technical notes Journal of proteome research

Prediction of Functional Class of the SARS Coronavirus Proteins by a Statistical Learning Method

C. Z. Cai,[†,‡] L. Y. Han,[†] X. Chen,[§] Z. W. Cao,[||] and Y. Z. Chen*[,†,||]

Bioinformatics and Drug Design Group, Department of Computational Science, National University of Singapore, Blk SOC1, Level 7, 3 Science Drive 2, Singapore 117543, Department of Applied Physics, Chongqing University, Chongqing 400044, Peoples Republic of China, College of Life Science, Zhejiang University, Hangzhou, Zhejiang 310029, Peoples Republic of China, and Shanghai Center for Bioinformatics Technology, 100 Qinzhou Road, Shanghai 200235, Peoples Republic of China

Received April 20, 2005

Abstract: The complete genome of severe acute respiratory syndrome coronavirus (SARS-CoV) reveals the existence of putative proteins unique to SARS-CoV. Identification of their function facilitates a mechanistic understanding of SARS infection and drug development for its treatment. The sequence of the majority of these putative proteins has no significant similarity to those of known proteins, which complicates the task of using sequence analysis tools to probe their function. Support vector machines (SVM), useful for predicting the functional class of distantly related proteins, is employed to ascribe a possible functional class to SARS-CoV proteins. Testing results indicate that SVM is able to predict the functional class of 73% of the known SARS-CoV proteins with available sequences and 67% of 18 other novel viral proteins. A combination of the sequence comparison method BLAST and SVMProt can further improve the prediction accuracy of SMVProt such that the functional class of two additional SARS-CoV proteins is correctly predicted. Our study suggests that the SARS-CoV genome possibly contains a putative voltage-gated ion channel, structural proteins, a carbon−oxygen lyase, oxidoreductases acting on the CH−OH group of donors, and an ATP-binding cassette transporter. A web version of our software, SVMProt, is accessible at http://jing.cz3.nus.edu.sg/cgi-bin/svmprot.cgi.

Keywords: SARS Coronuvirus • distantly related protein • protein functional characterization • support vector machines • SVMProt

Introduction

Following the identification of a novel coronavirus as the cause of severe respiratory syndrome (SARS),[1−3] the complete genome of this virus has been determined[4,5] and its sequence variations in different isolates have been analyzed.[6] The SARS coronavirus (SARS-CoV) genome contains five major open-reading frames (ORFs) which encode the replicase polyprotein, spike glycoprotein (S), small envelop protein (E), membrane glycoprotein (M), and nucleocapsid protein (N) found in other coronaviruses.[4−6] Moreover, nine potential ORFs unique to SARS-CoV have been identified.[5] While it is unclear which of these ORFs are translated in infected cells, the possibility that some of them may serve novel functions[5] raises great interest in probing their function.

The sequence of the majority of these putative proteins has no significant similarity to those of known proteins,[5] which complicates the task of using sequence analysis tools to probe their potential function. A statistical learning method, support vector machines (SVM), has recently been applied to protein functional classification,[7−9] fold recognition,[10] analysis of solvent accessibility,[11] prediction of secondary structures,[12] and protein−protein interactions.[13,14] As a method that uses sequence-derived physicochemical properties of proteins as the basis for classification, SVM has shown some potential for predicting the functional class of distantly related proteins and homologous proteins of different functions.[8,9,15] It may thus be a useful method to complement sequence alignment, clustering, and motif-based methods in functional characterization of novel proteins. A web-based SVM protein functional classification software SVMProt (http://jing.cz3.nus.edu.sg/cgi-bin/svmprot.cgi)[8] is used in this work to ascribe possible functional roles of SARS-CoV putative proteins.

Method

SVMProt[8] is a group of integrated classification systems that use a statistical learning method—support vector machines (SVM)[16,17]—for predicting the functional class of a protein from its primary sequence, irrespective of sequence similarity. It currently covers 97 protein functional classes including 46 enzyme families, 21 channel/transporter families, 5 RNA-binding protein families, DNA-binding proteins, G-protein-coupled receptors, nuclear receptors, tyrosine receptor kinases, cell adhesion proteins, coat proteins, envelope proteins, transmembrane proteins, outer membrane proteins, structural proteins, growth factors, and antigens. Until now, the majority of known types of viral proteins are included in these classes.

Representative proteins of a particular functional class (positive samples) and those which 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

* To whom correspondence should be addressed. Tel: 65-6874-6877. Fax: 65-6774-6756. E-mail: yzchen@cz3.nus.edu.sg.
[†] National University of Singapore.
[‡] Chongqing University.
[§] Zhejiang University.
[||] Shanghai Center for Bioinformatics Technology.

