GWFASTA: Server for Eukaryotic and Microbial Genomes

BioTechniques 33:__-__ (September 2002)

ABSTRACT

Similarity searches are a powerful method for solving important biological problems such as database scanning, evolutionary studies, gene prediction, and protein structure prediction. FASTA is a widely used sequence comparison tool for rapid database scanning. Here we describe the GWFASTA server that was developed to assist the FASTA user in similarity searches against partially and/or completely sequenced genomes. GWFASTA consists of more than 60 microbial genomes, eight eukaryote genomes, and proteomes of annotated genomes. In fact, it provides the maximum number of databases for similarity searching from a single platform. GWFASTA allows the submission of more than one sequence as a single query for a FASTA search. It also provides integrated post-processing of FASTA output, including compositional analysis of proteins, multiple sequences alignment, and phylogenetic analysis. Furthermore, it summarizes the search results organism-wise for prokaryotes and chromosome-wise for eukaryotes. Thus, the integration of different tools for sequence analyses makes GWFASTA a powerful tool for biologists.

INTRODUCTION

In the past few years, a considerable number of complete genome sequences of different organisms belonging to archaea, bacteria, and eukaryotes have been reported. The availability of these genome sequences provides us with an opportunity to find homologous sequences throughout the genomes using comparative similarity searches. Homologous sequences are those sequences that are related by distant ancestry. Similarity between two sequences reflects similar compositional properties, but the evolutionary placement may or may not be related. Although homologous sequences have sequence similarity in general, not all similar sequences are homologous. Searching for similar sequences in different organisms is a widely used method for gene characterization and annotation and for detecting homologs across genomes. This is because different organisms sharing a distant ancestor encode in their genomes similar proteins with high sequence similarity (13,25).

One powerful algorithm to calculate optimal alignment or similarity between two sequences is the Needleman-Wunsch method (16). This method is an efficient algorithm for creating global alignment between two sequences of similar lengths. However, it is unsuitable for searching databases because they contain sequences of different lengths. A variety of algorithms has been developed to conduct similarity based on local alignment strategies [e.g., Smith-Waterman, FASTA, and BLAST algorithms (2,3,20,23,27)]. The Smith-Waterman method introduced the concept of similarity between a pair of segments from two long sequences, which is called local alignment using dynamic programming (23).

BLAST and FASTA are two rapid approximation techniques that are widely used for database searching (2,3,20,27). Both algorithms have their advantages and disadvantages. For example, BLAST outperforms FASTA and SSEARCH in terms of speed, and the latest version of BLAST performs better than FASTA when one is using default parameters and comparing protein sequences (1,19). FASTA performs better than BLAST on nucleotide sequences; FASTA was found to have 43.2% coverage on default parameters compared to 21.6% for BLAST (4). Another advantage of the FASTA method is its ability to create a full-length alignment of a pair of sequences rather than several short high-segment pairs.

There are more than 50 Web servers worldwide that provide BLAST searches against various sequence databases, including NCBI, TIGR, and SANGER. In comparison with the number of BLAST servers, there are only a few servers that allow FASTA searches against various databases (http://www.imtech.res.in/raghava/gwfasta/links.html). For this reason, we use FASTA for sequence similarity searches.

As the number of genome sequences in public databases continues to grow, the output generated by typical FASTA searches is voluminous. It has made the manual parsing of FASTA reporting increasingly difficult. However, such a parsed report should not lose the rich information content of the FASTA report. The existing FASTA search methods also lack a genomic perspective in their presentation of results. Any user who wants to post-process the FASTA report has to visit multiple servers, which is time-consuming and makes data transfer prone to manual error. A solution to such post-processing of similarity search reports is presented on the NPS@ server (http://npsa-pbil.ibcp.fr/). The NPS@ server addresses the problem of automated and continuous protein sequence analysis by integrating approximately 25 autonomous components or programs (6). However, these programs tend to focus on particular problems and are ineffective for particular studies. Thus, we need a reliable tool that can accurately combine evidence from genomic sequence comparisons with the traditional clues from intrinsic sequence properties and the results of protein and nucleotide database searches (14).

Here we describe GWFASTA, a Web server that was developed to eliminate some of the difficulties faced by database users. The server uses a FASTA3 software package for similarity searching. It offers a flexible and convenient user interface that supports searches against user-selected multiple genome and proteome databases; fully automated batch submission of query sequences; searches with multiple FASTA programs (Table 2); and convenient post-processing of FASTA output. This paper describes the architecture, options, and applications of the GWFASTA server.

