DCBASE Vol. 2(1) 1-5, 2011 (July)

Prediction of a protein sequence and its function using bioinformatics tools

Mitra Kabir, A. S. Md. Mukarram Hossain and Upama Kabir

Department of Computer Science and Engineering, University of Dhaka, Bangladesh

E-mail: mukarram@cse.univdhaka.edu

Received on 09.03.2010. Accepted for publication on 02.08.2011

Abstract

One of the major challenges in bioinformatics is to characterize novel protein sequences. It is determined by making comparison with known protein sequences stored in protein sequence databases and family databases. Protein sequences are compared mainly by means of sequence alignments. Now days, various bioinformatics pattern-recognition methods have been developed to encode conserved regions in the sequence alignments. These conserved regions are used to characterize particular protein families, which in turn highlights the function and structure of member proteins. This study concerns an investigation of a range of bioinformatics services, databases and software available on the Web in characterizing a novel protein sequence and its function. The potency of these bioinformatics tools has been assessed by identifying a retinal protein bacteriorhodopsin (BR) of the archaeon *Halobacterium salinarum*. It is a transmembrane protein found in the cellular membrane of these organisms and functions as a light-driven proton pump.

Keywords: Protein prediction, Bioinformatics tool, Protein family, Sequence database, Bacteriorhodopsin.

1. Introduction

Proteins are the biological molecules that carry out essential functions in every cell of living organisms. Each protein has a unique amino acid sequence that is specified by the nucleotide sequence of a gene encoding this protein. These long amino acids chains fold into complex structures allowing them to perform functions. One of the main aims of bioinformatics is to characterize novel protein sequences so that they can be served as useful information for future diagnosis.

A novel protein sequence can be characterized by making comparison with the known sequences stored in sequence databases and family databases. In recent time, various bioinformatics tools have been developed in support of that. These tools are mainly based on sequence alignment including pairwise and multiple sequence alignment and various scoring matrices. The most frequently used method for identifying proteins related to a query sequence is to search a sequence database using pairwise alignment tools, such as the BLAST and FASTA programs [1, 2]. They find regions of local similarity between sequences. Again, various other tools exist that employ different pattern-recognition methods to encode conserved regions in the multiple sequence alignments of a family of proteins. These conserved regions are used to characterize particular signatures of a protein family, which usually relate to potential functionality or structural significances of member proteins. The family based databases basically fall into three broad categories depending on their use of single motif (PROSITE, eMOTIF), multiple motifs (BLOCKS, PRINTS), or full domain alignments (Profiles, Pfam) for encoding the conserved regions in the alignment. The single-motif based approaches, for example, perform well in diagnosing short, functional sites; multiple-motif dependent approaches work best in the diagnosis of families and subfamilies; and the domain-based methods which creates profile on that, tend to work well in the diagnosis of superfamilies.

Understanding the relationships between proteins is often not straightforward. This relationship is mainly dependent on sequence homology. The patterns of conservation change as the regions of similarity become smaller and smaller when aligning increasingly divergent sequences. Ultimately, it can be very difficult or even impossible to discover whether the alignment represents a single but highly divergent family with a common function, or a significant superfamily composed of several closely related but functionally isolated families. This represents one of the major challenges in bioinformatics now a day. In this study, a range of bioinformatics services, databases and softwares available on the Web have been utilized in analyzing a protein sequence. This includes identification of the protein family which serves as the signature of its potential function and structure. The potency of these bioinformatics tools has been assessed by the recognition of a known retinal protein *bacteriorhodopsin* (BR) of the archaeon *Halobacterium salinarum*.

2. Materials and methods

In this study, a 262 amino acids long protein sequence with accession no. P02945 in UniProtKB, version 15.12 (http://www.uniprot.org/uniprot/), is selected as the target sequence. The BLAST tools blastp and PSI-BLAST of version 2.2.22 in NCBI (http://www.ncbi.nlm.nih.gov/BLAST/) are used to find sequence similarities with default parameters and BLOSUM62 substitution matrix [3]. The target database used in these searches is non-redundant sequence database (nr). It is a composite database containing non-identical sequences from publicly available

primary sources: GenBank CDS translations [4], Swiss-Prot [5], PDB [6], PIR [7]. BLAST can identify generic similarities between sequences based on high scoring sequence pairs. BLAST search has also provided a crude means of classifying the sequence based on the similarity.

