Available online at www.sciencedirect.com

SCIENCE DIRECT

Virology 331 (2005) 136–143

VIROLOGY

www.elsevier.com/locate/yviro

Prediction of functional class of novel viral proteins by a statistical learning method irrespective of sequence similarity

L.Y. Han[a], C.Z. Cai[a,b], Z.L. Ji[c], Y.Z. Chen[a,*]

[a]*Bioinformatics and Drug Design Group, Department of Computational Science, National University of Singapore, Block SOC1, Level 7, 3 Science Drive 2, Singapore 117543, Singapore*

[b]*Department of Applied Physics, Chongquing University, Chongquing 400044, PR China*

[c]*Department of Biology, School of Life Sciences, Xiamen University, Xiamen 361000, FuJian Province, PR China*

Received 17 July 2004; returned to author for revision 15 September 2004; accepted 9 October 2004
Available online 5 November 2004

Abstract

The function of a substantial percentage of the putative protein-coding open reading frames (ORFs) in viral genomes is unknown. As their sequence is not similar to that of proteins of known function, the function of these ORFs cannot be assigned on the basis of sequence similarity. Methods complement or in combination with sequence similarity-based approaches are being explored. The web-based software SVMProt (http://jing.cz3.nus.edu.sg/cgi-bin/svmprot.cgi) to some extent assigns protein functional family irrespective of sequence similarity and has been found to be useful for studying distantly related proteins [Cai, C.Z., Han, L.Y., Ji, Z.L., Chen, X., Chen, Y.Z., 2003. SVM-Prot: web-based support vector machine software for functional classification of a protein from its primary sequence. Nucleic Acids Res. 31 (13). 3692–3697]. Here 25 novel viral proteins are selected to test the capability of SVMProt for functional family assignment of viral proteins whose function cannot be confidently predicted on by sequence similarity methods at present. These proteins are without a sequence homolog in the Swissprot database, with its precise function provided in the literature, and not included in the training sets of SVMProt. The predicted functional classes of 72% of these proteins match the literature-described function, which is compared to the overall accuracy of 87% for SVMProt functional class assignment of 34 582 proteins. This suggests that SVMProt to some extent is capable of functional class assignment irrespective of sequence similarity and it is potentially useful for facilitating functional study of novel viral proteins.

Keywords: Novel viral proteins; Statistical learning; Sequence similarity

Introduction

The complete genomes of 1536 viruses have been sequenced (viral genomes at NCBI http://www.ncbi.nlm.nih.gov/genomes/static/vis.html). Knowledge of these genomes has facilitated mechanistic study of viral infections and provided important clues for searching molecular targets of antiviral therapeutics (Herniou et al., 2003; Marra et al., 2003; Miller et al., 2003). The function of over 15% of the putative protein-coding open reading frames (ORFs) in these viral genomes is unknown (Herniou et al., 2003; Marra et al., 2003; Miller et al., 2003). Determination of the function of these unknown ORFs is important for a more comprehensive understanding of the molecular mechanism of specific virus and for searching novel targets for antiviral drug development.

The sequence of many of these unknown ORFs has no significant similarity to proteins of known functions, and their functions are difficult to probe on the basis of sequence similarity. For instance, 50%, 100%, 20%, and 67% of the unknown ORFs in the recently determined genomes of Fer-de-lance virus (Makeyev and Bamford, 2004), Grapevine fleck virus (Sabanadzovic et al., 2001), Indian citrus ringspot virus (Rustici et al., 2002), and SARS coronavirus (He et al., 2004) are without a homolog in Swissprot database (Boeckmann et al., 2003) based on BLAST search against all Swissprot entries as of September 2004. This

* Corresponding author. Fax: +65 6774 6756.
E-mail address: csccyz@nus.edu.sg (Y.Z. Chen).

0042-6822/$ - see front matter © 2004 Elsevier Inc. All rights reserved.
doi:10.1016/j.virol.2004.10.020

suggests that a significant percentage of new viral proteins are likely to have no known sequence homolog. It is thus desirable to explore alternative methods or combination of methods for providing useful hint about the function of unknown viral ORFs.

Various alternative methods for probing protein function have been developed. These include evolutionary analysis (Benner et al., 2000; Eisen, 1998), hidden Markov models (Fujiwara and Asogawa, 2002), structural consideration (Di Gennaro et al., 2001; Teichmann et al., 2001), protein/gene fusion (Enright et al., 1999; Marcotte et al., 1999), protein–protein interactions (Bock and Gough, 2001), motifs (Hodges and Tsai, 2002), family classification by sequence clustering (Enright et al., 2002), and functional family prediction by statistical learning methods (Cai et al., 2003, 2004; Han et al., 2004; Jensen et al., 2002; Karchin et al., 2002).

In the absence of clear sequence or structural similarities, the criteria for comparison of distantly related proteins become increasingly difficult to formulate (Enright and Ouzounis, 2000). Moreover, not all homologous proteins have analogous functions (Benner et al., 2000). The presence of shared domain within a group of proteins does not necessarily imply that these proteins perform the same function (Henikoff et al., 1997). Therefore, careful evaluation is needed to determine which method or combination of methods is useful for facilitating functional study of novel proteins with no homology to proteins of known function.

