The Internet Journal of Genomics and Proteomics
Volume 4 Number 2

Function assignment to Influenza B virus proteins using Support Vector Machine- A clue for vaccine development

B Rathi, A Sarangi

Citation

B Rathi, A Sarangi. *Function assignment to Influenza B virus proteins using Support Vector Machine- A clue for vaccine development*. The Internet Journal of Genomics and Proteomics. 2008 Volume 4 Number 2.

Abstract

Purpose: Identification of protein functions of Influenza B virus to understand the biological processes and networks of the virus which lead to novel way of diagnostics and vaccine designing. Methods: The protein sequences of Influenza B virus was retrieved from NCBI and Swiss Protein Databank. Prediction and analysis of different protein function family of proteins of Influenza B virus was done by the software, SVMProt, Support vector machine which classifies a protein into functional families from its primary sequence based on physico-chemical properties of amino acids.Results: The studies from SVMProt on Influenza B proteins suggest quite disparity of protein functions between structural and non-structural classes of proteins. ATP Binding cassette is a protein function shared by both the classes of proteins. Functions performed exclusively by Structural proteins are aptamer binding, these proteins belong to major facilitator family and transferases- transferring one carbon groups, whereas non-structural proteins performs functions like nuclear receptor and zinc-binding.Discussion and Conclusion: Protein function predictied by SVMProt is different for structural and non-structural protein of Influenza B virus, some of which may be responsible for virulence or pathogenicity of the virus and others for replication of the virus in the host. Functional classification of the proteins helps to facilitate the study of various biological processes and search for new therapeutic targets for vaccine designing.

INTRODUCTION

Influenza B virus is exclusively a human pathogen. The only other animal known to be susceptible to influenza B virus infection is the seal [1]. The virus mutates at a rate 2–3 times lower than Influenza A virus and consequently is genetically less diverse, with only one influenza B virus serotype. As a result of this lack of antigenic diversity, a degree of immunity to influenza B is usually acquired at an early age. However, influenza B mutates enough that immunity is not lasting [2]. Influenza Type B viruses are not differentiated into subtypes but several strains are known. The Influenza Type B viruses can cause morbidity and mortality in humans.[34]

The Influenza B virus capsid is enveloped while its virion consists of an envelope, a matrix protein, a nucleoprotein complex, a nucleocapsid, and a polymerase complex. It is sometimes spherical and sometimes filamentous. It has nearly 500 surface projections, which are made of hemagglutinin and neuraminidase [5]. The influenza B virus genome is 14648 nucleotides long and consists of eight segments of linear negative-sense, single-stranded RNA . The multipartite genome is encapsidated, each segment in a separate nucleocapsid and the nucleocapsids are surrounded by one envelope [5].

The influenza B virus genome consists of 8 separate segments covered by the nucleocapsid protein. Together, these build the ribonucleoprotein (RNP) and each segment codes for a functionally important protein [6]. These proteins are Polymerase B2 protein (PB2) , Polymerase B1 protein (PB1) , Polymerase A protein (PA) ,Hemagglutinin (HA or H) , Nucleocapsid protein (NP) ,Neuraminidase (NA or N) , Matrix protein (M): M1 and M2 proteins together constitute the matrix protein. The matrix protein M2 (BM2) plays a vital role in virus assembly. Non-structural protein NS1 and NS2.

The RNA-dependent RNA polymerase of influenza virus is composed of three viral P proteins (PB1, PB2, and PA) and involved in both transcription and replication of the RNA genome. The PB1 subunit plays a key role in both the assembly of three P protein subunits and the catalytic function of RNA polymerization. [7].

Hemagglutinin(HA) is a glycoprotein containing either 2 or 3 glycosylation sites. It spans the lipid membrane , HA serves as a receptor by binding to sialic acid (N-acetylneuraminic acid) and have antigenic sites. Antigenic drift occurs in HA.

The Nucleoprotein (NP) is translated from segment 5 and is named after its major function, which is to bind and protect RNA.. NP and RNA together constitute the RNP. NP contains 498 residues and has a nuclear localization signal that allows the protein to actively migrate to the nucleus.