10.1021/pr050110a CCC: $30.25

Published on Web 08/10/2005

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 belonging 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 class is optimized and tested by using separate testing sets of both positive and negative samples. While possible, all of the remaining distinct proteins in each functional family (not in the training set of that family) are used as positive samples and all of 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 keyword "virus coat protein" is used to search the Swiss−Prot database, which finds 3012 entries. These entries are checked to remove noncoat 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 more evenly distributed family structures.

Nonetheless, prediction on the basis of the current structures provides a useful hint about the functional class of a protein. SVMProt is trained for protein classification in the following manner. First, every protein sequence is represented by a 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.[8] 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 (the protein is predicted to be a member of the class) or the negative group (the protein is predicted to not belong to the class).

Support Vector Machine (SVM) is a promising algorithm for binary classification by means of supervised learning, which was originally developed by Vapnik and co-workers. The theory of SVM has been described in the literature.[16] Thus, only a brief description is given here. SVM is based on the structural risk minimization (SRM) principle from statistical learning theory.[16] 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

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

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

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

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

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:[7,10−13,16,18]

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

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

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

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

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

under the following conditions:

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

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.

Scoring of the SVM classification of proteins has been estimated by a reliability index, and its usefulness has been demonstrated by statistical analysis.[8,12] A slightly modified reliability score, $R-$ value, is used in SVMProt

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

where d is the distance between the position of the vector of a classified protein and the optimal separating hyperplane in the

Table 1. Novel Viral Proteins, Their SVMProt Predicted Functional Classes and Functions Suggested from Experiment and/or Sequence Analysis[a]

Protein (SwissProt and NCBI ID)	Virus	Function suggested by experiment and/or sequence-analysis (reference)	Functional classes characterized by SVMProt (probability of correct characterization)	Prediction status
CRV3 Q80MM6 (AY234855)	*Cotesia rubecula* virus	A novel protein with homology to C-type lectin[21]	Lectin (99.0%) EC 3.1.-.-: Hydrolase − Acting on Ester Bonds (73.8%)	+
SPLT137 (NP_258405)	SpLtMNPV virus	A noval envelope protein[38]	TC 3.A.5: Type II (general) secretory pathway family (58.6%)	−
E1A 13S protein (Q8JSK1) (AF492353)	Human adenovirus type 21	Formed transcription complex[30]	DNA-binding Protein (99.2%) Cell adhesion (71.3%) Outer membrane (58.6%)	c
P14 (Q38563) (AAC60530)	Bacteriophage phi-6	A new small low-abundant nonstructural protein, facilitated packaging or host-cell membrane repair[32]	EC3.4.-.-: Peptidase (58.6%)	c
V cath (P25783)	AcMNPV	Cathepsin-like protease (EC3.4.22.50)[22]	EC 3.4.-.-: Hydrolases−Peptidase (99.0%) EC 4.1.-.-: Carbon−Carbon Lyases (68.5%) EC 1.2.-.-: Oxidoreductases−Acting on the aldehyde or oxo group of donors (68.5%) EC 2.1.-.-: Transferase of One-Carbon Groups (58.6%) TC 3.A.5: Type II (general) secretory pathway family (58.6%)	+
MotA protein (P22915)	bacteriophage T4	DNA-binding, transcription regulation[23]	DNA-binding Proteins (99.0%) EC 3.1.-.-: Hydrolase−Acting on Ester Bonds (68.5%) TC 3.A.5: Type II (general) secretory pathway family (58.6%) TC 3.A.1: ATP-binding cassette family (58.6%)	+
M3 protein (O41925)	Murine gamma herpesvirus 68	soluble chemokine receptor[24]	Transmembrane (99.0%) Cell adhesion (82.2%) EC 3.4.-.-: Peptidase (62.2%) TC 3.A.3: P-type ATPase family (58.6%) 7 transmembrane receptor (Secretin family) (58.6%) TC 1.C.: Channels/Pores−Pore-forming toxins (58.6%)	PC
BFRF1 protein (P03185)	Epstein-barr virus (strain B95−8)	localized on the plasma membrane and nuclear compartments of the cells and is a structural component of the viral particle[39]	EC 2.7.-.-: Transferases of Phosphorus-Containing Groups (88.1%) EC 4.1.-.-: Carbon−Carbon Lyase (58.6%)	-
VSVG (Q89570)	Vesicular stomatitis virus	Transmembrane glycoprotein[25]	Transmembrane (99.0%) Aptamer-binding protein (94.7%) EC 3.4.-.-: Peptidase (92.1%) Coat protein (88.1%) EC 2.7.-.-: Transferases of Phosphorus-Containing Groups (83.9%) EC 1.18.-.-: Oxidoreductases−Acting on iron−sulfur proteins as donors (73.8%)	+
VCP (P68639)	Vaccinia virus	A novel complement control protein, binds to C3b and C4b[40]	No function predicted	−
Major structural protein L1 (Q8V1L7) (AF459425)	Human papillomavirus	Self-assembles into viral particles and bind to a cell-surface receptor, coat protein, capsid formation[26]	Coat protein (73.8%)	+