The web-based software SVMProt (http://jing.cz3.nus.edu.sg/cgi-bin/svmprot.cgi) to some extent has shown some potential for assigning the functional class of distantly related proteins and homologous proteins of different functions as well as homologous proteins (Cai et al., 2003, 2004). It classifies proteins into functional classes defined from activities or physicochemical properties rather than sequence similarity (Bock and Gough, 2001; Cai et al., 2003, 2004; Han et al., 2004; Karchin et al., 2002). In developing SVMProt, proteins in a training set, represented by their sequence-derived physicochemical properties, are projected onto a hyperspace where proteins in a class are separated from those outside the class by a hyperplane. By projecting a new sequence onto the same hyperspace, SVMProt determines whether the corresponding protein is a member of that class based on its location with respect to the hyperplane. The accuracy of SVMProt depends on the diversity of the protein samples, the quality of the representation of protein properties, and the efficiency of the statistical learning algorithm. To some extent, no sequence similarity is required per se. Thus SVMProt may be potentially explored for facilitating functional assignment of proteins whose function cannot be assigned on the basis of sequence similarity.

This work evaluates the usefulness of SVMProt for predicting the functional class of viral ORFs of unknown function. It is assessed by using novel viral proteins that are without a single homolog in the SwissProt database (Boeckmann et al., 2003), with their precise function described in the literature, and are not included in the training sets of SVMProt. These proteins are collected from an unbiased search of Medline (Wheeler et al., 2003) and SwissProt database (Boeckmann et al., 2003). The SVMProt predicted functional classes of these proteins are compared with the function described in the literature and databases to evaluate to what extent SVMProt are useful for functional class assignment of novel viral proteins. The prediction accuracy for assignment of these novel proteins is compared with the overall accuracy of the SVMProt assignment of a large number of proteins to examine the level of sequence similarity independence of SVMProt classification.

Results and discussion

Table 1 gives SVMProt ascribed functional classes for each of the 25 novel viral proteins together with literature-described function. More than one class may be characterized by SVMProt and the probability of correct prediction for each class is also given in Table 1. There are 18 proteins with the top hit of the SVMProt assigned functional class matching the literature-described function, representing 72% of the novel viral proteins studied in this work. These proteins are MotA protein of bacteriophage T4 (Gerber and Hinton, 1996), outer capsid protein VP4 of bovine rotavirus (serotype 10/strain B223) (Hardy et al., 1992), ADOMetase of bacteriophage T3 (Hughes et al., 1987), R.CviJI of chlorella virus IL3A (Skowron et al., 1995), exonuclease of bacteriophage lambda (Sanger et al., 1982), R.CviAII of paramecium bursaria chlorella virus 1 (Zhang et al., 1992), ORF13 of haemophilus phage HP1 (Esposito et al., 1996), Protein kinase of enterobacteria phage T7 (Dunn and Studier, 1983), DNA-directed RNA polymerase of African swine fever virus (strain BA71V) (Yanez et al., 1995), AGT (Miller et al., 2003), BGT (Miller et al., 2003; Tomaschewski et al., 1985), DNK (Broida and Abelson, 1985), Endonuclease II (Sjoberg et al., 1986), Endonuclease V (Valerie et al., 1984), Gp61.9 (Valerie et al., 1986), IRF protein (Chu et al., 1986), and I-TevII (Tomaschewski and Ruger, 1987) of enterobacteria phage T4.

MotA protein of bacteriophage T4 has been found to be a transcription activator that binds to DNA (Gerber and Hinton, 1996) and the far-C-terminal region of the sigma70 subunit of *Escherichia coli* RNA polymerase (Pande et al., 2002). The top hit of SVMProt predicted functional class for this protein is the DNA-binding, which matches with literature-described functions. Bovine rotavirus is a double-stranded RNA virus that is naked. Thus, the outer capsid protein VP4 of bovine rotavirus (serotype 10/strain B223) is located at the viral surface acting as part of the viral coat (Hardy et al., 1992). This protein is predicted by SVMProt as a coat protein that is consistent with literature-described function. The other 14 proteins are enzymes, and these are all correctly assigned by SVMProt to the respective enzyme EC class.

Because these proteins have no homolog of known function in the SwissProt entries of Swissprot database based on PSI-BLAST search, our study suggests that SVMProt has certain level of capability for providing useful hint about the functional class of novel proteins with no or low homology to known proteins, and this capability is not based on sequence similarity or clustering. The overall accuracy of 72% for the assignment of the novel viral proteins is smaller, but not too far away, than that of 87% for SVMProt functional class assignment of 34 582 proteins. This indicates certain level of the sequence-similarity-independent nature of SVM protein classification.

Several factors may affect the accuracy of SVMProt for functional characterization of novel plant proteins. One is the diversity of protein samples used for training SVMProt. It is likely that not all possible types of proteins, particularly those of distantly related members, are adequately represented in some protein classes. This can be improved along with the availability of more protein data. Not all distantly related proteins of the same function have similar structural and chemical features. There are cases in which different functional groups, unconserved with respect to position in the primary sequence, mediate the same mechanistic role, due to the flexibility at the active site (Todd et al., 2002). This plasticity is unlikely to be sufficiently described by the physicochemical descriptors currently used in SVMProt. Therefore, SVMProt in the present form is not expected to be capable of classification of these types of distantly related enzymes.