The family and function of the intended sequence are identified by searching the protein family databases. To locate known sequence patterns (regex), PROSITE 20.0 database, released on November 2008, is searched through ScanProsite tool of ExPasy proteomics server (http://www.expasy.ch/tools/scnpsite.html). By searching Pfam 24.0 database [8], using the web service maintained by Sanger institute (http://pfam.sanger.ac.uk/), the known protein domains (based on profiles) belonging to the selected sequence are identified. In order to validate the results obtained from PROSITE and Pfam, PRINTS 39_0 database (based on fingerprints) [9] is searched through FingerPrintScan (Using FPScan option) suit (http://www.bioinf.manchester.ac.uk/fingerPRINTScan/) with default parameters and BLOSUM62 matrix. Fingerprints are groups of conserved ungapped motifs that are obtained from multiple sequence alignments and intended to derive potent signatures of family membership through iterative database scanning [10]. Additionally, the integrated protein family database i.e., InterPro [11] has also been investigated providing a more coherent view of the results.

The potential transmembrane (TM) domains in the selected sequence are highlighted by means of hydropathy profiles using ProtScale program of ExPasy proteomics server (http://www.expasy.ch/tools/protscale.html). JalView, an alignment editor (http://www.jalview.org/examples/applets.html), with automatic multiple alignment program CLUSTALW [12], is also used to represent multiple alignment of the selected protein with the other sequences of its family.

3. Results and discussion

In this section, the potency of different bioinformatics tools to characterize the target sequence as the bacteriorhodopsin protein is presented along with sufficient details. This investigation includes prediction of the protein family as well as prediction of its function along with some structural relevancy.

The target sequence is searched against nr(Non-redundant) database using the blastp program of NCBI. This has generated more than 50 significant hits with E-value smaller than 1×10^{-70}. All of these sequences belong to bacterial opsin type proteins (obtained from annotation). With the highest E-value 6×10^{-175} (95% identity), the protein bacteriorhodopsin (BACR_HALSA) in *Halobacterium salinarumhas* has been diagnosed as the most significant hit. In order to validate this search result, the iterative BLAST search program PSI-BLAST (based on PSSM) is also used. This BLAST search also diagnosed the same result. But, it is not always true that the top most statistically significant hit from BLAST search is also the most biologically relevant. So, it is likely that the query is a bacteriorhodopsin protein which belongs to bacteria opsin family.

In order to obtain more biologically relevant results, protein family databases searching are the most important means. By aligning a new sequence with the representative family members, it is possible to identify conserved regions which ensure its family membership and in turn indicate its critical structural significance or functional roles. Here, four different databases- PROSITE, PRINTS and InterPro have been searched to characterize the target sequence.

Through PROSITE search, two different hits are found by two distinct patterns. The first hit with accession no. PS00950 and identifier BACTERIAL_OPSIN_1 indicates the bacterial opsins signature. The other hit with accession no. PS00327 and identifier BACTERIAL_OPSIN_RET represents the bacterial opsins retinal binding site. These two patterns are very important in allowing the specific detection of bacterial opsins. These proteins are retinal-containing proteins found in extremely halophilic bacteria providing light-dependent proton transport and sensory functions. From this result, it is identified that the query protein is a bacterial opsins type protein. But there are at least three types of bacterial opsins including bacteriorhodopsin, halorhodopsin and sensory rhodopsin having the same intended patterns. So from this PROSITE search result, it is not possible to say specifically to which type of bacterial opsin, the query belongs. A more specific result is obtained by Pfam search (dependent on Hidden Markov Model and Profiles). It returned a match which represents a bacteriorhodopsin-like protein Domain (accession no. PF01036 and identifier Bac_rhodopsin). From this, it can be said that it is a bacteriorhodopsin protein.

Based on the database annotation, it is also known that all the bacterial opsins including bacteriorhodopsin are integral membrane proteins containing 7 helix transmembrane (TM) domains, the last of which contains the attachment point for a photoreactive chromophore, retinal [13]. That is, bacteriorhodopsin has a seven-helix structural motif, which ensures its belongingness to 7TM superfamily. The PRINTS database search can provide more efficient evidence supporting this.

