RESEARCH PAPER | Asian Journal of Bio Science, Volume **9** | Issue 1 | April, 2014 | 71-75
Received : 06.02.2014; **Revised :** 12.03.2014; **Accepted :** 22.03.2014

Study of transmembraneous protein using bioinformatics and data mining

SHAHEEN JAMAL **AND SUNITA**

Department of Biotechnology, Lovely Professional University, PHAGWARA (PUNJAB) INDIA
Email : nidhusunita@yahoo.com

Membrane proteins perform diverse functions in living organisms such as transporters, receptors and channels. The functions of membrane proteins have been investigated with several computational approaches, such as developing databases, analyzing the structure function relationship and establishing algorithms to discriminate different type of membrane proteins. However, compilation of bioinformatics resources for the functions of membrane proteins is not well documented compared with their structural aspects. The purpose of the present work was to assess the study of transmembraneous protein using bioinformatics and data mining. Bioinformatics is the application of information technology to the field of molecular biology. Protein structure prediction is the important application of bioinformatics. Bioinformatics also provide researchers with software's and tools for analyzing the sequence data and deriving biologically meaningful information from a string of letters. By using application of bioinformatics one can predict isoelectric point, molecular weight, transmembrane helix and secondary structure of transmembrane protein.

Key words : Membrane proteins, Bioinformatics, Data mining, Transmembrane helix

How to cite this paper : Jamal, Shaheen and Sunita (2014). Study of transmembraneous protein using bioinformatics and data mining. *Asian J. Bio. Sci.*, **9** (1) : 71-75.

INTRODUCTION

The study of membrane proteins is one of prime importance in all branches of proteomics. They are known to play crucial roles in various cellular functions. Information about their function can be derived from their structure, but knowledge of these proteins is limited, as their structures are difficult to obtain (Caffrey, 2003). A wide range of essential cellular functions are mediated by membrane proteins. For example, the exchange of membrane impermeable molecules between organelles and between a cell and its extracellular environment are facilitated by channels and pumps. In addition, transmembrane receptors sense changes in the environment and commence specific cellular responses typically via their associated proteins. Membrane proteins are also of great diagnostic and therapeutic importance, so that they are targets of >50 per cent of all current drugs (Stagljar and Fields,2002; Ge *et al*., 2003; Clapham and Neer, 1997).

Membrane proteins are integral to all cellular functions acting as mediators between the cell and its environment. A transmembrane protein (TP) is an integral membrane protein (*i.e.*, proteins that penetrate into or through the membrane bilayer) that spans from the internal to the external surface of the biological membrane or lipid bilayer in which it is embedded (Li *et al.*, 2004).Transmembrane proteins have three regions or domains that can be defined: the domain in the bilayer, the domain outside the cell (called the extracellular domain), and the domain inside the cell (called the intercellular domain). Many transmembrane proteins function as gateways or "loading docks" to deny or permit the transport of specific substances across the biological membrane, to get into the cell, or out of the cell as in the case of waste byproducts. As a response to the shape of certain molecules these "freight handling" transmembrane protein may have special ways of folding up or bending that will move a substance through the biological membrane (Kosugi *et al.*, 1994; Hordijk *et al.*, 1994). It is a polytopic protein that spans an entire biological membrane. Transmembrane proteins aggregate and precipitate in water. They require detergents or nonpolar solvents for extraction; although some of them (beta-barrels) can be also extracted using denaturing agents (Berman *et al.*, 2000). These remarkable proteins play important roles in energy transduction, cell signaling, and maintaining the integrity of the cells' internal environment. However, there is still very

little known about their function since many of their structures remain unknown. Since structure leads to function, discovering the structure of these proteins will help lead to understanding their function and will aid in creating drugs for a host of diseases. They play several roles in the functioning of cells (Kosugi *et al.*, 2001). Bioinformatics and data mining play very important role in the prediction of protein structure. It provides software's and tools for analyzing the sequence data and deriving biologically meaningful information from a string of letters (Bordner, 2009; Tusnady *et al.*, 2004; Bradford and Westhead, 2005).

RESEARCH METHODOLOGY

A transmembrane protein was collected for this study. The nucleotide and protein sequences were retrieved from National Center for Biotechnology Information (NCBI) databases. Protein sequence was submitted to NCBI (National Center for Biotechnology Information). BLAST (Basic Local Alignment Search Tool) to find homologous sequence with known structure and function, Protparam Compute (pI/Mw) to calculate isoelectric point (pI) and molecular weight (Altschul *et al.*, 1997). Protscale to know the hydrophobicity and presence of transmembraneous domains in the protein. The homologus sequence of proteins were submitted to clustal web server to do multiple sequence alignment.