MATERIAL AND METHODS

Architecture

GWFASTA can be downloaded free of charge at http://www.imtech.res.in/raghava/gwfasta. The common gateway interface script of GWFASTA is writ-

Table 1. Genomes and other Databases Available in GWFASTA

Database Type	Release	Center[a]
11 archaeal, 54 bacterial, and 8 eukaryotic fully or partially sequenced genomes and their proteomes if the annotation is available	–	NCBI and JGI
Nonredundant Protein Database	Feb. 20, 2002	NCBI
Protein Data Bank Sequences	Feb. 20, 2002	NCBI
Swissprot Sequences	Feb. 20, 2002	NCBI
Patented DNA and Amino Acids Sequences	Feb. 20, 2002	NCBI
Sequence-Tagged Sites	Feb. 20, 2002	NCBI
Human ALU Sequences	Feb. 20, 2002	NCBI
Vector DNA Sequences	Feb. 20, 2002	NCBI
PRODOM Protein Sequences	2001.2	INRA
Intron Database	GenBank® Release 116	BIC, NUS

[a] The different databases are automatically updated weekly.

NCBI, National Center for Biotechnology Information; INRA, National Institute of Agronomic Research, France. BIC, NUS, Intron Database at Bioinformatics Center of National University of Singapore.

Table 2. FASTA Programs

Program	Query Type	Target Type	Remarks
fasta34	Protein or Nucleotide	Protein or Nucleotide, respectively	Scan a protein or DNA sequence library for similar sequences.
tfasta34	Protein	Nucleotide (translated)	Compare a protein sequence to a DNA sequence library, translating the DNA sequence library "on-the-fly".
tfastx34	Protein	Nucleotide (translated)	Compare a protein sequence to a DNA sequence database, calculating similarities with frameshifts to the forward and reverse orientations.
tfasty34	Protein	Nucleotide (translated)	Compare a protein sequence to a DNA sequence database, calculating similarities with frameshifts to the forward and reverse orientations.
fastx34	Nucleotide (translated)	Protein	Compare a DNA sequence to a protein sequence database, comparing the translated DNA sequence in forward and reverse frames.
fasty34	Nucleotide (translated)	Protein	Compare a DNA sequence to a protein sequence database, comparing the translated DNA sequence in forward and reverse frames.
ssearch34	Protein or Nucleotide	Protein or Nucleotide, respectively	Compare a protein or DNA sequence to a sequence database using the Smith-Waterman algorithm.

ten in PERL version 5.03. The GWFASTA server is installed on a Sun Server (420E) under a UNIX (Solaris 7) environment. The server has four 450-MHz UltraSparc II CPUs with 4 MB L2 Cache and 2 GB (8 × 256 MB) RAM. It has two internal Ultra SCSI hard disks of 18 GB each (10 000 rpm), a Fiber-Channel RAID Storage Array (9 × 36 GB), and a redundant power supply to keep the server working in the event of power failures. The GWFASTA server is capable of handling a heavy load of queries from users. The Apache Web server was installed to launch the GWFASTA server, which incorporates the FASTA3 version 3.4 (obtained from ftp://ftp.virginia.edu/ pub/fasta).

Databases

The server maintains 65 microbial genomes, including 11 from archaea and 54 from bacteria (Table 1). GWFASTA also provides eight eukaryotic genomes including assembly sequences of *Fugu rubripes*, the Japanese Puffer fish (Joint Genome Institute, The University of California, The US Department of Energy). All the annotated genomes have their proteomes available for FASTA searches in the GWFASTA server. There are 53 microbial proteomes for similarity searching, including 11 archaea and 42 bacteria proteomes. Proteomes for seven eukaryotes are available for similarity searching, including GeneScan®-based predictions for proteins in *F. rubripes* assembly sequences. The GWFASTA server is among the few sites to offer the maximum number of databases available for searches on a single platform. An important requirement for any dynamic server that provides for similarity searches is constant updating of the databases. We have installed the mirror package that maintains the mirror site of these databases so that databases can be updated weekly.

Batch Processing and E-Mail Reply

A significant feature of the GWFASTA server is the batch-processing capability. Users that have multiple query sequences to be processed can submit their sequences in a single visit.

The server provides for low-priority searches for such users. This allows users with a small number of sequences to get their searches done without being delayed by multiple query searches. The server queues all the jobs, single or multiple, and a job identification number for each query is generated that can later be used to retrieve the results for further analysis. The server provides an option that allows users to obtain the results of a FASTA search via e-mail, preserving the typical FASTA output and providing a URL link for further processing.

RESULTS AND DISCUSSION

Output from Whole Genome and Proteome FASTA Searches

The output from a genomic or proteomic search helps one to understand the FASTA report while retaining its rich content information. The GWFASTA parses the FASTA report differently for protein and nucleotide queries. For the proteins, the server summarizes the FASTA output proteome-wise (Figure 2), while for nucleotides, it tabulates the results chromosome-wise for eukaryotes and genome-wise for prokaryotes. This output format not only helps to compare the FASTA results genome to genome but also aids the biologists in localizing the query sequence to a particular region in the genome. The E-value and score of the top FASTA hit in a given genome or proteome are the two basic pieces of information provided in the report.

