A Support Vector Machine Approach for assigning novel function to HPV Proteins

A Sarangi, B Rathi

Citation

A Sarangi, B Rathi. *A Support Vector Machine Approach for assigning novel function to HPV Proteins*. The Internet Journal of Infectious Diseases. 2008 Volume 7 Number 2.

Abstract

Human papillomaviruses (HPVs) are small, non-enveloped double-stranded DNA viruses. HPV infection is associated with more than 90% of cervical cancer, which is the second leading cause of cancer death among women worldwide. Identification of functions of different proteins will help enable to understand the mechanism of HPV infection and will provide further clues to vaccine development. Analysis of the HPV protein sequences by SVMPort insinuate that the structural and nonstructural proteins of the HPV genome possibly belong to diverse protein functions anticipated to play a role in the life cycle of HPV. Protein function common to both structural and nonstructural proteins are zinc binding, calcium binding, and metal binding. Structural proteins also performs function like coat protein, zinc binding, copper binding and calcium binding. Non structural proteins performs function like zinc binding, copper binding calcium binding, magnesium binding, metal binding, DNA binding, DNA repair, DNA replication, repressors, antigen, cell adhesion, actin binding, immune response, DNA condensation. Nonstructural proteins also perform functions like pore forming toxins and participate in type II (general) secretary pathway.

INTRODUCTION

Human Papillomaviruses (HPVs) cause a diverse range of epithelial lesions and are the prime cause of cervical cancer, which is the second most widespread cancer in women worldwide. HPV genome contains a double-stranded DNA with about 8Kb nucleotides which encodes eight open reading frames and is organized into three regions: early genes (E1 to E7), late genes (L1 and L2) and the noncoding long control region (LCR) which regulates viral replication and gene expression. Early genes codes for the six non structural proteins (E1, E2, E4, E5, E6, and E7) each of which contribute to the viral replication and late genes codes for the two structural proteins (L1 and L2).

E1 is a nuclear phosphoprotein having ATPase and helicase activity [1] which binds to the origin of replication in the LCR [2], where it associates with DNA polymerase alpha and initiates transcription [3]. E2 is also a nuclear phosphoprotein and act as a DNA binding transcription factor interacting with ACCN6GGT motifs in the viral LCR [4]. E4 is a nonstructural protein which is expressed in the later stages of infection after completion of the virion assembly. The functions of the E4 protein is unknown, although this protein does associate with keratin intermediate filaments [67] as well as forming cytoplasmic and nuclear inclusions [5] which probably interfere with normal epithelial differentiation, allowing the virus to complete its life cycle. The E5 protein is a small, hydrophobic, membrane-associated protein which can regulate cell signaling pathways, [8] interfere with cell-cell communication [9] and associate with surface growth factor receptors [10]. E5 protein is distributed predominantly in the endoplasmic reticulum, the golgi and the cytoplasmic membrane [11] and can activate EGFR. Activation of EGFR by E5 leads to over expression of a variety of proto-oncogenes [12]. E6 is the primary oncogene of HPV which is expressed in the early stages in the viral life cycle. Expression of E6 blocks apoptosis, alters the transcription machinery, disturbs cell – cell interaction, increases the life span of the cell. E7 is a nuclear protein and is divided in to three domains CR1, CR2 and CR3. E7 interacts with retinoblastoma tumor suppressor suppresser family proteins (Rb, p107, p130, HDAC, TATA box binding protein (TBP), cyclins, cyclin dependent kinases and cdk. These interactions results in increased cell proliferation, immortalization and finally transformation of the epithelial cells [13].

HPV genes are differentially expressed both temporally and spatially throughout the infectious cycle. Functional significance of HPV proteins in different stages of its

lifecycle has not yet been determined experimentally. Hence in this study an attempt is made to find the novel function of different structural and non structural proteins of HPV from its primary amino acid sequence. By using SVMPort we were able to assign certain functional property to each protein. Novel therapeutic vaccine candidate can be prepared targeting the function of these proteins.

