysis program to be run, (3) read in multiple sequences from one or more specified input files, (4) perform the specified search in the background for each sequence (one at a time), and (5) store the result files as individual HTML documents that can be opened and read using a WWW browser such as NCSA Mosaic or Netscape. The batch client uses the READSEQ sequence file conversion program (Gilbert 1990) to read in sequences, so that any sequence file type compatible with READSEQ (e.g., Pearson/FASTA, GenBank, EMBL, NBRF/PIR, IntelliGenetics, etc.) can be used as an input file type.

On the Macintosh, the batch client has a very easy-to-use "drag-and-drop" interface: One simply drops a text file containing one or more sequences (in a READSEQ-compatible format) onto the icon of the Search Launcher batch client. A menu of all currently available services is displayed to the user, and the user simply selects one of the programs from the menu to start the search process. On Unix systems, a command-line interface is available that allows fully automated searches to be implemented.

The software for the BCM Search Launcher batch client is available via anonymous ftp at ftp://gc.bcm.tmc.edu/pub/software/search-launcher. A direct network connection to the Internet is required. Unix systems require the Perl package to be pre-installed. Macintoshes do not require Perl (a Perl run time is included) but do require AppleScript (included in MacOS 7.5) and MacTCP.

ACKNOWLEDGMENTS

We thank Drs. Istvan Ladunga and Karen Kabnick for critical reading of the manuscript. This work was supported by grants to the Baylor Human Genome Center (P30-HG00210), to R.F.S. (1R01-HG00973-01), and to K.C.W. (1F32-HG00133-01) from the National Center for Human Genome Research, National Institutes of Health; a grant to D.B.D. and R.F.S. from the Office of Health and Environmental Research, U.S. Dept. of Energy (DE-FG03-95ER62097/A000); 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.

Alschul, S.F., M.S. Boguski, W. Gish, and J.C. Wootton. 1994. Issues in searching molecular sequence databases. *Nature Genet.* **6:** 119–129.

Boguski, M.S. 1992. Computational sequence analysis revisited: New databases, software tools, and the research opportunities they engender. *J. Lipid Res.* **33:** 957–974.

———. 1994. Bioinformatics. *Curr. Opin. Genet. Dev.* **4:** 383–388.

Brenner, S.E. 1995a. World Wide Web and molecular biology. *Science* (letter) **268:** 622–623.

———. 1995b. BLAST, Blitz, BLOCKS and BEAUTY: Sequence comparison on the Net. *Trends Genet.* **11:** 330–331.

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

Coutinho, P.M. 1995. Pedro's BioMolecular Research Tools—A collection of WWW links to information and services useful to molecular biologists. URL: http://www.public.iastate.edu/~pedro/research_tools.html.

deBoer, T. 1993. MailFasta. Dept. of Microbiological Physiology, Vrije universiteit, Amsterdam, The Netherlands. URL: ftp://ftp.ebi.ac.uk/pub/software/unix/mailfasta.shar.

Epstein, J.A., J.A. Kans, and G.D. Schuler. 1994. WWW *Entrez*, A hypertext retrieval tool for molecular biology. Electronic Proceeding of the Second World Wide Web Conference '94: Mosaic and the Web. URL: http://www3.ncbi.nlm.nih.gov/www2/www2_paper2.html.

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

Fuchs, R. 1994. Sequence analysis by electronic mail: A tool for accessing Internet e-mail servers. *Comput. Appl. Biosci.* **10:** 413–417.

Gilbert, D.G. 1990. ReadSeq. Biology Department, Indiana University, Bloomington. URL: http://iubio.bio.indiana.edu:80/1/IUBio-Software%2bData/molbio/readseq.

Harper, R. 1994. Access to DNA and protein databases on the Internet. *Curr. Opin. Biotechnol.* **5:** 4–18.

———. 1995. World Wide Web resources for the biologist. *Trends Genet.* **11:** 223–228.

Huang, X. 1994. On global sequence alignment. *Comput. Appl. Biosci.* **10:** 227–235.

Jacobson, D. 1994. The World Wide Web for biologists. *Protein Sci.* **3:** 2159–2161.

Ladunga, I., B.A. Wiese, and R.F. Smith. 1996. FASTA-SWAP and FASTA-PAT: Pattern database searches using combinations of aligned amino acids, and a novel scoring theory. *J. Mol. Biol.* (in press).

McKinney, W.P., J.M. Wagner, G. Bunton, and L.M. Kirk. 1995. A guide to Mosaic and the World Wide Web for physicians. *MD Computing* **12:** 109–114, 141.

Recipon, H. and W. Gish. 1995. Basic Local Alignment Search Tool (BLAST) Notebook Home Page. National Center for Biotechnology Information, National Library of Medicine. URL: http://www.ncbi.nlm.nih.gov/Recipon/index.html.

Robison, K. 1995. The WWW Virtual Library: Biosciences. URL: http://golgi.harvard.edu/biopages.html.

Schatz, B.R. and J.B. Hardin. 1994. NCSA Mosaic and the World Wide Web: Global hypermedia protocols for the Internet. *Science* **265:** 895–901.

Scherer, S. 1995. Virtual Genome Center, University of Minnesota School of Medicine, Minneapolis, MN. URL: http://alces.med.umn.edu/VGC.html.

Smith, R.F. 1995. A brief guide to information resources supporting the Human Genome Project. *IEEE Eng. Med. Biol.* **14:** 760–761. (available as URL: http://gc.bcm.tmc.edu:8088/bio/hgp-resource-guide.html).

Smith, R.F., and T.F. Smith. 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.

Tisdall, J. 1995. DNA Workbench, Departments of Genetics and Computer and Information Science, University of Pennsylvania, Philadelphia, PA. URL: ftp://cbil.humgen.upenn.edu/pub/dnawork-bench/announce.

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

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

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:** 173–184.

Received February 26, 1996; accepted in revised form April 8, 1996.