A Graphical Web Server for the Analysis of Protein Sequences and Alignment

G. P. S. Raghava
Institute of Microbial Technology

Introduction

The presentation of the physical properties of a protein sequence graphically along its primary structure (e.g., hydropathy plot) can play a vital role in understanding the function of a protein (Hofman & Stoffel, 1992; Combat et al., 2000). In the past, different parameters have been used for sliding-window analysis of protein sequences. Hofman & Stoffel (1992) developed an interactive graphical tool for protein sequence analysis called PROFILEGRAPH, which integrated most of the existing sliding-windows methods. PROFILEGRAPH allows a wide range of analysis, but it has some limitations. For example, it cannot handle sequence alignment. There are a number of Web servers at http://www.expasy.ch/tools/#primary to compute and present the physical properties of amino acid sequences graphically (Wilkins et al., 1998). Each Web server allows limited analysis on protein sequence, and none of the existing servers allows the user to plot the amino acid properties along the sequence alignment.

The alignment of protein sequences is a central aid in understanding the function of a protein. A number of programs have been developed to beautify/analyze the alignment of proteins, for example, ALSCRIPT, AMAS, and SOMAP (Barton, 1990, 1993). Recently, Easy Sequencing in Postscript (ESPript) has been developed to format the multiple alignment in postscript (Gouet et al., 1999). The existing methods do not allow for the computation and presentation of the physiochemical properties of amino acids at each position in alignment.

This paper describes a Web server called Protein Sequence Analysis Web Server (PSAweb) that has been developed for analysis of protein sequences and alignments in the tradition of PROFILEGRAPH, ALSCRIPT, and AMAS. This server provides a number of additional features that are not available in existing programs/servers. This report describes the applications of our server.

System and Methods

All forms were designed using HyperText Markup Language (HTML) and Javascript. All CGI scripts were written in PERL (version 5.03) programming language. The graphical output was generated in GIF format using a graphics library, GD.PM, from Lincoln D. Stein (Cold Spring Harbor Laboratory). HTML is used to present the results in tabular form. For input of protein sequences and alignments, the READSEQ program of Don Gilbert (University of Indiana) is used.

Applications

Analysis of Protein Sequences

Users can input their protein sequence to the server as a file(s) or by using cut and paste, in one of the standard sequence formats (e.g., PIR, Fasta) or by simple amino acids only. Users can select the type of display (graphical or text) in which the analysis will be presented before submitting the sequence for analysis (Fig. 1).

Amino Acid Property Plot. The server allows the user to plot amino acid properties along the protein's primary structure (e.g., hydrophilicity/flexibility plot). One can select up to 4 properties at a time out of 36 available properties to plot within a single window or in multiple windows. This feature of the server allows the user to study and compare various properties of the protein. Users can select (1) the length of the window used for averaging, (2) the moment method to be applied before averaging, and (3) the method used for averaging (e.g., mean over window; hat over window; median; data-sieve) (Hofman & Stoffel, 1992).

Interesting/Unusual Residues. One can highlight a group of residues in the sequence that has a specific property (e.g., hydrophobic residues, polar residues, and small residues). This feature of the server assists the user in understanding the overall nature of the protein. Users can select a residue or a set of residues to be highlighted in the protein sequence, to determine the position of interesting/unusual residues in the sequence.

Analysis of Sequence Alignment

In order to analyze the sequence alignment, the user has to input the alignment and the appropriate parameters to be computed. The server allows the user to input alignments in one of the standard formats (PIR, Fasta, MSF) or in BLOCK format.

Property Plot of Each Sequence. The server allows the user to plot the amino acid properties of protein sequences along the sequence alignment. It creates a separate graph corresponding to each sequence in the alignment and presents all the user-selected properties in this graph. This option is very useful in studying each sequence in an alignment and in identifying the aligned regions that are highly similar or dissimilar for a given physiochemical property.

Mean Property Plot with Standard Deviation. The server computes the amino acid property of each position in the alignment in two steps. In the first step, it calculates the amino acid property for each sequence in the alignment as described. In the second step, it determines the mean value of the amino acid property for each position in the alignment by the following equation.

$$A_i = \frac{\sum_{j=1}^{N} P_{ij}}{N}$$

Where A_i is the mean value for position i in the alignment, P_{ij} is the value of the parameter (amino acid property for a specified window length) at position i in alignment of sequence j, and N is the total number of sequences in alignment.