The raw output for an individual genome is available to the user through a link in the tabular output that is generated by the server. This link provides the same option as the report generated by FASTA searches against standard databases. It is possible to extract the top or all hits for further analysis, which helps in eliminating spurious hits from subsequent analysis, while the raw output for a particular organism is also accessible through a link. Users also have the option of post-processing their FASTA report. The server extracts whole protein sequences from the databases at the user's request, while the nucleotide sequence is extracted from the alignment generated by the different FASTA programs.

Report from Searches against Standard Databases

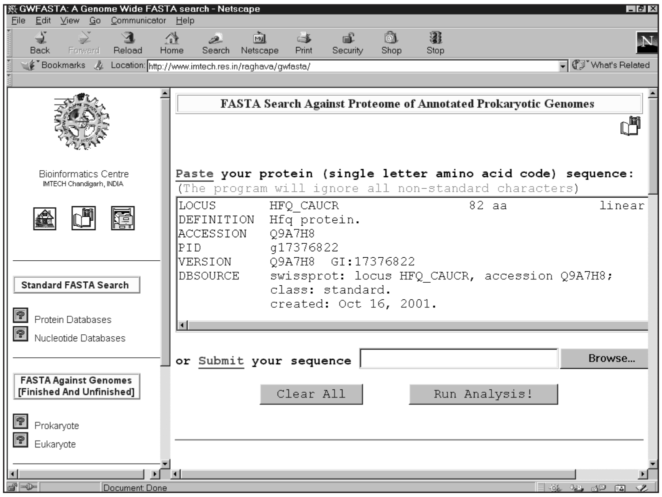
When any of the FASTA program runs competes against standard databases, the user is presented with a typical report with the usual FASTA findings, such as the E-value significance of the match, the number and length of the aligned sequences, and the alignment of query and database sequence at locally aligned regions (Figure 3). Biologists can select the best alignments according to their interpretation and proceed for analysis on the GWFASTA server itself. This saves time and prevents error caused by transferring sequence data from one program to another.

Visualization of Alignment

Mview is a program that allows the coloring of residues in an alignment, thereby helping to detect conserved regions and groups of residues with common properties such as hydrophobicity and polarity (5). The program provides options to select various parameters for beautiful presentations of alignments. The options are integrated into the server at two points. One option is selectable after FASTA searches for viewing FASTA alignments, and the other is selectable after multiple sequence alignments with ClustalW (Figure 4). These options provide users with useful and objective methods to view the alignment report.

Compositional Analysis and Thermostability

The frequency of individual amino acids or certain groups of residues (e.g., charged, polar, and hydrophobic)


GWFASTA: A Genome Wide FASTA search - Netscape
File Edit View Go Communicator Help
Back Forward Reload Home Search Netscape Print Security Shop Stop
Bookmarks Location: http://www.imtech.res.in/raghava/gwfasta/
What's Related
Bioinformatics Centre
IMTECH Chandigarh, INDIA
Standard FASTA Search
Protein Databases
Nucleotide Databases
FASTA Against Genomes [Finished And Unfinished]
Prokaryote
Eukaryote
Browse...
Clear All
Run Analysis!
Document: Done

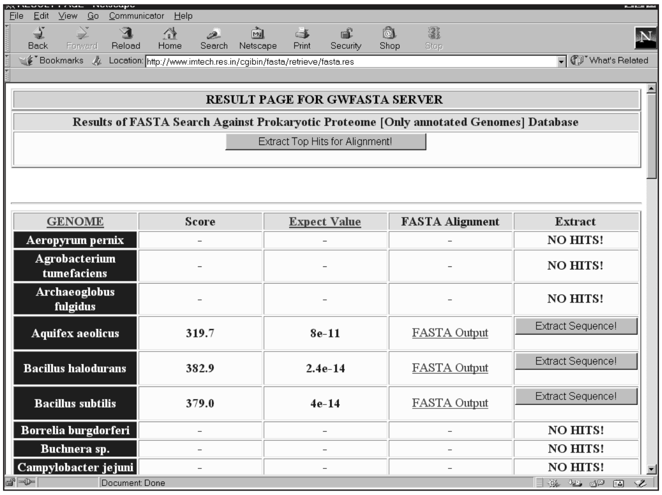
FASTA Search Against Proteome of Annotated Prokaryotic Genomes

Paste your protein (single letter amino acid code) sequence:
(The program will ignore all non-standard characters)

```
LOCUS       HFQ_CAUCR                 82 aa            linear
DEFINITION  Hfq protein.
ACCESSION   Q9A7H8
PID         g17376822
VERSION     Q9A7H8  GI:17376822
DBSOURCE    swissprot: locus HFQ_CAUCR, accession Q9A7H8;
            class: standard.
            created: Oct 16, 2001.
```

or Submit your sequence

Figure 1. Screenshot of the GWFASTA server showing two frames. The left frame shows the options for similarity searches, and the right frame shows the submission form for searching query sequence against prokaryotic proteome databases. This screenshot shows the Input sequence, Hfq RNA binding protein (82 aa) of *C. vibrioides* from Swissprot database with the accession no. Q9A7H8. The Fasta34 search was carried out against all prokaryotic proteome databases with parameters. ktup, 2; E-value, 10; and Weight matrix, BLOSUM50.