Like HA, Neuraminidase (NA) is a glycoprotein, which is also present as projections on the surface of the virus. The NA molecule presents its main part at the outer surface of the cell, spans the lipid layer. NA acts as an enzyme, cleaving sialic acid from the HA molecule, from other NA molecules and from glycoproteins and glycolipids at the cell surface. It also serves as an important antigenic site, and in addition, seems to be necessary for the penetration of the virus through the mucin layer of the respiratory epithelium. Antigenic drift can also occur in the NA.

Non-structural protein NS1 inhibit splicing of pre-mRNA and is probably able to suppress the interferon response in the virus-infected cells leading to unimpaired virus production [8]. Whereas NS2 function is believed to be to facilitate the transport of newly synthesized RNPs from the nucleus to the cytoplasm to accelerate virus production [6].

Among the two structural protein M1 and BM2 , M1 forms a coat inside the viral envelope. It binds to the viral RNA and binding to ribonucleocapsids (RNPs) in nucleus seems to inhibit viral transcription. The second structural protein BM2 protein is a proton selective ion channel protein, integral in the cell membrane of the influenza virus and also plays a crucial role in the virus assembly of Influenza B virus [9].

The purpose of the study is to classify the different Influenza B virus proteins into protein functional families. SVMProt used in the study assign certain functional properties to each protein. Novel vaccine candidates can be prepared targeting the functions of these proteins.

MATERIAL AND METHODS

RETRIEVAL OF THE PROTEIN SEQUENCES OF INFLUENZA B VIRUS

The protein sequences of Influenza B virus was retrieved from NCBI (www.ncbi.nlm.nih.gov) and Swiss Protein Databank (http://us.expasy.org/sprot).

PREDICTION AND ANALYSIS OF PROTEIN FAMILY FUNCTION

Prediction and analysis of different protein function family of proteins of Influenza B virus was done by the software, SVMProt [10], Support vector machine (SVM) which classifies a protein into functional families from its primary sequence based on physico-chemical properties of amino acids. SVMProt shows a certain degree of capability for the classification of distantly related proteins and homologous proteins of different function and thus is used as a protein function prediction tool that complements sequence alignment methods .

SVMProt classification system is trained from representative proteins of a number of functional families and seed proteins of Pfam curated protein families. The protein functional families included in SVMProt are families of enzymes from BRENDA [11], G-protein coupled receptors from GPCRDB [12], nuclear receptors from NucleaRDB [12], tyrosine receptor kinases derived from NCBI [15], families of channels and family of transporters from TCDB [13] and LGICdb [14] and DNA- and RNA-binding proteins derived from SWISS-PROT [16].

Scoring of SVM classification of proteins has been estimated by a reliability index and its usefulness has been demonstrated by statistical analysis [17]. R-Value is a scoring function for estimating the accuracy of support vector machine classification. It is defined as: where d is the distance between the position of the vector of a classified protein and the optimal separating hyperplane in the hyperspace. P-Value is expected classification accuracy (probability of correct classification). It is derived from the statistical relationship between the R-value and actual classification accuracy based on the analysis of 9,932 positive and 45,999 negative samples of proteins. As in the case of all discriminative methods, the performance of SVMProt classification can be measured by the quantity of true positives (TP), true negatives (TN), false positives (FP), false negatives (FN) and the overall accuracy

(Q) given below:

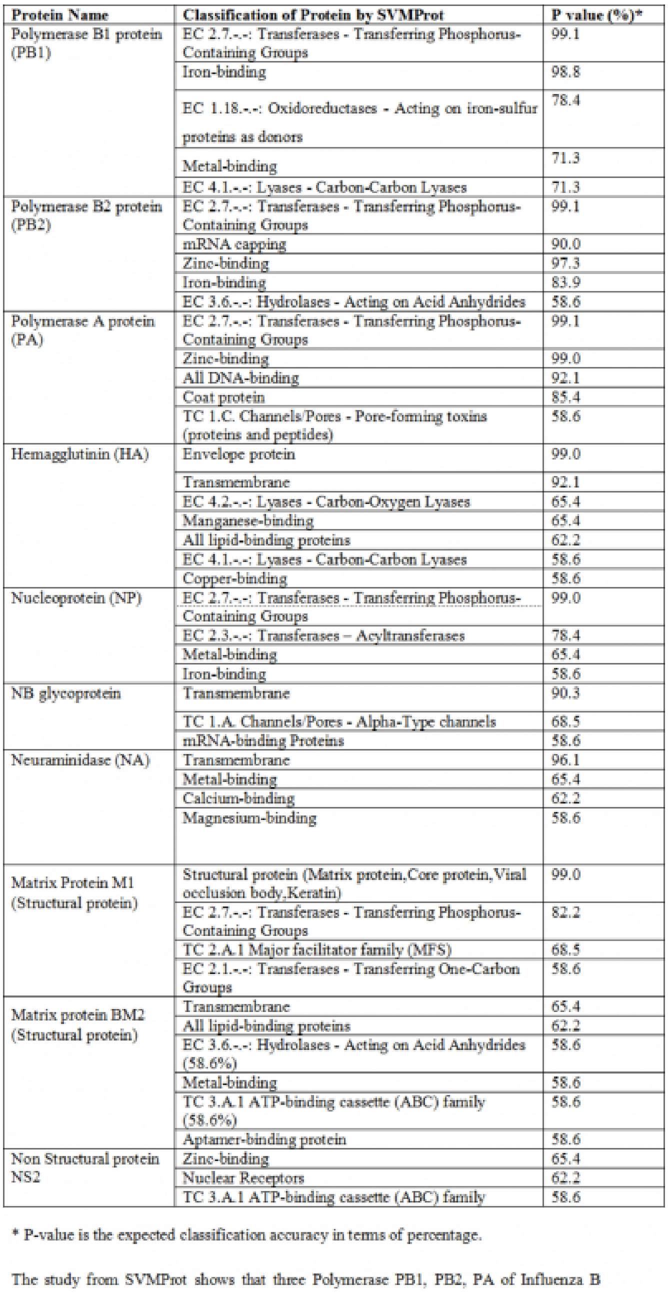
$Q = TP + TN / TP + TN + FN + FP.$

RESULT AND DISCUSSION

There has been no computational analysis done for the proteins of Influenza B virus. The study assigns the role of all the proteins of the virus. The functional analysis helps to specify the role of the proteins in the life cycle of the virus. The protein function family prediction done by SVMProt is given in the Table 1 for the proteins of Influenza B virus.

Figure 1

Table 1: SVMProt predicted protein function families for the different proteins of Influenza B virus.



Protein Name	Classification of Protein by SVMProt	P value (%)*
Polymerase B1 protein (PB1)	EC 2.7.-.-: Transferases - Transferring Phosphorus-Containing Groups	99.1
	Iron-binding	98.8
	EC 1.18.-.-: Oxidoreductases - Acting on iron-sulfur proteins as donors	78.4
	Metal-binding	71.3
	EC 4.1.-.-: Lyases - Carbon-Carbon Lyases	71.3
Polymerase B2 protein (PB2)	EC 2.7.-.-: Transferases - Transferring Phosphorus-Containing Groups	99.1
	mRNA capping	90.0
	Zinc-binding	97.3
	Iron-binding	83.9
	EC 3.6.-.-: Hydrolases - Acting on Acid Anhydrides	58.6
Polymerase A protein (PA)	EC 2.7.-.-: Transferases - Transferring Phosphorus-Containing Groups	99.1
	Zinc-binding	99.0
	All DNA-binding	92.1
	Coat protein	85.4
	TC 1.C. Channels/Pores - Pore-forming toxins (proteins and peptides)	58.6
Hemagglutinin (HA)	Envelope protein	99.0
	Transmembrane	92.1
	EC 4.2.-.-: Lyases - Carbon-Oxygen Lyases	65.4
	Manganese-binding	65.4
	All lipid-binding proteins	62.2
	EC 4.1.-.-: Lyases - Carbon-Carbon Lyases	58.6
	Copper-binding	58.6
Nucleoprotein (NP)	EC 2.7.-.-: Transferases - Transferring Phosphorus-Containing Groups	99.0
	EC 2.3.-.-: Transferases – Acyltransferases	78.4
	Metal-binding	65.4
	Iron-binding	58.6
NB glycoprotein	Transmembrane	90.3
	TC 1.A. Channels/Pores - Alpha-Type channels	68.5
	mRNA-binding Proteins	58.6
Neuraminidase (NA)	Transmembrane	96.1
	Metal-binding	65.4
	Calcium-binding	62.2
	Magnesium-binding	58.6
Matrix Protein M1 (Structural protein)	Structural protein (Matrix protein,Core protein,Viral occlusion body,Keratin)	99.0
	EC 2.7.-.-: Transferases - Transferring Phosphorus-Containing Groups	82.2
	TC 2.A.1 Major facilitator family (MFS)	68.5
	EC 2.1.-.-: Transferases - Transferring One-Carbon Groups	58.6
Matrix protein BM2 (Structural protein)	Transmembrane	65.4
	All lipid-binding proteins	62.2
	EC 3.6.-.-: Hydrolases - Acting on Acid Anhydrides (58.6%)	58.6
	Metal-binding	58.6
	TC 3.A.1 ATP-binding cassette (ABC) family (58.6%)	58.6
	Aptamer-binding protein	58.6
Non Structural protein NS2	Zinc-binding	65.4
	Nuclear Receptors	62.2
	TC 3.A.1 ATP-binding cassette (ABC) family	58.6