Some of the SVMProt functional classes are at the level of families and superfamilies that may include a broad spectrum of proteins. It has been shown that SVM works not as well as HMM for distinguishing proteins in a superfamily, but may be more accurate with subfamily discrimination (Karchin et al., 2002). Thus, the use of some large families and superfamilies as the basis for classification may affect the prediction accuracy of SVMProt to some extent.

SVMProt prediction may be further improved by using protein subfamilies as the basis of classification, more comprehensive set of protein samples, and more refined protein descriptors. SVMProt optimization procedure and feature vector selection algorithm may also be improved by adding additional constraints, and by incorporating independent component analysis and kernel PCA in the preprocessing steps.

Concluding remarks

SVMProt shows certain level of capability for predicting functional class of a number of novel viral proteins. This suggests that SVMProt is potentially useful to a certain extent for providing useful hint about the function of distantly related proteins in viruses as well as in other organisms. Further improvements in protein functional family coverage, sample collections, and SVM algorithm may enable the development of SVMProt into a practical tool for facilitating functional study of unknown ORFs in virus genomes and other genomes.

Methods

Selection of viral proteins

The key words, “novel protein virus” or “novel viral protein”, are used to search the Medline (Wheeler et al., 2003) and the Swissprot database (Boeckmann et al., 2003) for finding viral proteins that are both described as novel and with their precise function provided. As the search of the Medline is confined to the abstracts, those proteins whose function is not explicitly hinted in an abstract are not selected. Thus, the selected proteins likely account for a portion of the known novel viral proteins with available functional information. PSI_BLAST (Altschul et al., 1997) sequence analysis is subsequently conducted on each of these novel viral proteins against all SwissProt entries in the SwissProt protein database (Boeckmann et al., 2003) so that those with at least one sequence homolog of known function (including that of the same protein in different species) are removed. The commonly used criterion for homologs, the similarity score e-value < the inclusion threshold value of 0.005 (Altschul et al., 1997), is used in this work. Finally, those proteins that are in the training sets of SVMProt are removed. A total of 25 novel viral proteins are identified in this process, which together with their protein accession number and literature-described functional indications and related references are given in Table 1.

Computational method

SVMProt is based on a statistical learning method support vector machines (SVM) (Burges, 1998). In addition to the prediction of protein functional class (Cai et al., 2003, 2004; Han et al., 2004; Karchin et al., 2002), SVM has also been used for a variety of protein classification problems including fold recognition (Ding and Dubchak, 2001), analysis of solvent accessibility (Yuan et al., 2002), prediction of secondary structures (Hua and Sun, 2001), and protein–protein interactions (Bock and Gough, 2001). As a method that uses sequence-derived physicochemical properties of proteins as the basis for classification, SVM may be particularly useful for functional classification of distantly related proteins and homologous proteins of different functions (Cai et al., 2003, 2004).

There are 75 protein functional classes currently covered by SVMProt. These include 46 enzyme families, 13 channel/transporter families, 4 RNA-binding protein families, DNA-binding proteins, G-protein-coupled receptors, nuclear receptors, Tyrosine receptor kinases, cell adhesion proteins, coat proteins, envelope proteins, outer membrane

Table 1
Novel viral proteins, literature-described functional indications as suggested from experiment and/or sequence analysis, and SVMProt predicted functions