METHODOLOGY

RETRIEVAL OF THE PROTEIN SEQUENCES OF HPV

The primary protein sequences of HPV 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

The web based software SVMPort was used for the prediction and analysis of different protein function families of the structural and non structural proteins of HPV. Protein function prediction has immense importance in studying the underlying biological processes. SVMPort is based on Support Vector Machine algorithm 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 [14].

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 [15], G-protein coupled receptors from GPCRDB [16], nuclear receptors from NucleaRDB [16], tyrosine receptor kinases derived from NCBI [19], families of channels and family of transporters from TCDB [17] and LGICdb [18] and DNA- and RNA-binding proteins derived from SWISS-PROT [20].

Scoring of SVM classification of proteins has been estimated by a reliability index and its usefulness has been demonstrated by statistical analysis [21]. 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 (24,35), 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

No computational analysis has been done so far for the functional assignment of different proteins of HPV. The insilico functional analyses of HPV proteins help enable us understanding the role of different proteins in the life cycle of the virus.

With reference to protein function family detection by SVMProt, HPV proteins belong to different function families which are given in Table 1.

{image:1}

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

Functional analysis using SVMPort infer that the late gene products major (L1) and minor (L2) capsid proteins shows multiple functions. The capsid proteins functions as coat protein. The function common to the capsid protein is zinc binding activity. Analysis of L1 protein shows that it functions as coat protein (99.0%), involved in zinc binding (98.5%), metal binding (68.5%), copper binding (58.6%) and calcium binding (58.6%). The minor capsid protein (L2) functions as coat protein (97.5%), involved in zinc binding (85.4%) and all lipid binding proteins (65.4%).

E1 protein a nuclear phosphoprotein having ATPase and helicase activity [1], plays a role in replication and replication repressor [22]. Analysis of E1 protein by SVMPort reveals that it is an outer membrane protein, plays a major role in DNA replication (99.0%), having all DNA binding activity (98.6%) and may involve in DNA repair (58.6%) and belongs to zinc binding, metal binding, calcium binding, magnesium binding protein function families.

E2 is also a nuclear phosphoprotein and acts as regulator of

viral transcription and replication, control of early region viral gene expression, necessary for viral DNA replication together with E1 [23]. Analysis of E1 protein by SVMPort reveals that it is involved in viral replication (98.4%), act as repressor (98.2%) and belongs to zinc binding, all DNA binding and all lipid binding protein function families.

E4 is a nonstructural protein which is expressed as a late gene primarily in differentiating epithelium. This protein does associate with keratin intermediate filaments [67] as well as forming cytoplasmic and nuclear inclusions [5] which probably interfere with normal epithelial differentiation, allowing the virus to complete its life cycle. Analysis of E4 protein by SVMPort reveals that this protein belongs to iron binding, all lipid binding, and copper binding protein function families. It may act as an antigen in the host (58.6%) and may have cell adhesion properties. It has also toxin like pore forming properties which is a novel function derived by the analysis using SVMPort.

Analysis of the E5 protein by SVMPort shows that it is a transmembrane like protein (58.6%) and may act as an antigen in the host (58.6%). It can also act as pore forming toxins (proteins and peptides) and participate in type II (general) secretory pathway (IISP) family.

E6 protein contains two zinc finger binding motifs [24]. Analysis of the E6 protein by SVMPort shows that it may be a motor protein and belongs to zinc binding and actin binding protein function families. Analysis of E7 protein by SVMPort shows that this protein plays a role in immune response and may help in DNA condensation.

CONCLUSION

Cervical cancer is a world wide public health problem among women. To improve the control of cervical cancer new diagnostic and therapeutic strategies are required. Integration of immunology and proteomic biotechnology with computational tools like SVM has accelerated the understanding of the genetic and cellular basis of many cancer types. Protein function family predicted by SVMProt is different for each structural and non-structural protein of HPV, some of which may be responsible for virulence or pathogenicity of the virus and others for replication of the virus in the host. Prediction of the functional roles of pore-forming toxins is important for facilitating the study of various biological processes and the search for new therapeutic targets.