* P-value is the expected classification accuracy in terms of percentage.

The study from SVMProt shows that three Polymerase PB1, PB2, PA of Influenza B

The study from SVMProt shows that three Polymerase PB1, PB2, PA of Influenza B virus shows multiple functions in which certain functions are shared by the three polymerases and the other are performed individually by the three subunits of polymerases. The common functions are Transferases activity- transferring Phosphorous containing groups, iron binding and zinc binding. Analysis of PB1 shows that it is metal binding, acts enzyme in Oxidoreductases - acting on iron-sulfur proteins as donors (78.4%), Lyases - Carbon-Carbon Lyases (71.3%). PB2 is involved in mRNA capping (99%), enzyme for Hydrolases - acting on Acid Anhydrides. The acidic subunit of polymerase PA is all DNA binding (92.1%), it may also act as Channels/Pores - Pore-forming toxins (proteins and peptides) (58.6%).

Analysis of Hemagglutinin reveals it acts as envelope protein (99.0%), transmembrane (92.1%), that it is involved in manganese-binding (65.4%), all lipid binding (65.2%), Copper-binding (58.6%), lyases - carbon-carbon lyases. Nucleoprotein is metal binding, iron binding, Transferases - transferring phosphorus-containing groups, Transferases – acyltransferases.

The influenza B virus NB glycoprotein is abundantly expressed at the surface of virus-infected cells.NB spans the membrane once and has an 18 amino acid ectodomain, a 22 amino acid transmembrane domain, and a 60 amino acid cytoplasmic tail [18].The functional analysis by SVMProt also reveals it as transmembrane protein (90.3%). NB is mRNA-binding protein (58.6%)and involved in Channels/Pores - alpha-type channels (68.5%). SVMProt analysis Neuraminidase to be transmembrane protein (96.1%) with metal binding(65.4%), calcium binding (62.2%), magnesium binding properties (58.6%).

Matrix protein M1 is classified as structural protein (99.0%) having enzymatic activity of Transferases - transferring phosphorus-containing groups, Transferases - transferring one-carbon groups. Whereas matrix protein BM2 is transmembrane protein (65.4%) which is all lipid binding, metal binding, aptamer binding and have ATP -binding cassette (ABC) family. Non-structural protein NS2 acts a nuclear receptor and shows zinc- binding property. Protein family predicted by SVMProt shows Nonstructural and structural protein are different just share one common function of ATP -binding cassette (ABC) family.

From the above analysis it is evident that there is network of functions performed by Influenza B virus which accounts for the replication of the virus in the host and the virulence and pathogenesis of the virus. Knowing the biological function and the processes would help to understand the biology of the virus which will lead for better diagnostics and vaccine development.

ACKNOWLEDGEMENTS

The Bioinformatics Facility at our institution, where the work was done, is supported by the Indian Council of Medical Research, New Delhi and the Department of Biotechnology, Government of India, New Delhi.