Scan of sequence: USER_SEQUENCE

Highest scoring fingerprints for your query

Fingerprint	E-value	GRAPHScan	Motif3D
BACTRLOPSIN (relations)	8.593599e-62	Graphic	

Ten top scoring fingerprints for your query

Fingerprint	No. of Motifs	SumId	AveId	PfScore	Pvalue	Evalue	GRAPHScan	
BACTRLOPSIN	7 of 7	3.6e+02	51	3720	1.5e-67	8.6e-62	IIIIIII	Graphic
RHESUSRHD	2 of 12	56.12	28.06	367	4.6e-05	1.1e-02	iI.....	Graphic
DNAGYRASEB	2 of 11	73.33	36.67	348	4.9e-05	1.7e-02	..I.....I..	Graphic
ADENOVSFIBRE	2 of 7	69.82	34.91	321	6.9e-05	2.3e-02	..II...	Graphic
PNDRDTASEII	2 of 10	52.44	26.22	293	0.00014	3.9e-02	...i.....i	Graphic
EAGCHANLFMLY	2 of 10	62.43	31.22	402	0.00015	4.6e-02	...I.I....	Graphic
TYPE3IMPPROT	2 of 6	42.61	21.31	334	0.0004	7.2e-02	.i...i	Graphic
GLYCHORMONER	2 of 8	50.57	25.28	367	0.00041	1.2e-03	....i.I.	Graphic
TAP1PROTEIN	2 of 12	77.62	38.81	314	0.00076	2.1e-03	....I.I.....	Graphic
YEAST73DUP	2 of 13	48.25	24.13	210	0.0011	3.2e-03	..i...i.....	Graphic

Fig. 1: Searched output returned by FingerPrintScan

The algorithm employed by FingerPrintScan to search PRINTS provides rank-ordered hits based on combined motif E-values. For the query sequence, the PRINT search has revealed only one diagnosis named BACTRLOPSIN (accession no. PR00251) which is a 7-element (motif) fingerprint for the bacterial opsins. This fingerprint is also a signature of having 7 membrane-spanning domains or transmembrane α-helices (20-24 residues long each). A better appreciation of what such diagnose means is gained by plotting graphical view of this match as shown in Figure 2. Wherever a motif matches above a given threshold, a shaded block is plotted to mark its location, from N to C-terminus, indicating matches with each of the seven transmembrane domains.

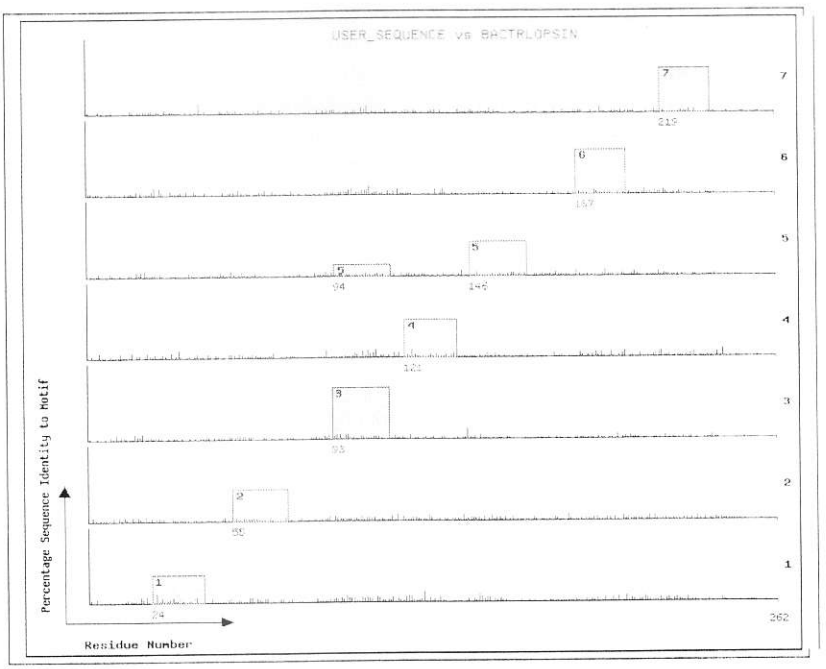

Fig. 2: GRAPHScan plots showing the results of scanning the selected sequence against fingerprint BACTRLOPSIN (The bacterial opsin family signature). The Y-axis represents identity score (in percentage) of each fingerprint element (0–100 per motif), whereas the X-axis denotes the amino acid number in the sequence. A shaded block is drawn with the width of the motif and with the height of the % identity of the match wherever a motif matches.

A more coherent biological interpretation (for functional and structural relevancy) of Figure 2 can be found from Figure 3, which represents the multiple alignments of sequences from bacterial opsin family.