Table 1. (Continued)

Protein (SwissProt and NCBI ID)	Virus	Function suggested by experiment and/or sequence-analysis (reference)	Functional classes characterized by SVMProt (probability of correct characterization)	Prediction status
FALPE (Q65010)	Amsacta moorei Entomopoxvirus	Associated with unique cytoplasmic structures, filament-associated protein[41]	EC 2.7.-.-: Transferases of Phosphorus-Containing Groups (58.6%)	−
EUS2 (ORF69) (P28926)	Equine herpesvirus-1	Serine-threonine kinase (EC 2.7.1.37)[27]	EC 2.7.-.-: Transferase of Phosphorus-Containing Groups (99.0%)	+
Virulence factor ICP34.5 (P36313)	Human herpes simplex virus 1	Complexes with PCNA hich forms part of replication machinery[31]	DNA-binding Proteins (62.2%) Cell adhesion (58.6%)	PC
Putative BARF0 protein (Q8AZJ4)	Epstein−Barr virus	Membrane associated and encodes three arginin-rich motifs of RNA-binding properties[42]	EC 4.1.-.-: Carbon−Carbon Lyase (58.6%) TC 3.A.15: The Outer Membrane Protein Secreting Main Terminal Branch family (58.6%)	−
35k myristylprotein (O93122)	Shope fibroma poxvirus	a soluble secreted form of an acquired cellular receptor for tumor necrosis factor[43]	Transmembrane (62.2%) Outer membrane (58.6%)	−
ICP6 (O39263)	HSV-1	a novel protein kinase enzymatic activity (EC 1.17.4.1)[28]	EC 1.17.-.-: Oxidoreductase− Acting on CH2 groups (99.1%) Outer membrane (58.6%)	+
Protein IRS1 (P09715)	Human cytomegalovirus	Competes for binding to DNA recognition site of another protein[29]	DNA-binding Protein (83.9%) EC 3.1.-.-: Hydrolase− Acting on Ester Bonds (76.2%) Outer membrane (58.6%)	+

[a] The symbols +, C, PC, and − represent the cases in which one of the SVMProt predicted functional class is in agreement, consistent, partially consistent, and not matching the suggested function from experiment and/or sequence analysis.

hyperspace. There is a statistical correlation between the *R*-value and the expected classification accuracy (probability of correct classification).[8,12] Thus, another quantity, the *P*-value, is introduced to indicate the expected classification accuracy. The *P*-value is derived from the statistical relationship between the *R*-value and the actual classification accuracy based on the analysis of 9932 positive and 45 999 negative samples of proteins.[8]

Results and Discussion

Test of the Capability of SVMProt for Predicting Functional Class of Novel Viral Proteins. Eighteen novel viral proteins with available sequence and function information, searched from Medline[19] abstracts published during from 1987 to 2003 are used to test SVMProt. These proteins are described as novel or new in their respective abstracts. BLAST[20] analysis shows that 10 of these are with no significant sequence similarity to known proteins, and three others are with homology to no more than five distinct proteins (sequence similarity score *e*-value < 0.05). The remaining five proteins possess homology to a relatively small number of proteins from primarily a few viruses in several different species. Table 1 gives SVMProt ascribed functional class for each protein together with literature described function. More than one functional class may be characterized by SVMProt and the probability of correct prediction for each functional class can be estimated by using a statistical method,[8] which is also given in Supporting Information Table 1.

There are nine proteins with one of its SVMProt characterized functional classes matching that described in the literature. These are CRV3 of *Cotesia rubecula* virus,[21] V-cath of AcMNPV,[22] MotA protein of bacteriophage T4,[23] M3 protein of Murine gamma herpesvirus 68,[24] VSVG of Vesicular stomatitis virus,[25] major structural protein L1 of Human papillomavirus,[26] EUS2 (ORF69) of Equine herpesvirus-1,[27] ICP6 of HSV-1,[28] and Protein IRS1 of Human cytomegalovirus.[29] Two other proteins, E1A 13S protein of Human adenovirus 21[30] and virulence factor ICP34.5 of Herpes simplex virus,[31] are characterized as DNA-binding, which is consistent with the finding that the first protein is part of a transcription complex, and the second is part of a replication complex that bind to the viral DNA. Moreover, P14 of bacteriophage phi-6 is predicted to be a protein in the EC3.4 family. A likely candidate for this protein is metalloproteinase, having a function consistent with the reported role of P14 in facilitating viral packaging or host-cell membrane repair.[32] Overall, 67% of the novel viral proteins have one of its SVMProt characterized functional classes to be consistent with that described in the literature.