Protein (SwiMSProt or NCBI accession number)	Virus	Literature-described function (reference)	Function characterized by SVMProt (probability of correct characterization *P* value)	Predict on status
ADOMetase (P07693)	Bacteriophage T3	Adenosylmethionine hydrolase (EC 3.3.1.2) (Hughes et al., 1987)	EC 3.3: hydrolase of ether bonds (99.0%); EC 2.7: transferase of phosphorus-containing groups (71.3%); DNA-binding proteins (65.4%);	M
AGT (P04519)	Enterobacteria phage T4	DNA alpha-glucosyltransferase (EC 2.4.1.26) (Miller et al., 2003)	EC 2.4: glycosyltransferase (80.4%); EC 2.7: transferase of phosphorus-containing groups (68.5%)	M
BGT (P04547)	Enterobacteria phage T4	DNA beta-glucosyltransferase (EC 2.4.1.27) (Miller et al., 2003; Tomaschewski et al., 1985)	EC 2.4: glycosyltransferases (95.7%); EC 2.5: transferase of alkyl or aryl groups, other than methyl groups (80.4 %)	M
DNA-directed RNA polymerase (P42488)	African swine fever virus (strain BA71V)	DNA-directed RNA polymerase, subunit 10 homolog (EC 2.7.7.6) (Yanez et al., 1995)	EC 2.7: transferase of phosphorus-containing groups (99.0%)	M
DNK (P04531)	Enterobacteria phage T4	dNMPkinase (EC 2.7.4.13) (Broida and Abelson, 1985)	EC 2.7: transferase of phosphorus-containing groups (99.0%); EC 2.4: glycosyltransferase (96.4%); EC 1.1: oxidoreductase of the CH–OH group of donors (71.3%)	M
Endonuclease II (P07059)	Enterobacteria phage T4	Endonuclease II (EC 3.1.21.1) (Sjoberg et al., 1986)	EC 3.1: hydrolase of ester bonds (99.0%)	M
Endonuclease IV (P39250)	Enterobacteria phage T4	Endonuclease IV (EC 3.1.21.-) (Miller et al., 2003)	No function predicted	NM
Endonuclease V (P04418)	Enterobacteria phage T4	Endonuclease V (EC 3.1.25.1) (Valerie et al., 1984)	EC 3.1: hydrolase of ester bonds (99.0%)	M
Exonuclease (P03697)	Bacteriophage lambda	Exonuclease (EC 3.1.11.3) (Sanger et al., 1982)	EC 3.1: hydrolase of ester bonds (99.0%); EC 4.1: carbon–carbon lyases (88.1%); EC 2.7: transferase of phosphorus-containing groups (68.5%); EC 1.1: oxidoreductase of the CH–OH group of donors (58.6%)	M
FALPE (Q65010)	Amsacta moorei Entomopoxvirus	Associated with unique cytoplasmic structures, filament-associated protein (Alaoui-Ismaili and Richardson, 1996)	No function predicted	NM
Gp61.9 (P13312)	Enterobacteria phage T4	Ribonuclease (EC 3.1.-.-) (Valerie et al., 1986)	EC 3.1: hydrolase of ester bonds (99.0%)	M
IRF protein (P13299)	Enterobacteria phage T4	Intron-associated endonuclease 1 (EC 3.1.-.-) (Chu et al., 1986)	EC 3.1: hydrolase of ester bonds (99.0 %); DNA-binding protein (83.9%)	M
I-TevII (P07072)	Enterobacteria phage T4	Intron-associated endonuclease 2 (EC 3.1.-.-) (Tomaschewski and Ruger, 1987)	EC 3.1: hydrolase of ester bonds (99.0%)	M
MotA protein (P22915)	bacteriophage T4	DNA-binding, transcription regulation (Gerber and Hinton, 1996)	DNA-binding proteins (99.0 %); EC 3.1: hydrolase acting on ester bonds (68.5%)	M
ORF13 (P51715)	Haemophilus phage HP1	Putative adenine-specific methylase (EC 2.1.1.72) (Esposito et al., 1996)	EC 2.1: transferase of one-carbon groups (99.0%); outer membrane (58.6%); mRNA-binding protein (58.6%)	M
Outer capsid protein VP4 (P35746)	Bovine rotavirus (serotype 10/strain B223)	surface outer capsid protein (Hardy et al., 1992)	Coat protein (99.0%)	M
Possible CC chemokine (NP_042976)	Human herpesvirus 6	chemokine like (Luttichau et al., 2003)	No function predicted	NM

(continued on next page)

Table 1 (*continued*)

Protein (SwiMSProt or NCBI accession number)	Virus	Literature-described function (reference)	Function characterized by SVMProt (probability of correct characterization *P* value)	Predict on status
Protein kinase (P00513)	Enterobacteria phage T7	Protein kinase (EC 2.7.1.37) (Dunn and Studier, 1983)	EC 2.7: transferase of phosphorus-containing groups (99.0 %)	M
Putative BARF0 protein (Q8AZJ4)	Epstein–Barr virus	Membrane associated and encodes three arginine-rich motifs of RNA-binding properties (Fries et al., 1997)	EC 4.1.-.-: carbon–carbon lyase (58.6%)	NM
R.CviAII (P31117)	Paramecium bursaria Chlorella virus 1	Endonuclease CviAII (EC 3.1.21.4) (Zhang et al., 1992)	EC 3.1: hydrolase of ester bonds (99.0%)	M
R.CviJI (P52283)	Chlorella virus IL3A	Type II restriction enzyme CviJI (EC 3.1.21.4) (Skowron et al., 1995)	EC 3.1: hydrolase of ester bonds (99.0%); rRNA-binding proteins (98.8%); EC 3.4: peptidase (68.5%)	M
SeMNPV ORF18 (AAF33548)	Spodoptera exigua nucleopolyhedrovirus	Transferase (Wilfred et al., 2002)	No function predicted	NM
SPLT13 (NP_258405)	SpLtMNPV virus	A noval envelope protein (Yin et al., 2003)	No function predicted	NM
TRL10 (AAL27474)	Human cytomegalovirus (HCMV)	Structural envelop glycoprotein (Spaderna et al., 2002)	Transmembrane (98.2%)	NM

The SVMProt predicted functions are categorized in one of the four classes: The first class is M (matched), in which all of the literature-described functional indications are predicted. The second is PM (partially matched), in which some of the literature-described functional indications are predicted. The third is WC (weakly consistent), in which some of the predicted functions can be considered to be consistent with literature-described functional indications on an inconclusive basis. The fourth is NM (not matched), in which No function predicted of the literature-described functions matched or consistent with a predicted function.

proteins, structural proteins, and growth factors. Two broadly defined families of antigens and transmembrane proteins are also included. The majority of known types of viral proteins are included in these classes.