File Edit View Go Communicator Help
Back Forward Reload Home Search Netscape Print Security Shop Stop
Bookmarks Location: http://www.imtech.res.in/cgibin/fasta/retrieve/fasta.res
What's Related
Extract Top Hits for Alignment!
Document: Done

RESULT PAGE FOR GWFASTA SERVER

Results of FASTA Search Against Prokaryotic Proteome [Only annotated Genomes] Database

GENOME	Score	Expect Value	FASTA Alignment	Extract
Aeropyrum pernix	-	-	-	NO HITS!
Agrobacterium tumefaciens	-	-	-	NO HITS!
Archaeoglobus fulgidus	-	-	-	NO HITS!
Aquifex aeolicus	319.7	8e-11	FASTA Output	Extract Sequence!
Bacillus halodurans	382.9	2.4e-14	FASTA Output	Extract Sequence!
Bacillus subtilis	379.0	4e-14	FASTA Output	Extract Sequence!
Borrelia burgdorferi	-	-	-	NO HITS!
Buchnera sp.	-	-	-	NO HITS!
Campylobacter jejuni	-	-	-	NO HITS!

Figure 2. Proteome-wise summary of the result for FASTA searches of Hfq RNA binding protein against all the prokaryotic proteome databases.

is a valuable indicator of the thermostability of the proteins (9). Observations that on average thermostable proteins have more charged residues and fewer polar residues suggest that a compositional profile of the residues in the query protein and FASTA hits could be useful for the deduction of the thermostability of query proteins. The GWFASTA server allows one to calculate the compositional frequency of amino acids or groups of residues of various user-selected proteins from FASTA hits.

Multiple Sequence Alignment and Phylogenetic Analysis

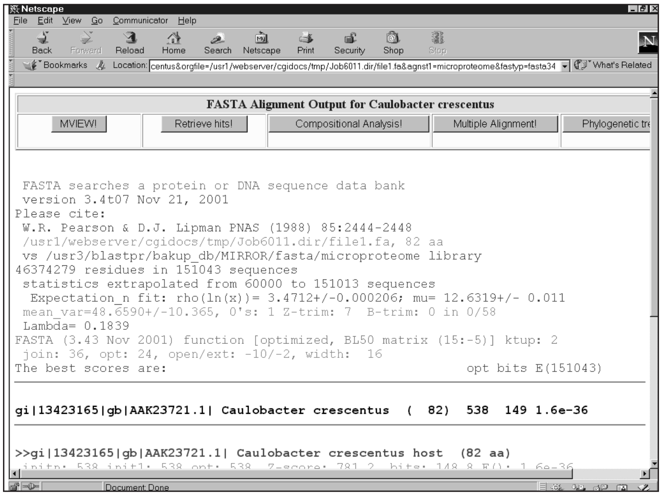
The multiple sequence alignment of biological sequences has been used to find characteristic motifs and conserved regions in protein families, the determination of evolutionary linkage, and the improved prediction of protein secondary and tertiary structure. One of the most used programs for multiple sequence alignment is ClustalW (24). This program is integrated in the GWFASTA server, which allows its users to perform multiple sequence alignment on selected top hits from a FASTA output. The phylogenetic analysis can be performed using the tree generated by the ClustalW. A Web server, Phylodendron, is integrated with our server and allows the creation and visualization of a phylogenetic tree. Phylodendron provides several options including the selection of the type of tree (e.g., cladogram, phenogram, and swoopogram) and the format for the generated image (e.g., GIF, Postscript, or PDF).

Editing of Alignment

Jalview, an alignment-editing server, is integrated with GWFASTA to edit and manipulate the multiple sequence alignment of FASTA hits. Jalview is a Java-based tool that easily manipulates the protein alignment for the user. It combines display speed and consensus color schemes with easy access to the public databases using CORBA or CGI. The server colors the residues by the physicochemical properties of amino acid, similarity to consensus sequence, hydrophobicity, or secondary structure. It performs additional pairwise alignment using the Smith-Waterman algorithm and can send colored postscripts of the output by e-mail.