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.

References

1. Hughes FJ, Ramanos MA. E1 protein of human papillomavirus is a DNA helicase/ATPase. Nuc Acids Res 1993; 21:5817-5823.
2. Sarafi TR and McBride AA. Domains of BPV-1 E1 replication protein required for origin-specific DNA binding and interaction with the E2 transactivator. Virology 1995; 211:385-396.
3. Park P, Copeland W, Yang L. The cellular DNA polymerase alpha-primase is required for papillomavirus DNA replication and associates with the viral E1 helicase. Proc Natl Acad Sci USA 1994; 91:8700-8704.
4. McBride AA, Romanczuk H, Howley PM. The papillomavirus E2 regulatory proteins. J Biol Chem 1991;266:18411-4
5. Doorbar J. The E4 proteins and their role in the viral life cycle. In:Lacey C, ed. Papillomavirus reviews: Current Research on Papillomaviruses. Leeds: Leeds University Press, 1996:31-38.
6. Roberts S, Ashmole I, Rookes SM et al. Mutational analysis of the Human papillomavirus type 16 E1-E4 protein shows that the C terminus is dispensable for keratin cytoskeleton association but is involved in inducing disruption of the keratin filaments. J Virol 1997; 71:3554-3562.
7. Doorbar J, Ely S, Sterling J et al. Specific interaction between HPV-16 E1-E4 and cytokeratins results in collapse of the epithelial cell intermediate filament network. Nature 1991; 352:824-827
8. Ghai J, Ostrow RS, Tolar J et al. The E5 gene product of rhesus papillomavirus is an activator of endogenous Ras and phosphatidylinositol-3'-kinase in NIH 3T3 cells. Proc Natl Acad Sci USA 1996; 93:12879-12884.
9. Oelze I, Kartenbeck J, Crusius K et al. Human papillomavirus type 16 E5 protein affects cell-cell communication in an epithelial cell line. J Virol 1995; 69:4489-4494.
10. Hwang ES, Nottoli T, Dimaio D. The HPV16 E5 protein: Expression, detection, and stable complex formation with transmembrane proteins in COS cells. Virology 1995; 211:227-233.
11. Oetke C, Auvinen E, Pawlita M, Alonso A. Human papillomavirus type 16 E5 protein localizes to the Golgi apparatus but does not grossly affect cellular glycosylation. Arch Virol 2000; 145:2183-91.
12. Pim D, Collins M, Banks L. Human papillomavirus type 16 e5 gene stimulates the transforming activity of the epidermal growth factor receptor. Oncogene 1992;7:27-32
13. Furukawa T, Duguid WP, Rosenberg L et al. Long-term culture and immortalization of epithelial cells from normal adult human pancreatic ducts transfected by the E6E7 gene of human papilloma virus 16. Am J Pathol 1996; 148:1763-1770.
14. 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
15. Schomburg,I., Chang,A. and Schomburg,D. BRENDA, enzyme data and metabolic information. Nucleic Acids

Res.2002, 30,47–49.
16. 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.
17. Saier,M.H. Jr A functional-phylogenetic classi cation system for transmembrane solute transporters.Microbiol.Mol. Biol. Rev.2000, 64, 354–411.
18. Le Novere,N. and Changeux,J.-P. LGICdb: the ligand-gated ion channel database. Nucleic Acids Res.2001, 29, 294–295.
19. 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.
20. 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.
21. 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.
22. Doobar J. The papillomavirus life cycle. J Clin Virol 2005;32 Suppl 1:S7-15
23. Chiang CM, Dong G, Broker TR, Chow LT. Control of human papillomavirus type 11 origin of replication by the E2 family of transcription regulatory proteins. J Virol 1992;66:5224-31.
24. Lipari F, McGibbon GA, Wardrop E, Cordingley MG. Purification and biophysical characterization of a minimal functional domain of an N-terminal Zn+2 –binding fragment from the human papillomavirus type 16 E6 protein. Biochemistry 2001;40:1196-204

Author Information

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

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