SVMProt Prediction of the Functional Class of SARS-CoV Proteins. The sequence of each individual protein or putative protein contained in the NCBI entry NC_004718 of the complete SARS-CoV genome[5] is used for predicting its functional class. The SVMProt characterized functional classes of SARS-CoV proteins are given in Table 2 together with the estimated probability of correct prediction and the suggested function from experiment or sequence alignment for some of these proteins.[5] There are fifteen proteins with function derived from experiment or sequence analysis.[5] These are 3C-like proteinase, NSP3, NSP4, NSP6, NSP9, NSP10, NSP13, NSP14, NSP15, RNA-dependent RNA polymerase, putative ribose 2′-*O*-methyltrans-

Table 2. SVMProt Predicted Functional Class of SARS-Associated Coronavirus Proteins Together with the Function Suggested by Experiment or Sequence Analysis[a]

Protein	Function suggested by experiment, or sequence analysis from ref 5 and description in NCBI entry NC_004718	Functional classes characterized by SVMProt (probability of correct characterization)	Prediction status
Counterpart of MHV p65 protein	Unknown	EC 2.6.-.-: Transferases of Nitrogenous Groups (86.8%) EC 1.1.-.-: Oxidoreductases−Acting on the CH−OH group of donors (62.2%)	?
3CL-PRO	3C-like proteinase (EC 3.4.22.-)	EC 3.4.-.-: Peptidase (62.2%)	+
nsp3	predicted phosphoesterase (EC3.1.-.-) (similar to the ppr-1′−p processing enzyme) formerly known as 'X-domain', PL-PRO (EC 3.4.22.-) similar to that of MHV PL2-PRO, Y-domain; Transmembrane domain 1; adenosine diphosphate-ribose 1′′-phosphatase (ADPR) (EC 3.6.1.-)	Transmembrane (98.5%) EC 3.6.-.-: Hydrolases (83.9%) Outer membrane (58.6%) TC 3.A.3: P-type ATPase family (58.6%)	PC
nsp4	Contains transmembrane domain 2	G Protein Coupled Receptors (98.9%) Transmembrane (98.8%) 7 transmembrane receptor (rhodopsin family and chemoreceptor) (62.2%) 7 transmembrane receptor (metabotropic glutamate family) (58.6%) 7 tarnsmembrane receptor (Secretin family) (58.6%) TC 2.A.1: Major facilitator family (58.6%) TC 3.A.5: Type II (general) secretory pathway family (58.6%)	+
nsp6	putative transmembrane domain	Transmembrane (99.1%) TC 2.A.1: Major facilitator family (58.6%)	+
nsp7	Unknown	TC 3.A.15: The Outer Membrane Protein Secreting Main Terminal Branch family (58.6%) TC 3.A.1: ATP-binding cassette family (58.6%)	?
nsp8	Unknown	EC 2.3.-.-: Acyltransferases (58.6%) EC 4.2.-.-: Carbon−Oxygen Lyases (58.6%)	?
nsp9	ssRNA-binding protein (experimental)[36,37]	EC 2.4.-.-: Glycosltransferases (76.2%) DNA-binding Proteins (58.6%)	−
nsp10	formerly known as growth-factor-like protein	Outer membrane(58.6%)	−
nsp12 (RNA-dependent RNA polymerase)	RNA-dependent RNA polymerase (EC 2.7.7.48)	Transmembrane (83.9%) EC 4.1.-.-: Carbon−Carbon Lyases (76.2%) EC 2.7.-.-: Transferases of Phosphorus-Containing Groups (62.2%)	+
nsp13	zinc-binding domain (ZD), NTPase/helicase[44] domain (EC3.6.1.-). RNA 5′-triphosphatase (EC 3.6.1.-)	Transmembrane (97.3%) Envelope protein (96.7%) EC 3.6.-.-: Hydrolases − Acting on Acid Anhydrides (62.2%)	+
nsp14	3′-to-5′ exonuclease (EC 3.1.-.-)	EC 3.4.-.-: Peptidases (76.2%)	−
nsp15	uridylate-specific endoribonuclease NendoU (EC3.1.-.-)	EC 1.1.-.-: Oxidoreductases−Acting on the CH−OH group of donors (97.0%) EC 2.7.-.-: Transferases of Phosphorus-Containing Groups (78.4%) EC 4.2.-.-: Carbon−Oxygen Lyases (65.4%)	−

Table 2. (Continued)