Representative proteins of a particular functional class (positive samples) and those do not belong to this class (negative samples) are needed to train a SVMProt classifier for this class. The positive samples of a class are constructed by using all of the known distinct protein members in that class. Because of the enormous number of proteins, the size of negative samples needs to be restricted to a manageable level by using a minimum set of representative proteins. One way for choosing representative proteins is to select one or a few proteins from each protein domain family. The negative samples of a class are selected from seed proteins of the 7316 curated protein families (domain-based) in the Pfam database excluding those families that have at least one member belong to the functional class. Pfam families are constructed on the basis of sequence similarity. The purpose of using Pfam proteins is to ensure that the negative samples are evenly distributed in the protein space. Sequence similarity is not required for selecting positive samples. In this sense, SVMProt is to some extent independent of sequence similarity.

The SVMProt training system for each family is optimized and tested by using separate testing sets of both positive and negative samples. While possible, all the remaining distinct proteins in each functional family (not in the training set of that family) are used as positive samples and all the remaining representative seed proteins in Pfam curated families are used to construct negative samples in a testing set. The performance of SVMProt classification is further evaluated by using independent sets of both positive and negative samples. There is no duplicate protein in each training, testing, or independent evaluation set.

Data set construction can be demonstrated by an illustrative example of viral coat proteins. The key word "virus coat protein" is used to search the Swissprot, which finds 3012 entries. These entries are checked to remove non-coat proteins, redundant entries, and putative proteins, which gives 848 positive samples. These positive samples cover 140 Pfam families; thus, 14 758 seed proteins of the remaining 7176 Pfam families are used as the negative samples. These positive and negative samples are further divided into 346 and 1474 training, 305 and 8370 testing, and 197 and 4914 independent evaluation sets using the procedure described above.

Not all of the SVMProt classes are at the same hierarchical level. These classes are mixtures of subfamilies, families, and superfamilies. Some classes, such as antigen, need to be more clearly defined into specific subclasses. While it is desirable to define all of the classes at the same level, this is not yet possible because of insufficient data for the subhierarchies of some families and superfamilies. Effort is being made to collect sufficient data so that SVMProt classification systems can be constructed on the basis of a more evenly distributed family structures.

Nonetheless, prediction on the basis of the current structures provides useful hint about the function of a protein.

SVMProt is trained for protein classification in the following manner. First, every protein sequence is represented by specific feature vector assembled from encoded representations of tabulated residue properties including amino acid composition, hydrophobicity, normalized Van der Waals volume, polarity, polarizability, charge, surface tension, secondary structure, and solvent accessibility for each residue in the sequence (Cai et al., 2003). The feature vectors of the positive and negative samples are used to train a SVMProt classifier. The trained SVMProt classifier can then be used to classify a protein into either the positive group (protein is predicted to be a member of the class) or the negative group (protein is predicted to not belong to the class).

The theory of SVM has been described in the literature (Burges, 1998). Thus, only a brief description is given here. SVM is based on the structural risk minimization (SRM) principle from statistical learning theory (Burges, 1998). In linearly separable cases, SVM constructs a hyperplane that separates two different groups of feature vectors with a maximum margin. A feature vector is represented by $\mathbf{x}_i$, with physicochemical descriptors of a protein as its components. The hyperplane is constructed by finding another vector $\mathbf{w}$ and a parameter b that minimizes $||\mathbf{w}||^2$ and satisfies the following conditions:

$$\mathbf{w}\cdot\mathbf{x}_i + b \geq +1, \text{ for } y_i = +1 \qquad \text{Group1 (positive)} \qquad (1)$$

$$\mathbf{w}\cdot\mathbf{x}_i + b \leq -1, \text{ for } y_i = -1 \qquad \text{Group2 (negative)} \qquad (2)$$

where y_i is the group index, $\mathbf{w}$ is a vector normal to the hyperplane, $|b| / ||\mathbf{w}||$ is the perpendicular distance from the hyperplane to the origin and $||\mathbf{w}||^2$ is the Euclidean norm of $\mathbf{w}$. After the determination of $\mathbf{w}$ and b, a given vector $\mathbf{x}$ can be classified by:

$$\text{sign}[(\mathbf{w}\cdot\mathbf{x}) + b] \qquad (3)$$

In nonlinearly separable cases, SVM maps the input variable into a high dimensional feature space using a kernel function $K(\mathbf{x}_i, \mathbf{x}_j)$. An example of a kernel function is the Gaussian kernel that has been extensively used in different protein classification studies (Bock and Gough, 2001; Burges, 1998; Cai et al., 2002; Ding and Dubchak, 2001; Hua and Sun, 2001; Karchin et al., 2002; Yuan et al., 2002):

$$K(\mathbf{x}_i, \mathbf{x}_j) = e^{-||\mathbf{x}_j - \mathbf{x}_i||^2/2\sigma^2} \qquad (4)$$

Linear support vector machine is applied to this feature space and then the decision function is given by:

$$f(\mathbf{x}) = \text{sign}\left(\sum_{i=1}^{1} \alpha_i^0 y_i K(\mathbf{x}, \mathbf{x}_i) + b\right) \qquad (5)$$

where the coefficients α_i^0 and b are determined by maximizing the following Langrangian expression:

$$\sum_{i=1}^{l} \alpha_i - \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l} \alpha_i\alpha_j y_i y_j K(\mathbf{x}_i, \mathbf{x}_j) \qquad (6)$$

under conditions:

$$a_i \geq 0 \text{ and } \sum_{i-1}^{l} \alpha_i y_i = 0 \qquad (7)$$

A positive or negative value from Eq. (3) or Eq. (5) indicates that the vector $\mathbf{x}$ belongs to the positive or negative group, respectively. To further reduce the complexity of parameter selection, hard margin SVM with threshold instead of soft margin SVM with threshold is used in SVMProt.