Software used :

Protparam :

Protparam used for the prediction of physico-chemical parameters of a protein sequence (amino-acid and atomic compositions, pI, extinction co-efficient, etc.). Protein molecular weight is used to predict the location of a protein of interest on a gel in relation to a set of protein standards. At a pH below their isoelectric point, proteins carry a net positive charge; above their isoelectric point they carry a net negative charge. Proteins can thus, be separated according to their isoelectric point (overall charge) on a polyacrylamide gel using a technique called isoelectric focusing, which uses a pH gradient to separate proteins. Isoelectric focusing is also the first step in 2-D gel polyacrylamide gel electrophoresis.

Compute pI/Mw :

Compute pI/Mw is a tool which allows the computation of the theoretical pI (isoelectric point) and Mw (molecular weight) for a list of UniProt Knowledge base.

Prot scale analysis :

The protscale is used for the view of hydrophobicity. Hydrophobicity predicts the spatial arrangement of amino acid in proteins. Protscale allow to compute and represent the profile produced by any amino acid scale on selected protein. An amino acid scale is defined by a numerical value assigned to each type of amino acid. The most frequently scale are hydrophobicity or hydrophilicity (Tusnady *et al.*, 2005; Qin *et al.*, 2007).

Prediction of secondary structure of proteins :

Secondary structure prediction is a set of techniques in bioinformatics that aim to predict the local secondary

Table A : Software's used for the prediction of secondary structure

Name of software for prediction of secondary structure	Accuracy	Comments
PHDsec: http://www.emblheidelberg.de/predictprotein/	>72% (+/10%, one standard deviation)	Multiple alignment-based neural network system (Filmore, 2004; Gilman, 1987)
NSSP: http://dot.imgen.bcm.tmc.edu:9331/pssprediction/pssp.html	>71%.Evaluated on >200 unique proteins.	Multiple alignment-based nearest-neighbor method (Filmore, 2004; Gilman, 1987)
SOPM: http://www.ibcp.fr/predict.html	>70%.	Multiple alignment-based method combining various other prediction(Filmore, 2004; Gilman, 1987)
HNN (Hierarchical Neural Network)	70%	More successful in predicting alpha helices than beta sheets, regions(King *et al.*, 2003)
SSPRED: http://www.embl-heidelberg.de/sspred/ssp_mul.html	>70%.	Multiple alignment-based program using statistics. (Vapnik, 1998)
MultiPredict: http://kestrel.ludwig.ucl.ac.uk/zpred.html	>65%	Multiple alignment-based method using physico-chemical information from a set of aligned sequences and statistical secondary structure decision constants (Palczewski *et al.*, 2000)
PSA: http://bmerc-www.bu.edu/psa/	>70%	The PSA server analyzes amino acid sequences to predict secondary structures and folding classes (Li, *et al.*, 2004; Chang *et al.*, 2006; Negi *et al.*, 2007)
NNPREDICT: http://www.cmpharm.ucsf.edu/~nomi/nnpredict.html	>65%	Single-sequence based neural network prediction (Li *et al.*, 2004)

structures of proteins and RNA sequences based only on knowledge of their primary structure - amino acid or nucleotide sequence, respectively. For proteins, a prediction consists of assigning regions of the amino acid sequence as likely alpha helices, beta strands (often noted as "extended" conformations), or turns. The success of a prediction is determined by comparing it to the results of the dictionary of protein secondary structure (DSSP) algorithm applied to the crystal structure of the protein. The dictionary of protein secondary structure method is commonly used to describe the protein secondary structure with single letter codes as shown in Table A (Neuvirth, 2004; Landau *et al.*, 2005).

RESEARCH FINDINGS AND ANALYSIS

The prediction of physico-chemical parameters of a protein sequence *i.e.* is amino-acid and atomic compositions, isoelectric point(pI), extinction co-efficient, etc. are important to co-relate the protein structure and function in a biological system. By using software (Protpram and Compute pI) the pI and molecular weight of the protein was found out to be ~ 8.67 and 37876.9, respectively (Table 1).

Table 1: Physico-chemical parameters of protein sequence

Parameter	Protparam	Compute pI/Mw
Number of amino acids	332	332
Molecular weight	37876.9	37876.93
Theoretical pI	8.67	8.67

Phosphorylated sites on the protein :

Knowledge about the phosphorylation site gives an idea about the regulation and activity of the protein by phosphorylation. The results show that the protein molecule contains eight site for the phosphorylation (Table 2). Also phosphorylated protein results showed decrease in the isoelectric point when number of phosphate group increase because addition of negative charge on this protein.

Table 2 : Phosphorylated sites on the protein

# Phosphates	Molecular Weight	Isoelectric Point
0	45076.6853	8.27
1	45154.6493	7.80
2	45232.6133	7.33
3	45310.5773	7.03
4	45388.5413	6.83
5	45466.5053	6.68
6	45544.4693	6.55
7	45622.4333	6.44
8	45700.3973	6.34

Protscale analysis :

Hydrophobicity predicts the spatial arrangement of amino acid in proteins. The Fig. 1 showing the total six peaks of hydrophobicity.