The server presents the mean value of parameter (A_i) along the alignment graphically, for example, hydrophobic plot of multiple alignment. The user can identify the interesting regions (e.g., hydrophobic, hydrophilic, flexible, surface) in the alignment from these plots. However, the mean/average plot of an alignment presents the overall characteristics of each position, but it provides no knowledge about similarity/dissimilarity in the sequences at that position. In order to measure magnitude of similarity or dissimilarity among sequences in the alignment, we compute the standard deviation (SD). The SD_i at a position i in the alignment can be calculated using the following equation.

$$SD_i = \sqrt{\frac{\sum_{j=1}^{N} (P_{ij} - A_i)^2}{N}}$$

The server presents the mean value and SD in a graphical form that allows the user to understand the overall nature and level of conservation of each position in the alignment.

Highlight Conserved Positions in Alignment. The user can set the similarity level above which residues are color-coded to highlight the conserved positions within the alignment. To compute conserved positions, the server first calculates the similarity scores (Sc) in all possible pairs of residues by using a user-specified weight matrix (BLOSUM62, PAM250, or identity scoring matrix). Then, it calculates the Sc for a given position in the alignment by the following equation.

$$Sc = \frac{\sum_{j=1}^{N-1} \sum_{k=j+1}^{N} S_{jk}}{N(N-1)/2}$$

where S_{jk} is the score of aligning a residue in sequence i with a residue in sequence j, using a similarity score matrix. Sc is the mean score at a position, and N is the total number of sequence in the alignment.

The server highlights all the residues at a given position if the score Sc is above a user-defined threshold at that position. The server also provides an option to highlight a residue or group of residues in the alignment that has a specific property.

Position-Specific Matrix. To study the composition of the amino acids at each position in the alignment, the server computes the position-specific matrix. Thus, the user can easily detect the consensus sequence in the alignment.