Scoring of SVM classification of proteins has been estimated by a reliability index and its usefulness has been demonstrated by statistical analysis (Cai et al., 2003; Hua and Sun, 2001). A slightly modified reliability score, R value, is used in SVMProt:

$$R - \text{value} = \begin{cases} 1 & \text{if} \quad 0 < d < 0.2 \\ d/0.2 + 1 & \text{if} \quad 0.2 \leq d < 1.8 \\ 10 & \text{if} \quad d \geq 1.8 \end{cases} \qquad (8)$$

where d is the distance between the position of the vector of a classified protein and the optimal separating hyperplane in the hyperspace, $d > 0$ indicates the sample belongs to the positive group and $d < 0$ the negative group. There is a statistical correlation between R value and expected classification accuracy (probability of correct classification) (Cai et al., 2003; Hua and Sun, 2001). Thus, another quantity, P value, is introduced to indicate the expected classification accuracy. P value is derived from the statistical relationship between the R value and actual classification accuracy based on the analysis of 9932 positive and 45,999 negative samples of proteins (Cai et al., 2003).

References

Alaoui-Ismaili, M.H., Richardson, C.D., 1996. Identification and characterization of a filament-associated protein encoded by *Amsacta moorei* entomopoxvirus. J. Virol. 70 (5), 2697–2705.

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

Benner, S.A., Chamberlin, S.G., Liberles, D.A., Govindarajan, S., Knecht, L., 2000. Functional inferences from reconstructed evolutionary biology involving rectified databases—an evolutionarily grounded approach to functional genomics. Res. Microbiol. 151 (2), 97–106.

Bock, J.R., Gough, D.A., 2001. Predicting protein–protein interactions from primary structure. Bioinformatics 17 (5), 455–460.

Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M.C., Estreicher, A., Gasteiger, E., Martin, M.J., Michoud, K., O'Donovan, C., Phan, I.,

Pilbout, S., Schneider, M., 2003. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Res. 31 (1), 365–370.

Broida, J., Abelson, J., 1985. Sequence organization and control of transcription in the bacteriophage T4 tRNA region. J. Mol. Biol. 185 (3), 545–563.

Burges, C., 1998. A tutorial on support vector machine for pattern recognition. Data Min. Knowl. Disc. 2, 121–167.

Cai, Y.D., Liu, X.J., Xu, X.B., Chou, K.C., 2002. Support vector machines for predicting HIV protease cleavage sites in protein. J. Comput. Chem. 23 (2), 267–274.

Cai, C.Z., Han, L.Y., Ji, Z.L., Chen, X., Chen, Y.Z., 2003. SVM-Prot: web-based support vector machine software for functional classification of a protein from its primary sequence. Nucleic Acids Res. 31 (13), 3692–3697.

Cai, C.Z., Han, L.Y., Ji, Z.L., Chen, Y.Z., 2004. Enzyme family classification by support vector machines. Proteins 55 (1), 66–76.

Chu, F.K., Maley, G.F., West, D.K., Belfort, M., Maley, F., 1986. Characterization of the intron in the phage T4 thymidylate synthase gene and evidence for its self-excision from the primary transcript. Cell 45 (2), 157–166.

Di Gennaro, J.A., Siew, N., Hoffman, B.T., Zhang, L., Skolnick, J., Neilson, L.I., Fetrow, J.S., 2001. Enhanced functional annotation of protein sequences via the use of structural descriptors. J. Struct. Biol. 134 (2–3), 232–245.

Ding, C.H., Dubchak, I., 2001. Multi-class protein fold recognition using support vector machines and neural networks. Bioinformatics 17 (4), 349–358.

Dunn, J.J., Studier, F.W., 1983. Complete nucleotide sequence of bacteriophage T7 DNA and the locations of T7 genetic elements. J. Mol. Biol. 166 (4), 477–535.

Eisen, J.A., 1998. Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis. Genome Res. 8 (3), 163–167.

Enright, A.J., Ouzounis, C.A., 2000. GeneRAGE: a robust algorithm for sequence clustering and domain detection. Bioinformatics 16 (5), 451–457.

Enright, A.J., Iliopoulos, I., Kyrpides, N.C., Ouzounis, C.A., 1999. Protein interaction maps for complete genomes based on gene fusion events. Nature 402 (6757), 86–90.

Enright, A.J., Van Dongen, S., Ouzounis, C.A., 2002. An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res. 30 (7), 1575–1584.