Protein	Function suggested by experiment, or sequence analysis from ref 5 and description in NCBI entry NC_004718	Functional classes characterized by SVMProt (probability of correct characterization)	Prediction status
putative ribose 2′-O-methyltransferase	2′-O-ribose methyltransferase (EC 2.1.1.-)	EC 2.6.-.-: Transferases of Nitrogenous Groups (92.1%) EC 3.1.-.-: Hydrolases−Acting on Ester Bonds (73.8%) EC 4.1.-.-: Carbon−Carbon Lyase (62.2%) EC 2.4.-.-: Glycosyltransferases (58.6%) EC 2.1.-.-: Transferase of One-Carbon Groups (58.6%)	+
S protein	spike glycoprotein[45]	Transmembrane (99.0%) Aptamer-binding protein (93.6%) Envelope protein (92.9%)	+
Orf3 (sars3a)	No significant similarity to known proteins, three transmembrane regions, signal peptide, ATP-binding properties	Transmembrane (97.5%) TC 1.A.1: Voltage-gated ion channel family (58.6%)	+
Orf4 (sars3b)	No significant similarity to known proteins, a single transmembrane helix	Transmembrane (58.6%) TC 3.A.1: ATP-binding cassette family (58.6%) TC 1.C.: Channels/Pores−Pore-forming toxins (58.6%)	+
E protein	Small envelope protein	Transmembrane (99.0%) EC 1.9.-.-: Oxidoreductases−Acting on a heme group of donors (62.2%) EC 3.6.-.-: Hydrolases−Acting on Acid Anhydrides (58.6%) Envelope protein (58.6%)	+
M protein	membrane glycoprotein; Matrix Protein	Transmembrane (98.8%) Structural protein (Matrix protein, Core protein,Viral occlusion body,Keratin) (89.3%)	+
Orf7 (sars6)	No significant similarity to known proteins, a likely transmembrane helix between residue 3 and 22.	No function predicted	−
Orf8 (sars7a)	No significant similarity to known proteins, a cleaved signal sequence, a transmembrane helix	Transmembrane (76.2%)	+
Orf9 (sars7b)	weakly similar to sterol-C5 desaturase (EC1.3.-.-), a single strong transmembrane helix	Transmembrane (58.6%)	+
Orf10 (sars8a)	no significant sequence similarity to known proteins, a transmembrane helix with one end within viral particle	Transmembrane (58.6%)	+
Orf11 (sars8b)	Matches to a region of human coronavirus E2 glycoprotein precursor, a soluble protein	Outer membrane (58.6%) RNA-binding Protein (58.6%)	?
N protein (sars9a)	nucleocapsid protein[45]	RNA-binding Protein (58.6%)	C
Orf13 (sars9b)	Unknown, no transmembrane helix	Structural protein (Matrix protein, Core protein,Viral occlusion body, Keratin) (71.3%) Outer membrane (58.6%) EC 4.1.-.-: Carbon−Carbon Lyase (58.6%) EC 2.8.-.-: Transferases of Sulfur-Containing Groups (58.6%) EC 4.2.-.-: Carbon−Oxygen Lyase (58.6%)	?

[a] The symbols +, C, PC, and − are described in Table 1. The symbol "?" indicates that the currently available information is insufficient to determine prediction status.

ferase, and the S, E, M, and N proteins. The sequences of all of the above proteins are given in NCBI entry NC_004718. Eleven out of these fifteen proteins with known sequence information have one of the SVMProt ascribed functions, consistent with that predicted from experiment or sequence analysis.

The replicase polyprotein is known to contain three other potential protein-encoding subunits. One of these is the counterpart of the MHV p65 protein. SVMProt ascribed it as either an EC2.6 transferase of nitrogenous groups (87%) or an EC1.1 oxidoreductase (62%). Viral protein with EC1.1 oxidoreductase has been found,[33] and there is no report about a viral protein that belongs to the EC2.6 family. Therefore, it seems that this protein is an enzyme with EC1.1 oxidoreductase activity. The other two proteins are the nonstructural proteins nsp7 and NSP8. The function of NSP7 and NSP8 has not been determined and no functional motif/domain for each of these proteins has been reported. SVMProt characterizes NSP7 as a member of the transporter TC3.A.15 family or TC3.A.1 family. There is evidence of viral protein in the ATP-binding cassette (TC3.A.1) family,[34] whereas there is no report about a viral outer membrane protein secreting main terminal branch. Hence, NSP7 is likely a transporter of the ATP-binding cassette family. NSP8 is classified as either an EC2.3 acyltransferase (59%) or an EC4.2 carbon−oxygen lyase (59%). Viral protein with EC4.2 lyase activity has been found,[35] and there is no report about that with acyltransferase activity. Thus, it is likely that NSP8 is an enzyme with EC4.2 lyase activity.