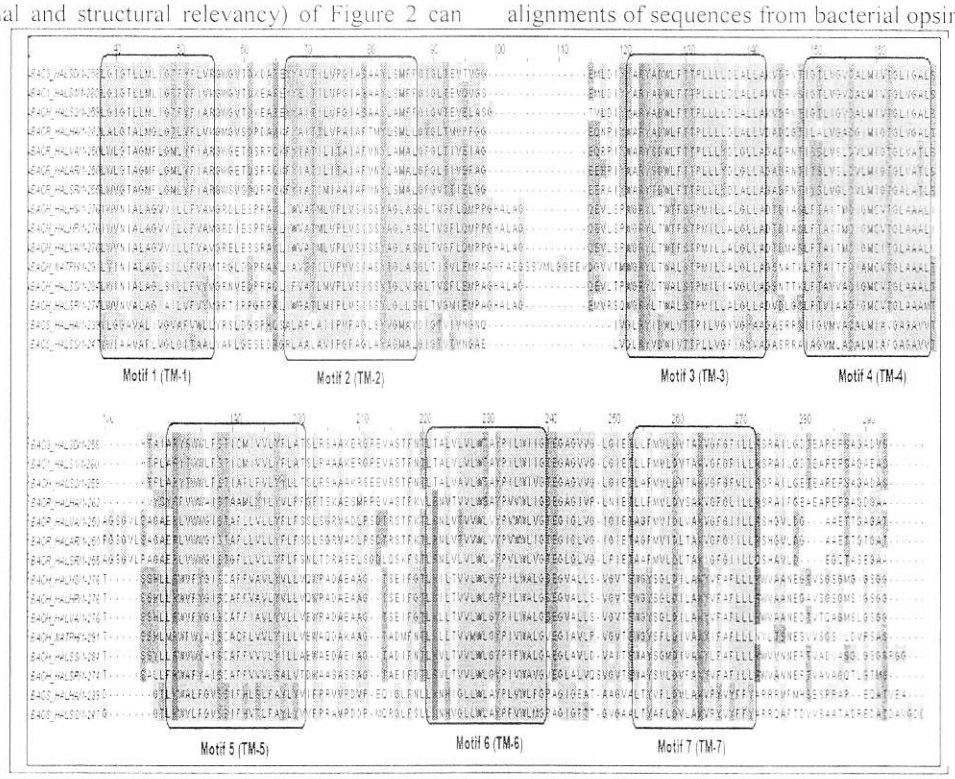

Fig. 3: Multiple alignment of a set of bacterial opsin protein sequences including BACR_HALSA protein. The motifs labelled as Motif 1- Motif 7 belong to fingerprint BACTRLOPSIN. This alignment is obtained from JalView.

Based on the multiple alignments shown in Figure 3, one major feature that is observed is the occurrence of seven relatively hydrophobic regions in the primary sequences that contain large hydrophobic residues (residues marked as blue). The periodic distribution of hydrophobicity is consistent with an alpha-helical conformation. This feature provides that the query contains seven transmembrane helices (core region). It is also proved through the hydropathy profiles (Kyte & Doolittle) of the sequence generated with ProtScale. As shown in Figure 3, motif 3 (called helix C), the region encoded by PROSITE pattern BACTERIAL_OPSIN_RET, includes a conserved functionally important arginine (R) residue which seems to be involved in the release of a proton from the Schiff base to the extracellular medium [14]. The seventh transmembrane domain (called helix G), encoded by PROSITE pattern BACTERIAL_OPSIN_RET, includes a conserved retinal binding lysine (K) residue. That is, the retinal chromophore is attached as a Schiff base to Lysine in this putative helix G. Proline (P) residues in transmembrane α helices are the general structural feature of integral membrane proteins. Bacteriorhodopsin folds into a seven transmembrane helix topology with short interconnecting loops due to the presence of proline residues. From Figure 3, it is observed that TM-2, TM-3 and TM-7 i.e. the membrane spanning helices B, C and F contain conserved proline residues in the middle. A proline residue in the middle of an alpha-helix produces a bend in the helix (provides flexibility in protein backbone) which aids folding. That is, these residues help the helix to adopt the optimal disposition to establish important helix-helix packing interactions. As a result, these contribute to protein folding and forming retinal binding pocket. Helix C, D and G contain conserved aspartic acid (D) residues which are important for proton translocation to the retinal Schiff base [15]. Amino acids on TM-3 and TM-7 (helix C and G) form the components of a hydrogen bonding network that provides a pathway along which the proton is translocated. This path contains charged residues.

Additionally, InterPro database search has validated the above identification with accession no. IPR001425 and IPR018229. Thus from all the results obtained from BLAST search as well as protein family database analysis, it is concluded that, the query sequence is a bacteriorhodopsin protein structurally having a topology of seven transmembrane (TM) helices and functions as a light-driven proton-translocating pump by converting the energy of protons into an electrochemical potential.