Analysis of Multiple Sequence Alignment

Another GWFASTA integrated server, AMAS, has a strategy based on a flexible, set-based description of amino acid properties that defines the conservation between any groups of amino acids (12). The sequences in the alignment are in subgroups based on sequence similarity, functional, evolutionary, or other criteria. The comparison of all pairs of subgroups highlights positions that confer the subgroup's unique features. AMAS provides a textual summary of the analysis and an annotated (boxed, shaded, and/or colored) multiple sequence alignment. The server simplifies the analysis of multiple sequence data by condensing the mass of information present and thus allows the rapid identification of substitutions of structural and functional importance.



centus&orgfile=/usr1/webserver/cgidocs/tmp/Job6011.dir/file1.fa&agnst1=microproteome&fastyp=fasta34

FASTA Alignment Output for Caulobacter crescentus

```
 FASTA searches a protein or DNA sequence data bank
 version 3.4t07 Nov 21, 2001
Please cite:
 W.R. Pearson & D.J. Lipman PNAS (1988) 85:2444-2448
 /usr1/webserver/cgidocs/tmp/Job6011.dir/file1.fa, 82 aa
 vs /usr3/blastpr/bakup_db/MIRROR/fasta/microproteome library
46374279 residues in 151043 sequences
 statistics extrapolated from 60000 to 151013 sequences
  Expectation_n fit: rho(ln(x))= 3.4712+/-0.000206; mu= 12.6319+/- 0.011
 mean_var=48.6590+/-10.365, 0's: 1 Z-trim: 7  B-trim: 0 in 0/58
 Lambda= 0.1839
FASTA (3.43 Nov 2001) function [optimized, BL50 matrix (15:-5)] ktup: 2
 join: 36, opt: 24, open/ext: -10/-2, width:  16
The best scores are:                                   opt bits E(151043)

gi|13423165|gb|AAK23721.1| Caulobacter crescentus  (  82)  538  149 1.6e-36

>>gi|13423165|gb|AAK23721.1| Caulobacter crescentus host  (82 aa)
 initn: 538 init1: 538 opt: 538  Z-score: 781.2  bits: 148.8 E(): 1.6e-36
```

Figure 3. Parsed FASTA report of query Hfq RNA binding protein for hits against *C. cresentus* obtained after clicking the "FASTA Output" in the proteome-wise FASTA report from Figure 3. The report is similar in structure to that obtained from typical FASTA searches against standard protein databases. The screenshot shows the various options available for the post-processing of FASTA output.

Property Plots

GWFASTA also allows one to plot the amino acid properties of protein sequences along sequence alignment through another integrated server called the PSA (22). PSA allows one to plot separate graphs that correspond to each sequence in multiple sequence alignments. This server computes and presents the overall property of each position in the alignment, highlights the conserved residues in the alignment, highlights residues in the alignment that exhibit specific functions, computes a position specific score matrix, and exhibits similarity among the sequences in the alignment. The server also helps to identify the protein from the multiple sequence alignment that has the highest similarity with other sequences, indicating the representative protein.

How to Use the GWFASTA Server

GWFASTA is an integrated server with several programs that consists of two frames. The left frame (Figure 1) shows the various options that include FASTA searches against standard protein databases, including standard nucleotide, prokaryotic genome, prokaryotic proteome, eukaryotic genome, and eukaryotic proteome databases. Users can select any of the above options to perform sequence similarity searches for their protein or DNA query sequence and can search their protein or nucleotide query sequence against standard/genomic protein or nucleotide databases, respectively. Users can also search their protein query against translated standard/genomic protein and nucleotide databases when required. Options have been provided to the users to search their translated nucleotide query against standard/genomic protein databases or against translated standard or genomic nucleotide databases (Table 2).

For example, if users want to perform sequence similarity searches for their protein against prokaryotic proteome databases, then they should select the fourth option listed above. Users will observe a submission form on the right frame of the browser (see Figure 1). They can paste their sequence in the submission form and select various search parameters such as the k-tuple, matrix to be used, and format type of the query sequence. Users should provide their e-mail addresses to obtain the results. The submission form also allows users to select the particular genomes/proteomes against which they want to perform the search.

On the submission of the form, the user will get a summarized report (Figure 2). Similar sequences in the proteome of selected organisms are listed as a table along with their score and E-value. The server allows the extraction of all FASTA hits in selected proteomes or hits against any individual proteome (see Figure 2). In case the user clicks on “FASTA Output” for *Caulobacter cresentus*, they will observe the parsed FASTA report as shown (Figure 3). The parsed FASTA report is similar to the typical FASTA report that is generated by searches against standard databases.