The putative protein ORF3 is characterized as a transmembrane protein (97.5%) and a transporter in the TC1.A.1 family (voltage-gated ion channel) by SVMProt. The predicted transmembrane property is consistent with the reported identification of three transmembrane regions within this protein.[5] Since proteins of the TC1.A.1 family have been found in a wide range of viruses as well as bacteria, archaea, and eukaryotes (http://www.tcdb.org), it is possible that ORF3 is also a voltage-gated ion channel. SVMProt ascribed three function classes for ORF4, which are transmembrane (59%), TC3.A.1 ATP-binding cassette (59%) and TC1.C. Channels/Pores-Pore forming toxins (59%). The characterized transmembrane property is consistent with the described identification of a single transmembrane helix about this protein.[5] No viral protein of TC1.C (Channels/Pores-Pore forming toxins) family has been found. Thus, it is possible that ORF4 is a protein of the TC3.A.1 ATP-binding cassette family. ORF4 overlaps with ORF3 and E protein, but no potential TRS sequence can be found at the 5′ end of this protein, which led to the suggestion that it might be expressed from the ORF3 mRNA using an internal ribosomal entry site.[5]

SVMProt fails to ascribe a function for ORF7. This putative protein has no significant sequence similarity with known proteins. An analysis of this putative protein indicated a likely transmembrane helix between residues 3 and 22 with the N terminus located outside the viral particle.[5] ORF8, ORF9, and ORF10 are characterized as transmembrane proteins by SVMProt. The predicted transmembrane property for these putative proteins is consistent with reports from a study of SARS-CoV genome, which shows that each of these putative proteins contains a transmembrane helix.[5] FASTA analysis suggested weak similarities of ORF9 with a sterol-C5 desaturase and a hypothetical *Clostridium perfringens* protein.[5] But SVMProt is unable to provide additional information about the function for this and the other two putative proteins.

ORF11 is predicted as either an outer membrane protein (59%) or an RNA-binding protein (59%). A section of this putative protein is known to match to a region of human coronavirus E2 glycoprotein precursor. But it is unclear which function is more likely for this protein. ORF13 is characterized as either a structural protein (71%), or an EC4.1 carbon−carbon lyase (59%), or an EC2.8 transferase of sulfur-containing groups (59%), or an EC4.2 carbon−oxygen lyase (59%), or an outer membrane protein (59%). There has been no report about the possible function of this putative protein other than the finding that no transmembrane helix is detected within this protein.[5] Structural proteins include matrix proteins, core proteins, and viral occlusion body proteins found in various viruses. No viral protein of the EC4.1 or EC2.8 family has been found. So it is likely that ORF13 is either a structural protein, or an outer membrane protein, or an EC4.2 carbon−oxygen lyase. But it remains to be determined which is a more likely function for this protein.

Can Combination of BLAST and SVMProt Improve the Prediction Accuracy of SVMProt? It is of interest to examine whether SVMProt prediction of the SARS-CoV proteins can be further improved if it is combined with sequence alignment. A BLAST search is conducted to find similarity proteins in other coronaviruses for each of the 4 SARS-CoV proteins whose functional class is incorrectly predicted by SVMProt. SVMProt is then used to predict the functional class of the corresponding similarity proteins of each of these proteins. It is found that the functional class of two of these 6 SARS CoV proteins, NSP9 and NSP14, can be correctly predicted by such an approach. P12 chain in pp1ab of HCoV-229E, which is similar in sequence to NSP9 based on BLAST search result, is predicted to be an RNA-binding protein by SVMProt, which is consistent with experimental suggestion that NSP9 is a single-stranded RNA-binding protein.[36,37] A fragment in Replicase 1b of HCoV, which is similar in sequence to NSP14 based on BLAST search result, is predicted to be an EC 3.1 hydrolase acting on an ester bond, which is consistent with the annotation that NSP14 is a 3′−5′ exonuclease (EC 3.1.−.−). Therefore, our study suggests that combination of BLAST and SVMProt can to some extent improve the prediction accuracy of SVMProt. SVMProt prediction of the functional classes of the encoded proteins in all of the available coronavirus genomes are given in the Supporting Information.

Conclusion

SVMProt shows a certain level of capability for predicting the functional class of a number of novel viral proteins, the majority of which are distantly related proteins. It is also able to predict the functional class of 73% of the SARS-CoV proteins with known function. Our analysis provides the functional classes of some of the putative proteins in SARS-CoV, which is subject to further validation. The functional class of two additional SARS-CoV proteins can be predicted by combining the BLAST sequence comparison method with SVMProt. Our study suggests that combined use of different methods can facilitate the functional study of SARS-CoV proteins and other novel proteins, which assists in the mechanistic study of SARS and other diseases and the development of therapeutics for their treatment.