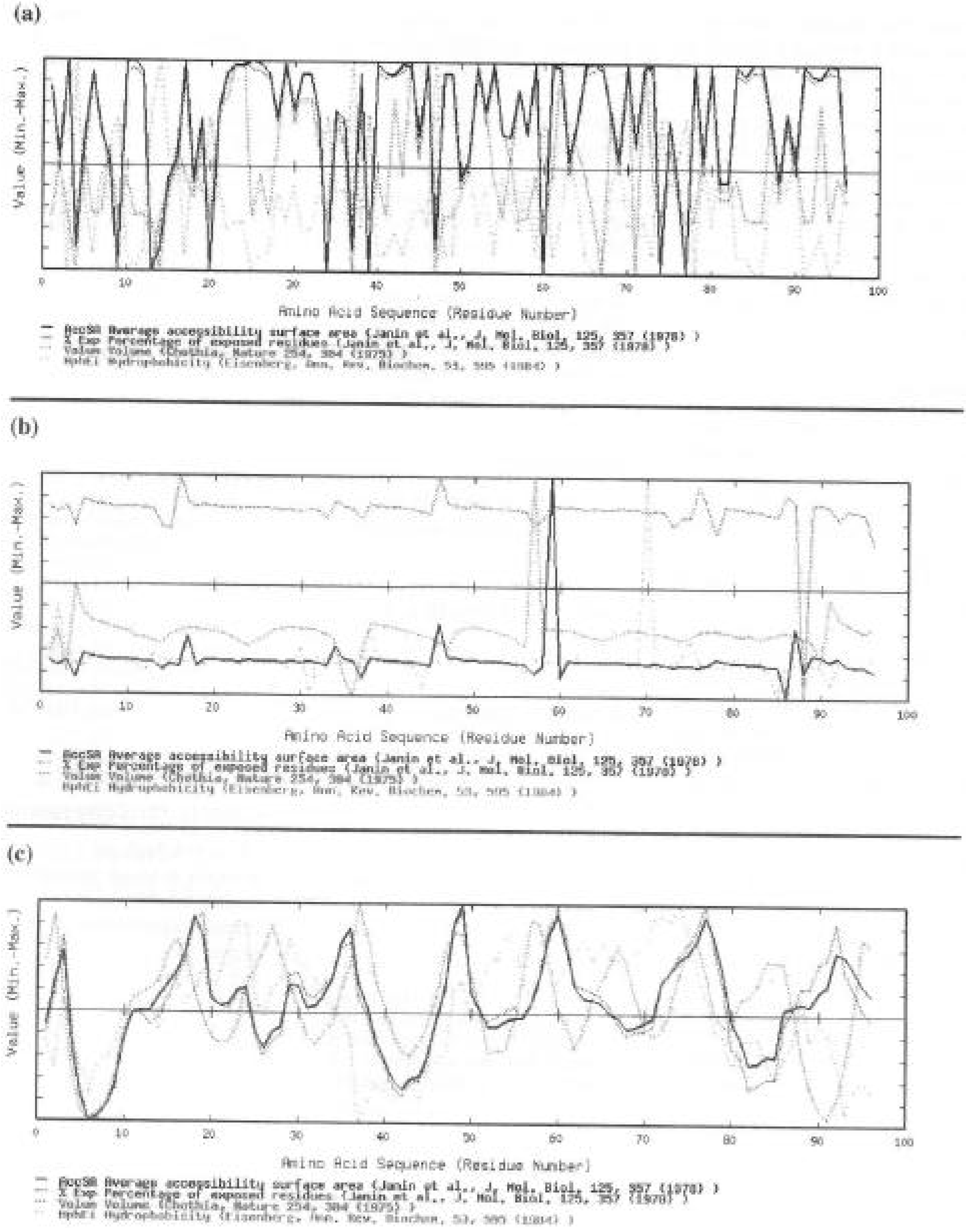

Figure 1. A plot shows the various properties of amino acids of GROES protein (Ch10_actac: 10 kDA Caperonin protein). Here, 'mean over window', is used for averaging with window size 7. a) no-moment method with window size 7; b) moment angle with window size 21; c) normalize moment with window size 21. This is a direct output of PSAweb server.

Sequence Similarity Among Sequences. Users can compute the percent similarity among the sequences in an alignment. This assists the user in clustering the sequences in the alignment and in identifying the consensus sequence that has maximum similarity with the rest of the sequences. It also computes the maximum, minimum, and average similarity of a query sequence to other sequences in the alignment.

Summary

A Web server has been developed to analyze the amino acid sequence and multiple sequence alignment of proteins. This is a comprehensive on-line Internet tool that allows for the rapid visualization of an analysis in GIF format. It assists the user in analyzing and presenting the primary structure of proteins. It (1) allows the computation of the physical properties of amino acids, (2) presents the properties along the primary structure of a protein in graphical and tabulated format, and (3) highlights a residue or group of residues in the sequence that exhibits a specific function. This Web server also allows for analysis of multiple sequence alignments of proteins and provides the following options: (1) compute and present the overall property of each position in the alignment, (2) highlight the conserved residues in the alignment, (3) highlight a residue in the alignment that exhibits a specific function, (4) compute a position-specific score matrix, and (5) compute the similarity among the sequences present in the alignment (Fig. 1).

In brief, PSAweb allows for the comprehensive analysis of protein sequences and alignments. It allows a user to analyze a protein sequence and presents the results graphically in a manner similar to PROFILEGRAPH and performs analysis/beautification/format of a multiple sequence alignment, similar to ALSCRIPT and ESPript.

Availability

The tool is available from http://www.imtech.res.in/raghava/psa/.

Contact Information

E-mail: raghava@imtech.res.in

Acknowledgments

This is communication No. 8/2000 from the Institute of Microbial Technology, Chandigarh, supported by grants from the Department of Biotechnology and Council of Scientific Industrial Research, Government of India.

References

1. Barton, G. J. (1990). Protein multiple sequence alignment and flexible pattern matching. *Methods in Enzymology* 183, 403–428.
2. Barton, G. J. (1993). ALSCRIPT a tool to format multiple sequence alignments. *Protein Eng.* 6, 37–40.
3. Combet, C., Blanchet, C., Geourjon, C. and Deleage, G. (2000). NSP@: Network protein sequence analysis network. *Trends in Biochem. Sci.* 25, 147–50.
4. Gouet, P., Courcelle, E., Stuart, D.I. and Metoz, F. (1999). ESPript: analysis of multiple sequence alignment in PostScript. *Bioinformatics* 15, 05–308.
5. Hofman, K. and Stoffel. W. (1992). PROFILEGRAPH: An interactive graphical tool for protein sequence analysis. *Comput. Applic. Biosci.* 8, 331–337.
6. Kyte, J. and Doolittle, R.F. (1982). A simple method for displaying hydrophobic character of protein. *J. Mol. Biol.* 157, 105–132.
7. Livingstone, C. L. and Barton, G.J. (1993). Protein sequence alignments: a strategy for the hierarchial analysis of residue conservation. *Comput. Applic. Biosci.* 9, 745–756.
8. Wilkins M.R., Gasteiger, E., Bairoch, A., Sanchez, J.C., Williams, K.L., Appel, R.D. and Hochstrasser D.F.(1998). Protein identification and analysis tools in the ExPASy server in: 2-D Proteome Analysis Protocols. Editor A.J. Link. Humana Press, New Jersey.

Address for Correspondence:

Dr. G P S Raghava
Scientist & Head, Bioinformatics Centre
Institute of Microbial Technology
Sector 39A, Chandigarh, INDIA
Fax: +91-172-690632 or 690585
Phone: +91-172-690557 or 695225
E-mail: raghava@imtech.res.in
Web: http://imtech.res.in/raghava/