The server allows the post-processing of FASTA search results (Figure 3) that includes (*i*) viewing of FASTA alignment using Mview; (*ii*) ClustalW for multiple sequence alignment; (*iii*) generation of phylogenetic tree; (*iv*) compositional analysis of user-selected sequences; and (*v*) editing and analysis of multiple sequence alignment. Users can generate and view the multiple sequence alignment of FASTA hits by clicking on “Multiple Alignment” (Figure 4a) and generate the phylogenetic tree by clicking on “View Phylogenetic Tree” using Phylodendron (Figure 4b).

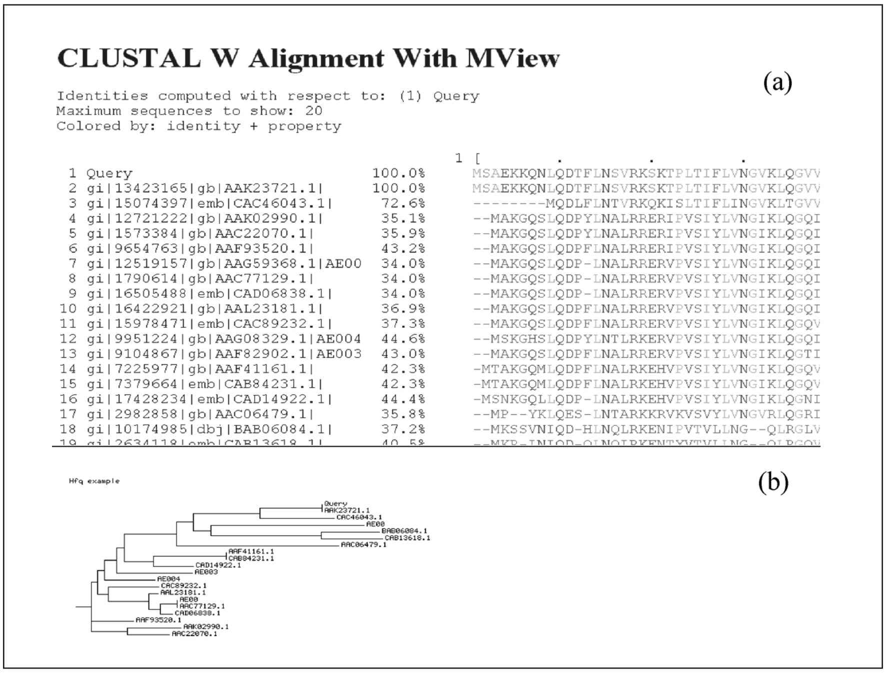
Gene Characterization

The annotation of available sequences from increasingly rapid genome sequencing projects is a major problem. A quick method for the reliable and accurate characterization of genomic sequences is sequence similarity searches with available annotated databases (13). It is generally accepted that the probability of correct annotation of newly sequenced genomic sequences increases if known sequences are present in the genomes of other organisms. Numerous databases in the GWFASTA server ensure that the user can confirm the annotation of sequences in multiple ways. While an indication of the functionality of the sequence can be observed from similarity searching against nonredundant databases, the presence of homologous sequences in the same or different genome helps to derive the sequence function. Specialized databases such as ProDom, Intron, and ALU can help to locate the functionally or biologically significant domains in the sequence.

Evolutionary Studies and Phylogenetic Analysis

Linking orthologs from a set of sequences is a prerequisite for the meaningful extrapolation of gene functional studies from invertebrates to humans (26). The formatted GWFASTA report also ensures the detection of xenologs present in genomes because of the horizontal gene transfer from other species. Similarity searches against genome databases can also detect paralogs (10). However, a phylogenetic analysis allows a better understanding of the molecular relationships between the sequences (11). Phylogenetic divergence can be gauged from the number or type of changes in the aligned residues in a multiple sequence alignment. While the alignment provides a prediction as to which residues correspond, each column in the alignment shows the mutations that occurred during the evolution of a sequence family. A group of sequence families may share some unique features that were not present in their distant ancestors. The phylogenetic tree helps the user to visually appreciate the degree of divergence of the aligned sequences.

One should note that the word “similarity” in computer parlance refers to the compositional matches between any two sequences, and these sequences are not “homologs” of each other (21). Homologous sequences are those that share some common ancestry or origin. The GWFASTA server can help to identify similar sequences for a user query that can potentially be homologs of the query.