Acknowledgment. This work was supported in part by grants from Singapore ARF R-151-000-031-112, the Shanghai Commission for Science and Technology (04DZ19850), and the "973" National Key Basic Research Program of China (2004CB715901).

Supporting Information Available: SVMProt predicted functional class of the encoded proteins in five coronavirus genome entries in NCBI (Supporting Table 1). This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Ksiazek, T. G.; Erdman, D.; Goldsmith, C. S.; et al. A novel coronavirus associated with severe acute respiratory syndrome. *N. Engl. J. Med.* **2003**, *348*, 1953−1966.

(2) Drosten, C.; Gunther S.; Preiser, W.; et al. Identification of a novel coronavirus in patients with severe acute respiratory syndrome. *N. Engl. J. Med.* **2003**, *348*, 1967−1976.

(3) Poutanen, S. M.; Low, D. E.; Henry, B.; et al. Identification of severe acute respiratory syndrome in Canada. *N. Engl. J. Med.* **2003**, *348*, 1995−2005.

(4) Rota, P. A.; Oberste, M. S.; Monroe S. S.; et al. Characterization of a novel coronavirus associated with severe acute respiratory syndrome. *Science* **2003**, *300*, 1394−1399.

(5) Marra, M. A.; Jones S. J.; Astell, C. R.; et al. The Genome sequence of the SARS-associated coronavirus. *Science* **2003**, *300*, 1399−1404.

(6) Ruan, Y. J.; Wei, C. L.; Ee, A. L.; et al. Comparative full-length genome sequence analysis of 14 SARS coronavirus isolates and common mutations associated with putative origins of infection. *Lancet* **2003**, *361*, 1779−1785.

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

(8) Cai, C. Z.; Han, L. Y.; Ji, Z. L.; Chen, X.; Chen, Y. Z. 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.

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

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

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

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

(13) Bock, J. R.; Gough, D. A. Predicting protein−protein interactions from primary structure. *Bioinformatics* **2001**, *17*, 455−460.

(14) Lo, S. L.; Cai, C. Z.; Chen, Y. Z.; Chung, M. C. M. Effect of training datasets on support vector machine prediction of protein−protein interactions. *Proteomics* **2005**, *5*, 876−884.

(15) Han, L. Y.; Cai, C. Z.; Ji, Z. L.; Cao, Z. W.; Cui, J.; Chen, Y. Z. Predicting functional family of novel enzymes irrespective of sequence similarity: a statistical learning approach. *Nucleic Acids Res.* **2004**, *32*, 6437−6444.

(16) Burges, C. A tutorial on support vector machine for pattern recognition. *Data Min. Knowl. Discuss.* **1998**, *2*, 121−167.

(17) Cai, C. Z.; Wang, W. L.; Sun, L. Z.; Chen, Y. Z. Protein function classification via support vector machine approach. *Math. Biosci.* **2003**, *185*, 111−122.

(18) Cai, Y. D.; Lin, S. L. Support vector machines for predicting rRNA-, RNA-, and DNA-binding proteins from amino acid sequence. *Biochim. Biophys. Acta* **2003**, *1648*, 127−133.

(19) Wheeler, D. L.; Church, D. M.; Federhen, S.; et al. Database resources of the National Center for Biotechnology. *Nucleic Acids Res.* **2003**, *31*, 28−33.

(20) Sonnhammer, E. L.; von Heijne, G.; Krogh, A. A hidden Markov model for predicting transmembrane helices in protein sequences. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **1998**, *6*, 175−182.

(21) Glatz, R.; Schmidt, O.; Asgari, S. Characterization of a Novel Protein with Homology to C-type Lectins Expressed by the Cotesia rubecula Bracovirus in Larvae of the Lepidopteran Host, *Pieris rapae*. *J. Biol. Chem.* **2003**, *278*, 19743−19750.

(22) Lanier, L. M.; Slack, J. M.; Volkman, L. E. Actin binding and proteolysis by the baculovirus AcMNPV: the role of virion-associated V-CATH. *Virology* **1996**, *216*, 380−388.

(23) Gerber, J. S.; Hinton, D. M. 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.* **1996**, *178*, 6133−6139.

(24) Parry, C. M.; Simas, J. P.; Smith, V. P.; et al. A broad spectrum secreted chemokine binding protein encoded by a herpesvirus. *J. Exp. Med.* **2000**, *191*, 573−578.

(25) Sevier, C. S.; Machamer, C. E. p38: A novel protein that associates with the vesicular stomatitis virus glycoprotein. *Biochem. Biophys. Res. Commun.* **2001**, *287*, 574−582.