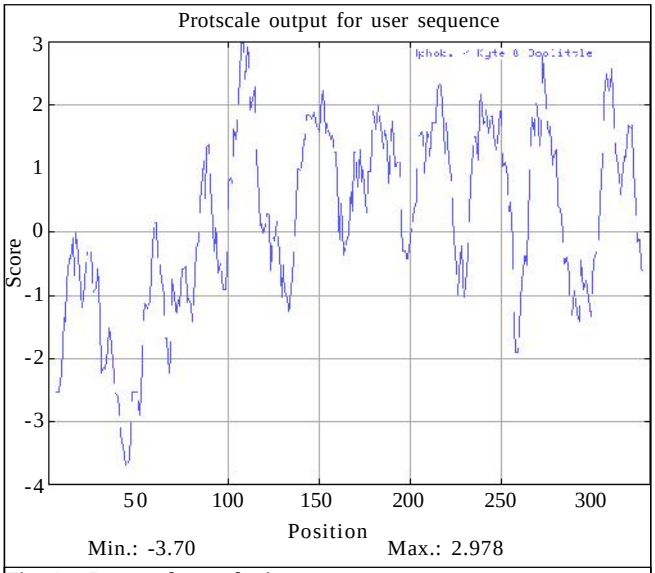

Fig. 1 : Protscale analysis

Prediction of transmembrane helices in protein:

Transmembrane proteins are associated with controlling the exchange of materials across the membrane. Therefore, prediction of the transmembraneous helices (Fig. 2) would be important to study how the different material and signal pass through these proteins.

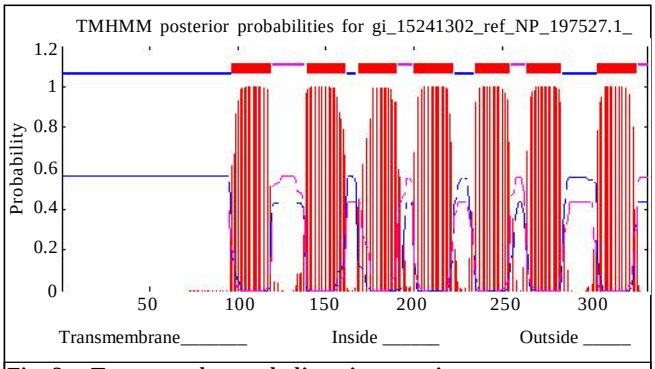

Fig. 2 : Transmembrane helices in protein

TMHMM Server v. 2.0

gi_15241302_ref_NP_197527.1_ Length: 332

gi_15241302_ref_NP_197527.1_ Number of predicted TMHs: 7

gi_15241302_ref_NP_197527.1_ Exp number of AAs in TMHs: 153.57402

gi_15241302_ref_NP_197527.1_ Exp number, first 60 AAs: 5e-05

gi_15241302_ref_NP_197527.1_ Total prob of N-in: 0.56496
gi_15241302_ref_NP_197527.1_ TMHMM2.0 inside 1 96
gi_15241302_ref_NP_197527.1_ TMHMM2.0 TMhelix 97 119
gi_15241302_ref_NP_197527.1_ TMHMM2.0 outside 120 138
gi_15241302_ref_NP_197527.1_ TMHMM2.0 TMhelix 139 161
gi_15241302_ref_NP_197527.1_ TMHMM2.0 inside 162 167
gi_15241302_ref_NP_197527.1_ TMHMM2.0 TMhelix 168 190
gi_15241302_ref_NP_197527.1_ TMHMM2.0 outside 191 199
gi_15241302_ref_NP_197527.1_ TMHMM2.0 TMhelix 200 222
gi_15241302_ref_NP_197527.1_ TMHMM2.0 inside 223 234
gi_15241302_ref_NP_197527.1_ TMHMM2.0 TMhelix 235 254
gi_15241302_ref_NP_197527.1_ TMHMM2.0 outside 255 263
gi_15241302_ref_NP_197527.1_ TMHMM2.0 TMhelix 264 283
gi_15241302_ref_NP_197527.1_ TMHMM2.0 inside 284 303
gi_15241302_ref_NP_197527.1_ TMHMM2.0 TMhelix 304 326
gi_15241302_ref_NP_197527.1_ TMHMM2.0 outside 327 332
Number of transmembrane helices: 7
Transmembrane helices: 97-118 139-160 175-194 203-222 235-254 263-284 305-324
Total entropy of the model: 17.0107
Entropy of the best path: 17.0122.