CORRESPONDENCE TO

Bhawna Rathi, Biomedical Informatics Centre, Sanjay Gandhi Postgraduate Institute of
Medical Sciences, Lucknow 226014, India
Bhawna Rathi (MSc), Aditya N Sarangi (MSc)
Email : bhawna_rathi3@rediffmail.com

References

1. Osterhaus, A; Rimmelzwaan G, Martina B, Bestebroer T, Fouchier R ."Influenza B virus in seals." Science 2000. 288 (5468): 1051–3.
2. R, Webster; Bean W, Gorman O, Chambers T, Kawaoka Y . "Evolution and ecology of influenza A viruses.". Microbiol 1992. 56 (1): 152–79.
3. Carrat, F., and A. J. Valleron. 1995. Influenza mortality among the elderly in France, 1980-90: how many deaths may have been avoided through vaccination? J. Epidemiol. Community Health 49:419-425.
4. Sullivan, K. M., A. S. Monto, and I. M. Longini, Jr. 1993. Estimates of the US health impact of influenza. Am. J. Public Health
5. Büchen-Osmond, C. (Ed). ICTVdB Virus Description -00.046.0.04. Influenzavirus B. ICTVdB - The Universal Virus Database, 2006 version 4. Columbia University, New York, USA.
6. Nicholson KG, Webster RG, Hay AJ. Textbook of Influenza. Blackwell Science, Oxford, 1998.
7. Honda, A., Mizumoto, K., Ishihama, A..Minimum molecular architectures for transcription and replication of the influenza virus. Proc. Natl. Acad. Sci. USA 2002. 99: 13166-13171.
8. Wetherall NT, Trivedi T, Zeller J, Hodges-Savola C, McKimm-Breschkin JL, Zambon M, Hayden FG. Evaluation of neuraminidase enzyme assays using different substrates to measure susceptibility of influenza virus clinical isolates to neuraminidase inhibitors: report of the neuraminidase Inhibitor susceptibility network. J Clin Microbiol 2003; 41: 742-750.
9. Horwarth, C.M., M.A. Williams, and R.A. Lamb, .Eukaryotic Coupled translation of tandem cistrons: identification of the Influenza B viruse BM2 polypeptide.EMBO J. 1990 9:2639-2647.
10. C.Z. Cai, L.Y. Han, Z.L. Ji, X. Chen, Y.Z. Chen. SVM-Prot: Web-Based Support Vector Machine Software for Functional Classification of a Protein from Its Primary Sequence. Nucleic Acids Res.2003 , 31: 3692-3697
11. Schomburg,I., Chang,A. and Schomburg,D. BRENDA, enzyme data and metabolic information. Nucleic Acids Res.2002, 30,47–49.
12. Horn,F., Vriend,G. and Cohen,F.E. Collecting and harvesting biological data: the GPCRDB and NucleaRDB information systems. Nucleic Acids Res.2001, 29, 346–349.
13. Saier,M.H. Jr A functional-phylogenetic classification system for transmembrane solute transporters.Microbiol.Mol. Biol. Rev.2000, 64, 354–411.
14. Le Novere,N. and Changeux,J.-P. LGICdb: the ligand-gated ion channel database. Nucleic Acids Res.2001, 29, 294–295.
15. Wheeler,D.L., Church,D.M., Federhen,S., Lash,A.E., Madden,T.L., Pontius,J.U., Schuler,G.D., Schriml,L.M., Sequeira,E., Tatusova,T.A. and Wagner,L. Database resources of the National Center for Biotechnology. Nucleic Acids Res.2003, 31,28–33.
16. Boeckmann,B., Bairoch,A., Apweiler,R., Blatter,M.-C., Estreicher,A., Gasteiger,E., Martin,M.J., Michoud,K., O'Donovan,C., Phan,I. et al. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Res., 31, 365–370.
17. Hua,S.J. and Sun,Z.R. A novel method of protein secondary structure prediction with high segment overlap measure: support vector machine approach. J. Mol. Biol.2001, 308, 397–407.
18. Brassard DL, Leser GP, Lamb RA. Influenza B virus NB glycoprotein is a component of the virion. Virology. 1996 15;220(2):350-60

Author Information

Bhawna Rathi
Biomedical Informatics Centre, Sanjay Gandhi Postgraduate Institute of Medical Sciences

Aditya N. Sarangi
Biomedical Informatics Centre, Sanjay Gandhi Postgraduate Institute of Medical Sciences