CLUSTAL W Alignment With MView

```
Identities computed with respect to: (1) Query
Maximum sequences to show: 20
Colored by: identity + property

                                           1 [         .         .         .
 1 Query                           100.0%    MSAEKKQNLQDTFLNSVRKSKTPLTIFLVNGVKLQGVV
 2 gi|13423165|gb|AAK23721.1|      100.0%    MSAEKKQNLQDTFLNSVRKSKTPLTIFLVNGVKLQGVV
 3 gi|15074397|emb|CAC46043.1|      72.6%    --------MQDLFLNTVRKQKISLTIFLINGVKLTGVV
 4 gi|12721222|gb|AAK02990.1|       35.1%    --MAKGQSLQDPYLNALRRERIPVSIYLVNGIKLQGQI
 5 gi|1573384|gb|AAC22070.1|        35.9%    --MAKGQSLQDPYLNALRRERIPVSIYLVNGIKLQGQI
 6 gi|9654763|gb|AAF93520.1|        43.2%    --MAKGQSLQDPFLNALRRERIPVSIYLVNGIKLQGQI
 7 gi|12519157|gb|AAG59368.1|AE00   34.0%    --MAKGQSLQDP-LNALRRERVPVSIYLVNGIKLQGQI
 8 gi|1790614|gb|AAC77129.1|        34.0%    --MAKGQSLQDP-LNALRRERVPVSIYLVNGIKLQGQI
 9 gi|16505488|emb|CAD06838.1|      34.0%    --MAKGQSLQDP-LNALRRERVPVSIYLVNGIKLQGQI
10 gi|16422921|gb|AAL23181.1|       36.9%    --MAKGQSLQDPFLNALRRERVPVSIYLVNGIKLQGQI
11 gi|15978471|emb|CAC89232.1|      37.3%    --MAKGQSLQDPFLNALRRERVPVSIYLVNGIKLQGQV
12 gi|9951224|gb|AAG08329.1|AE004  44.6%    --MSKGHSLQDPYLNTLRKERVPVSIYLVNGIKLQGQI
13 gi|9104867|gb|AAF82902.1|AE003  43.0%    --MAKGQSLQDPFLNALRRERVPVSIYLVNGIKLQGTI
14 gi|7225977|gb|AAF41161.1|        42.3%    -MTAKGQMLQDPFLNALRKEHVPVSIYLVNGIKLQGQV
15 gi|7379664|emb|CAB84231.1|       42.3%    -MTAKGQMLQDPFLNALRKEHVPVSIYLVNGIKLQGQV
16 gi|17428234|emb|CAD14922.1|      44.4%    -MSNKGQLLQDP-LNALRKEHVPVSIYLVNGIKLQGNI
17 gi|2982858|gb|AAC06479.1|        35.8%    --MP--YKLQES-LNTARKKRVKVSVYLVNGVRLQGRI
18 gi|10174985|dbj|BAB06084.1|      37.2%    --MKSSVNIQD-HLNQLRKENIPVTVLLNG--QLRGLV
19 gi|2634118|emb|CAB13618.1|       40.5%    --MKP-INIQD-QLNQLRKENTYVTVLLNG--QLRGQV
```

Hfq example

Figure 4. ClustalW alignment with Mview. (a) Screenshot of multiple sequence alignment generated by ClustalW of the selected FASTA hits for Hfq RNA binding protein against prokaryotic protein databases (above E-value = Xe-10, where X is any integer) and viewed using the Mview option in Figure 3. The aligned residues are shown in different colors, while the unaligned residues are in gray. (b) Screenshot of the phylogenetic tree (phenogram) generated from the multiple sequence alignment in the above example. The GIF image was generated using the Phylodendron server integrated with the GWFASTA server.

Protein Structure Prediction

For several years, homology searching has been used in protein structure prediction, based on the detection of significant similarities at the sequence level. One can achieve a higher level of accuracy on the detection of a significant match with a sequence of known structure in the database (18). An alignment of similar sequences and subsequent profile analysis can help in protein structure prediction (7). Similarly, using multiple sequence alignment of homologous sequences to detect conserved structural elements makes such predictions more accurate (8). GWFASTA not only helps to find homologous sequences but also aids in the multiple sequence alignment of similar sequences from a FASTA search.

Locating Proteins in Cell

Intracellular and extracellular proteins have different amino acid compositions, and their location may therefore be discernible from composition data alone (17). The presence of more charged residues in a protein tends to prevent its transport across the membrane, while more hydrophobic residues aid in trapping the protein in the membrane. Smaller amounts of hydrophobic and charged residues are ideally suited for protein transport across the membrane (15). The presence of amino acid residues (e.g., Pro and Cys) and polar residues (e.g., Ser, Thr, Asn, and Gln) retards the kinetics of protein folding. GWFASTA allows the compositional analysis of selected FASTA hits from protein databases and helps in predicting structural classes of proteins, their location in cells, and the kinetics of their folding.