(26) Gornemann, J.; Hofmann, T. G.; Will, H.; Muller, M. Interaction of human papillomavirus type 16 L2 with cellular proteins: identification of novel nuclear body-associated proteins. *Virology* **2002**, *303*, 69−78.

(27) Colle, C. F.; O'Callaghan, D. J. Localization of the Us protein kinase of equine herpesvirus type 1 is affected by the cytoplasmic structures formed by the noval IR6 protein. *Virology* **1996**, *220*, 424−435.

(28) Peng, T.; Hunter, J. R.; Nelson, J. W. The novel protein kinase of the RR1 subunit of herpes simplex virus has autophosphorylation and transphosphorylation activity that differs in its ATP requirements for HSV-1 and HSV-2. *Virology* **1996**, *216*, 184−196.

(29) Romanowski, M. J.; Shenk, T. Characterization of the human cytomegalovirus irs1 and trs1 genes: a second immediate-early transcription unit within irs1 whose product antagonizes transcriptional activation. *J. Virol.* **1997**, *71*, 1485−1496.

(30) Richter, J. D.; Hurst, H. C.; Jones, N. C. Adenovirus E1A requires synthesis of a cellular protein to establish a stable transcription complex in injected *Xenopus laevis* oocytes. *Mol. Cell. Biol.* **1987**, *7*, 3049−3056.

(31) Brown, S. M.; MacLean, A. R.; McKie, E. A.; Harland, J. The herpes simplex virus virulence factor ICP34.5 and the cellular protein MyD116 complex with proliferating cell nuclear antigen through the 63-amino acid domain conserved in ICP34.5, MyD116, and GADD34. *J. Virol.* **1997**, *71*, 9442−9449.

(32) Casini, G.; Revel, H. R. A new small low-abundant nonstructural protein encoded by the L segment of the dsRNA bacteriophage phi 6. *Virology* **1994**, *203*, 221−228.

(33) Moore, J. B.; Smith, G. L. Steroid hormone synthesis by a vaccinia enzyme: a new type of virus virulence factor. *EMBO J.* **1992**, *11*, 1973−1980.

(34) Raoult, D.; Audic, S.; Robert, C.; et al. The 1.2-Mb Genome Sequence of Mimivirus. *Science* **2004**, *306*, 1344−1350.

(35) Suda, K.; Tanji, Y.; Hori, K.; Unno, H. Evidence for a novel Chlorella virus-encoded alginate lyase. *FEMS Microbiol. Lett.* **1999**, *180*, 45−53.

(36) Egloff, M. P.; Ferron, F.; Campanacci, V.; et al. The severe acute respiratory syndrome-coronavirus replicative protein nsp9 is a single-stranded RNA-binding subunit unique in the RNA virus world. *PNAS* **2004**, *101*, 3792−3796.

(37) Sutton, G.; Fry, E.; Carter, L.; et al. The nsp9 Replicase Protein of SARS−Coronavirus, Structure and Functional Insights. *Structure* **2004**, *12*, 341−353.

(38) Yin, C.; Yu, J.; Wang, L.; et al. Identification of a novel protein associated with envelope of occlusion-derived virus in Spodoptera litura multicapsid nucleopolyhedrovirus. *Virus Genes* **2003**, *26*, 5−13.

(39) Farina, A.; Santarelli, R.; Gonnella, R.; et al. The BFRF1 gene of Epstein−Barr virus encodes a novel protein. *J. Virol.* **2000**, *74*, 3235−3244.

(40) Al-Mohanna, F.; Parhar, R.; Kotwal, G. J. Vaccinia virus complement control protein is capable of protecting xenoendothelial cells from antibody binding and killing by human complement and cytotoxic cells. *Transplantation* **2001**, *71*, 796−801.

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

(42) Fries, K. L.; Sculley, T. B.; Webster-Cyriaque, J.; et al. Identification of a novel protein encoded by the BamHI A region of the Epstein−Barr virus. *J. Virol.* **1997**, *71*, 2765−2771.

(43) Smith, C. A.; Smith, T. D.; Smolak, P. J.; et al. Poxvirus genomes encode a secreted, soluble protein that preferentially inhibits beta chemokine activity yet lacks sequence homology to known chemokine receptors. *Virology* **1997**, *236*, 316−327.

(44) Thiel, V.; Lvanov, K. A.; Putics, A.; et al. Mechanisms and enzymes involved in SARS coronavirus genome expression. *J. Gen. Virol.* **2003**, *84*, 2305−2315.

(45) Krokhin, O.; Li, Y.; Andonov, A.; et al. Mass Spectrometric Characterization of Proteins from the SARS Virus: A Preliminary Report. *Mol. Cell. Proteomics* **2003**, *2*, 346−356.

PR050110A