Esposito, D., Fitzmaurice, W.P., Benjamin, R.C., Goodman, S.D., Waldman, A.S., Scocca, J.J., 1996. The complete nucleotide sequence of bacteriophage HP1 DNA. Nucleic Acids Res. 24 (12), 2360–2368.

Fries, K.L., Sculley, T.B., Webster-Cyriaque, J., Rajadurai, P., Sadler, R.H., Raab-Traub, N., 1997. Identification of a novel protein encoded by the *Bam*HI A region of the Epstein–Barr virus. J. Virol. 71 (4), 2765–2771.

Fujiwara, Y., Asogawa, M., 2002. Protein function prediction using hidden Markov models and neural networks. NEC Res. Dev. 43, 238–241.

Gerber, J.S., Hinton, D.M., 1996. An N-terminal mutation in the bacteriophage T4 motA gene yields a protein that binds DNA but is defective for activation of transcription. J. Bacteriol. 178 (21), 6133–6139.

Han, L.Y., Cai, C.Z., Lo, S.L., Chung, M.C., Chen, Y.Z., 2004. Prediction of RNA-binding proteins from primary sequence by a support vector machine approach. RNA 10 (3), 355–368.

Hardy, M.E., Gorziglia, M., Woode, G.N., 1992. Amino acid sequence analysis of bovine rotavirus B223 reveals a unique outer capsid protein VP4 and confirms a third bovine VP4 type. Virology 191 (1), 291–300.

He, R., Dobie, F., Ballantine, M., Leeson, A., Li, Y., Bastien, N., Cutts, T., Andonov, A., Cao, J., Booth, T.F., Plummer, F.A., Tyler, S., Baker, L., Li, X., 2004. Analysis of multimerization of the SARS coronavirus nucleocapsid protein. Biochem. Biophys. Res. Commun. 316 (2), 476–483.

Henikoff, S., Greene, E.A., Pietrokovski, S., Bork, P., Attwood, T.K., Hood, L., 1997. Gene families: the taxonomy of protein paralogs and chimeras. Science 278 (5338), 609–614.

Herniou, E.A., Olszewski, J.A., Cory, J.S., O'Reilly, D.R., 2003. The genome sequence and evolution of baculoviruses. Annu. Rev. Entomol. 48, 211–234.

Hodges, H.C., Tsai, J.W., 2002. 3D-Motifs: an informatics approach to protein function prediction. FASEB J. 16, A543.

Hua, S., Sun, Z., 2001. A novel method of protein secondary structure prediction with high segment overlap measure: support vector machine approach. J. Mol. Biol. 308 (2), 397–407.

Hughes, J.A., Brown, L.R., Ferro, A.J., 1987. Nucleotide sequence and analysis of the coliphage T3 S-adenosylmethionine hydrolase gene and its surrounding ribonuclease III processing sites. Nucleic Acids Res. 15 (2), 717–729.

Jensen, L.J., Gupta, R., Blom, N., Devos, D., Tamames, J., Kesmir, C., Nielsen, H., Staerfeldt, H.H., Rapacki, K., Workman, C., Andersen, C.A., Knudsen, S., Krogh, A., Valencia, A., Brunak, S., 2002. Prediction of human protein function from post-translational modifications and localization features. J. Mol. Biol. 319 (5), 1257–1265.

Karchin, R., Karplus, K., Haussler, D., 2002. Classifying G-protein coupled receptors with support vector machines. Bioinformatics 18 (1), 147–159.

Luttichau, H.R., Clark-Lewis, I., Jensen, P.O., Moser, C., Gerstoft, J., Schwartz, T.W., 2003. A highly selective CCR2 chemokine agonist encoded by human herpesvirus 6. J. Biol. Chem. 278 (13), 10928–10933.

Makeyev, E.V., Bamford, D.H., 2004. Evolutionary potential of an RNA virus. J. Virol. 78 (4), 2114–2120.

Marcotte, E.M., Pellegrini, M., Ng, H.L., Rice, D.W., Yeates, T.O., Eisenberg, D., 1999. Detecting protein function and protein–protein interactions from genome sequences. Science 285 (5428), 751–753.

Marra, M.A., Jones, S.J., Astell, C.R., Holt, R.A., Brooks-Wilson, A., Butterfield, Y.S., Khattra, J., Asano, J.K., Barber, S.A., Chan, S.Y., Cloutier, A., Coughlin, S.M., Freeman, D., Girn, N., Griffith, O.L., Leach, S.R., Mayo, M., McDonald, H., Montgomery, S.B., Pandoh, P.K., Petrescu, A.S., Robertson, A.G., Schein, J.E., Siddiqui, A., Smailus, D.E., Stott, J.M., Yang, G.S., Plummer, F., Andonov, A., Artsob, H., Bastien, N., Bernard, K., Booth, T.F., Bowness, D., Czub, M., Drebot, M., Fernando, L., Flick, R., Garbutt, M., Gray, M., Grolla, A., Jones, S., Feldmann, H., Meyers, A., Kabani, A., Li, Y., Normand, S., Stroher, U., Tipples, G.A., Tyler, S., Vogrig, R., Ward, D., Watson, B., Brunham, R.C., Krajden, M., Petric, M., Skowronski, D.M., Upton, C., Roper, R.L., 2003. The genome sequence of the SARS-associated coronavirus. Science 300 (5624), 1399–1404.