REFERENCES

1. **Agarwal, P. and D.J. States.** 1998. Comparative accuracy of methods for protein sequence similarity search. Bioinformatics *14*:40-47.
2. **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.
3. **Altschul, S.F., T.L. Madden, A.A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D.J. Lipman.** 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. *25*:3389-3402.
4. **Anderson, I. and A. Brass.** 1998. Searching DNA databases for similarities to DNA sequences: when is a match significant? Bioinformatics *14*:349-356.
5. **Brown, N.P., C. Leroy, and C. Sander.** 1998. MView: A Web-compatible database search or multiple alignment viewer. Bioinformatics *14*:380-381.
6. **Combet, C., C. Blanchet, C. Geourjon, and G. Deleage.** 2000. NPS@: network protein sequence analysis. Trends Biochem. Sci. *25*:147-150.
7. **Cuff, J.A. and G.J. Barton.** 2000. Application of multiple sequence alignment profiles to improve protein secondary structure prediction. Proteins *40*:502-511.
8. **Francesco, V.D., J. Garnier, and P.J. Munson.** 1996. Improving protein secondary structure prediction with aligned homologous sequences. Protein Sci. *5*:106-113.
9. **Fukuchi, S. and K. Nishikawa.** 2001. Protein surface amino acid compositions distinctively differ between thermophilic and mesophilic bacteria. J. Mol. Biol. *309*:835-843.
10. **Gogarten, J.P. and L. Olendzenski.** 1999. Orthologs, paralogs and genome comparisons. Curr. Opin. Genet. Dev. *9*:630-636.
11. **Le Gall, O., T. Candresse, and J. Dunez.** 1995. A multiple alignment of the capsid protein sequences of nepoviruses and comoviruses suggests a common structure. Arch. Virol. *140*:2041-2053.
12. **Livingstone, C.D. and G.J. Barton.** 1993. Protein sequence alignments: a strategy for the hierarchical analysis of residue conservation. Comput. Appl. Biosci. *9*:745-756.
13. **Manuel, A., D. Beaupain, P.H. Romeo, and N. Raich.** 2000. Molecular characterization of a novel gene family (PHTF) conserved from *Drosophila* to mammals. Genomics *64*:216-220.
14. **Miller, W.** 2000. Comparison of genomic DNA sequences: solved and unsolved problems. Bioinformatics *17*:391-397.
15. **Nakashima, H.** 1994. Discrimination of intracellular and extracellular proteins using amino acid composition and residue-pair frequencies. J. Mol. Biol. *238*:54-61.
16. **Needleman, S. and C. Wunsch.** 1970. A general method applicable to search for similarities in the amino acid sequences of two proteins. J. Mol. Biol. *48*:444-453.
17. **Nishikawa, K., Y. Kubota, and T. Ooi.** 1983. Classification of proteins into groups based on amino acid composition and other characters. II. Grouping into four types. J. Biochem. (Tokyo) *94*:997-1007.
18. **Ouzounis, C., C. Sander, M. Scharf, and R. Schneider.** 1993. Prediction of protein structure by evaluation of sequence-structure fitness. J. Mol. Biol. *232*:805-825.
19. **Pearson, W.R.** 1995. Comparison of methods for searching protein sequence databases. Protein Sci. *4*:1150-1160.
20. **Pearson, W.R. and D.J. Lipman.** 1988. Improved tools for biological sequence comparison. Proc. Natl. Acad. Sci. USA *85*:2444-2448.
21. **Pertsemlidis A. and J.W. Fondon, III.** 2001. Having a BLAST with bioinformatics (and avoiding BLASTphemy). Genome Biol. *2*:REVIEWS2002.
22. **Raghava, G.P.S.** 2001. A graphical web server for the analysis of protein sequences and alignment. Biotech. Software Internet Rep. *2*:255-258.
23. **Smith, T. and M. Waterman.** 1981. Identification of common molecular subsequences. J. Mol. Biol. *147*:195-197.
24. **Thompson, J.D., D.G. Higgins, and T.J. Gibson.** 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. *22*:4673-4680.
25. **Tomii, K. and M. Kanehisa.** 1998. A comparative analysis of ABC transporters in complete microbial genomes. Genome Res. *8*:1048-1059.
26. **Walchli, S., J. Colinge, and R. Hooft van Huijsduijnen.** 2000. MetaBlasts: tracing protein tyrosine phosphatase gene family roots from Man to *Drosophila melanogaster* and *Caenorhabditis elegans* genomes. Gene *253*:137-143.
27. **Zhang, Z., W. Pearson, and W. Miller.** 1997. Aligning a DNA sequence with a protein sequence. J. Comput. Biol. *4*:333-443.

The authors thank the Council of Scientific and Industrial Research (CSIR) and Department of Biotechnology (DBT), Government of India, for financial assistance. We also thank the developers of the servers/software integrated in GWFASTA. B.I. is a recipient of a fellowship from CSIR. This report has IMTECH communication no. 03/2002. Address correspondence to Dr. G.P.S. Raghava, Institute of Microbial Technology, Sector 39A, Chandigarh-160036, India. e-mail: raghava@imtech.res.in

Received 25 March 2002; accepted 13 June 2002.

Biju Issac and G.P.S. Raghava
Institute of Microbial Technology
Chandigarh, India