Conclusion :

In this study it is concluded that the protein has a isoelectric point(pI) and molecular weight ~ 8.67 and 37876.9, respectively. It has seven transmembrane therefore, it is present in the membrane with N-terminus inside. Also, it can concluded that this protein may be playing an important role either in the transport of the solutes and molecules or biological signals across the membrane.

LITERATURE CITED

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z. and Miller, W. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.,* **25**(17) : 3389–3402.

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. *Nucl. Acids Res.,* **28** : 235–242.

Bordner, A.J. (2009). Predicting protein-protein binding sites in membrane proteins. *BMC Bioinformatics,***10** : 312.

Bradford, J.R. and Westhead, D.R. (2005). Improved prediction of protein-protein binding sites using a support vector machines approach. *Bioinformatics*, **21**(8) : 1487–1494.

Caffrey, M. (2003). Membrane protein crystallization. *J. Structural Biol.,* **142** : 108–132.

Chang, D.T., Weng, Y.Z., Lin, J.H., Hwang, M.J. and Oyang, Y.J. (2006). Protemot : prediction of protein binding sites with automatically extracted geometrical templates. *Nucleic Acids Res.,* **34** (Web Server issue) : W303–309.

Clapham, D.E. and Neer, E.J. (1997). G Protein ßγ Subunits. *Annu. Rev. Pharmacol. Toxicol*., **37** : 167–203.

Filmore, David (2004). "It's a GPCR world". Modern Drug Discovery (American Chemical Society) 2004 (November): 24–28.

Ge, H., Walhout, A.J. and Vidal, M. (2003). Integrating 'omic' information: a bridge between genomics and systems biology. *Trends Genet.*, **19**(10) : 551–560.

Gilman, A.G. (1987). G Proteins: Transducers of Receptor-Generated Signals. *Annu. Rev. Biochem.,* **56** : 615-649.

Hordijk, P.L., Verlaan, I., Van Corven, E.J. and Moolenaar, W.H. (1994). Protein tyrosine phosphorylation induced by lysophosphatidic acid in Rat-1 fibroblasts. Evidence that phosphorylation of map kinase is mediated by the Gi-p21ras pathway. *J. Biol. Chem.*, **269** : 645–651.

King, N., Hittinger, C.T. and Carroll, S.B. (2003). Evolution of key cell signaling and adhesion protein families predates animal origins. *Science,* **301** (5631): 361–363.

Kosugi, S., Hines, J., Heerding, J.N., Fluharty, S.J. and Yee, D.K. (2001). Identification of Angiotensin II Type 2 (AT2) Receptor Domains Mediating High-Affinity CGP 42112A Binding and Receptor Activation. *J. Pharmacol. Exp.,* **298**(2) : 665-673.

Landau, M., Mayrose, I., Rosenberg, Y., Glaser, F., Martz, E. and Pupko, T. and Con surf (2005). The projection of evolutionary conservation scores of residues on protein structures. *Nucleic Acids Res.,* **33** (Web Server issue) : W299–302.

Li, J., Edwards, P.C., Burghammer, M., Villa, C. and Schertler, G.F.X. (2004). Structure of bovine rhodopsin in a trigonal crystal form. *J. Mol. Biol.,* **343**(5) : 1409-1438.

Negi, S.S., Schein, C.H., Oezguen, N., Power, T.D. and Braun, W. (2007). InterProSurf: a web server for predicting interacting sites on protein surfaces. *Bioinformatics*., **24** : 3397–3399.

Neuvirth, H., Raz, R. and Schreiber, G. (2004). ProMate: a structure based prediction program to identify the location of protein-protein binding sites. *J. Mol. Biol*., **338**(1) : 181–199.

Palczewski, K., Kumasaka, T., Hori, T., Behnke, C.A., Motoshima, H., Fox., B.A., Le Trong, I., Teller, D.C., Okada, T., Stenkamp, R.E., Yamamoto, M. and Miyano, M. (2000). Crystal structure of rhodopsin: A G protein-coupled receptor. *J. Sci.,* (**289)**: 739-745.

Qin, S., Zhou, H.X. and meta-PPISP (2007). A meta web server for protein-protein interaction site prediction. *Bio-informatics*., **23**(24) : 3386–3387.

Stagljar, I. and Fields, S. (2002). Analysis of membrane protein interactions using yeast-based technologies. *Trends Biochem. Sci.,* **27**(11) : 559–563.

Tusnady, G.E., Dosztanyi, Z. and Simon, I. (2004). Transmembrane proteins in the protein data bank: *Bioinformatics*., **20**(17) : 2964–2972.

Tusnady, G.E., Dosztanyi, Z. and Simon, I. (2005). Web server for detecting transmembrane regions of proteins by using their 3D coordinates. *Bioinformatics*., **21**(7) : 1276–1277.

Vapnik, V. (1998).*The nature of statistical learning theory*. New York: John Wiley and Sons, pp 158-160.