Miller, E.S., Kutter, E., Mosig, G., Arisaka, F., Kunisawa, T., Ruger, W., 2003. Bacteriophage T4 genome. Microbiol. Mol. Biol. Rev. 67 (1), 86–156. (Table of contents).

Pande, S., Makela, A., Dove, S.L., Nickels, B.E., Hochschild, A., Hinton, D.M., 2002. The bacteriophage T4 transcription activator MotA interacts with the far-C-terminal region of the sigma70 subunit of *Escherichia coli* RNA polymerase. J. Bacteriol. 184 (14), 3957–3964.

Rustici, G., Milne, R.G., Accotto, G.P., 2002. Nucleotide sequence, genome organisation and phylogenetic analysis of Indian citrus ringspot virus. Brief report. Arch. Virol. 147 (11), 2215–2224.

Sabanadzovic, S., Ghanem-Sabanadzovic, N.A., Saldarelli, P., Martelli, G.P., 2001. Complete nucleotide sequence and genome organization of Grapevine fleck virus. J. Gen. Virol. 82 (Pt. 8), 2009–2015.

Sanger, F., Coulson, A.R., Hong, G.F., Hill, D.F., Petersen, G.B., 1982. Nucleotide sequence of bacteriophage lambda DNA. J. Mol. Biol. 162 (4), 729–773.

Sjoberg, B.M., Hahne, S., Mathews, C.Z., Mathews, C.K., Rand, K.N., Gait, M.J., 1986. The bacteriophage T4 gene for the small subunit of ribonucleotide reductase contains an intron. EMBO J. 5 (8), 2031–2036.

Skowron, P.M., Swaminathan, N., McMaster, K., George, D., Van Etten, J.L., Mead, D.A., 1995. Cloning and applications of the two/three-base

restriction endonuclease R.CviJI from IL-3A virus-infected *Chlorella*. Gene 157 (1–2), 37–41.

Spaderna, S., Blessing, H., Bogner, E., Britt, W., Mach, M., 2002. Identification of glycoprotein gpTRL10 as a structural component of human cytomegalovirus. J. Virol. 76 (3), 1450–1460.

Teichmann, S.A., Murzin, A.G., Chothia, C., 2001. Determination of protein function, evolution and interactions by structural genomics. Curr. Opin. Struct. Biol. 11 (3), 354–363.

Todd, A.E., Orengo, C.A., Thornton, J.M., 2002. Plasticity of enzyme active sites. Trends Biochem. Sci. 27 (8), 419–426.

Tomaschewski, J., Ruger, W., 1987. Nucleotide sequence and primary structures of gene products coded for by the T4 genome between map positions 48.266 kb and 39.166 kb. Nucleic Acids Res. 15 (8), 3632–3633.

Tomaschewski, J., Gram, H., Crabb, J.W., Ruger, W., 1985. T4-induced alpha- and beta-glucosyltransferase: cloning of the genes and a comparison of their products based on sequencing data. Nucleic Acids Res. 13 (21), 7551–7568.

Valerie, K., Henderson, E.E., deRiel, J.K., 1984. Identification, physical map location and sequence of the denV gene from bacteriophage T4. Nucleic Acids Res. 12 (21), 8085–8096.

Valerie, K., Stevens, J., Lynch, M., Henderson, E.E., de Riel, J.K., 1986. Nucleotide sequence and analysis of the 58.3 to 65.5-kb early region of bacteriophage T4. Nucleic Acids Res. 14 (21), 8637–8654.

Wilfred, F.I.J., Roode, E.C., Goldbach, R.W., Vlak, J.M., Zuidema, D., 2002. Characterization of *Spodoptera exigua* multicapsid nucleopolyhedrovirus ORF17/18, a homologue of *Xestia* c-nigrum granulovirus ORF129. J. Gen. Virol. 83 (Pt. 11), 2857–2867.

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., Wagner, L., 2003. Database resources of the National Center for Biotechnology. Nucleic Acids Res. 31 (1), 28–33.

Yanez, R.J., Rodriguez, J.M., Nogal, M.L., Yuste, L., Enriquez, C., Rodriguez, J.F., Vinuela, E., 1995. Analysis of the complete nucleotide sequence of African swine fever virus. Virology 208 (1), 249–278.

Yin, C., Yu, J., Wang, L., Li, Z., Zhang, P., Pang, Y., 2003. Identification of a novel protein associated with envelope of occlusion-derived virus in *Spodoptera litura* multicapsid nucleopolyhedrovirus. Virus Genes 26 (1), 5–13.

Yuan, Z., Burrage, K., Mattick, J.S., 2002. Prediction of protein solvent accessibility using support vector machines. Proteins 48 (3), 566–570.

Zhang, Y., Nelson, M., Nietfeldt, J.W., Burbank, D.E., Van Etten, J.L., 1992. Characterization of *Chlorella* virus PBCV-1 CviAII restriction and modification system. Nucleic Acids Res. 20 (20), 5351–53563.