4. Conclusion

Proteins are the biological workhorses that carry out essential functions (individually or as a group) in every cell of living organisms. A major challenge in bioinformatics is to characterize novel protein sequences i.e., determining their families which aid the signature for their potential function and structure. In this study, the pairwise and family based search methods have proven their effectiveness in recognizing a known retinal bacteriorhodopsin (BR) protein (from its amino acid sequence) which functions as a light-driven proton pump in halobacteria. But this task would become difficult and more challenging when characterizing a more divergent member of a family. Still, the existing tools could not characterize a large number of unknown proteins. These challenges need improved diagnostic tools and analysis with a hope that in the near future all these unknown proteins having significant functionality on living organisms would be characterized significantly. Further, this result can be served as useful information in developing various bioinformatics diagnosis.

References

1. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D. J, 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res., 25(17)*. 3389-3402.
2. Pearson, W. R., 2000. Flexible sequence similarity searching with the FASTA3 program package. *Methods Mol. Biol.*, **132**, 185-219.
3. Henikoff, J. G. and Henikoff, S., 1992. Amino acid substitution matrices from protein blocks. ***PNAS,* 22(89), 10915-10919.**
4. Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., Rapp, B. A. and Wheeler, D. L., 2000. GenBank. *Nucleic Acids Res.*, **28**(1), 15-18.
5. Bairoch, A. and Apweiler, R., 2000. The SWISS-PROT protein sequence database and its supplement TrEMBL. *Nucleic Acids Res.*, **28**(1), 45-48.
6. Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. and Bourne, P. E., 2000. The Protein Data Bank. *Nucleic Acids Res.*, **28**(1), 235-242.
7. Barker, W. C., Garavelli, J. S., Hou, Z., Huang, H., Ledley, R. S., McGarvey, P. B., Mewes, H. W., Orcutt, B. C., Pfeiffer, F., Tsugita, A., Vinayaka, C. R., Xiao, C., Yeh, L. S. and Wu, C., 2001. Protein Information Resource: a community resource for expert annotation of protein data. *Nucleic Acids Res.*, **29**(1), 29-32.
8. Bateman, A., Birney, E., Cerruti, L., Durbin, R., Etwiller, L., Eddy, S. R., Griffiths-Jones, S., Howe, K. L., Marshall, M. and Sonnhammer, E. L., 2002. The Pfam protein family's database. *Nucleic Acids Res.*, **30**(1), 276-280.
9. Attwood, T. K., Avison, H., Beck, M. E., Bewley, M., Bleasby, A. J., Brewster, F., Cooper, P., Degtyarenko, K., Geddes, A. J., Flower, D. R., Kelly, M. P., Lott, S., Measures, K. M., Parry-Smith, D. J., Perkins, D. N., Scordis, P., Scott, D. and Worledge, C., 1997. The PRINTS database of protein fingerprints: a novel information resource for computational molecular biology. *J. Chem. Inf. Comput. Sci.*, **37**(3), 417-424.
10. Attwood, T. K. and Findlay, J. B., 1993. Fingerprinting G-protein-coupled receptors. Protein Eng., 7(2), 195-203.
11. Apweiler, R., Attwood, T. K., Bairoch, A., Bateman, A., Birney, E., Biswas, M., Bucher, P., Cerutti, L., Corpet, F., Croning, M. D., Durbin, R., Falquet, L., Fleischmann, W., Gouzy, J., Hermjakob, H., Hulo, N., Jonassen, I., Kahn, D., Kanapin, A., Karavidopoulou, Y., Lopez, R., Marx, B., Mulder, N. J., Oinn, T. M., Pagni, M. and Servant, F., 2001. The InterPro database, an integrated documentation resource for protein families, domains and functional sites. *Nucleic Acids Res.*, **29**(1), 37-40.
12. Thompson, J. D., Higgins, D. G. and Gibson, T. J., 1994. CLUSTALW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22**(22), 4673-4680.
13. Oesterhelt, D., Tittor, J. and Bamberg, E., 1992. A unifying concept for ion translocation by retinal proteins. *J. Bioenerg. Biomem.*, **24**, 181-191
14. Mogi, T., Marti T., and Khorana, G., 1989. Structure-Function Studies on Bacteriorhodopsin. *J. Biol. Chem.*, **24**(264), 14197-14201.
15. Soppa, J., Otomo, J., Straub, J., Tittor, J., Meessen, S., Oesterhelt, D., 1989. Bacteriorhodopsin mutants of Halobacterium sp. GRB. II. Characterization of mutants. *J. Biol. Chem.*, **264**(22), 13049–13056.