

# Application of computational biology in understanding emerging infectious diseases: inferring biological function for S-M complex of SARS-CoV

Xiu-Feng Wan<sup>1</sup>, Demet Ataman<sup>2</sup>, and Dong Xu<sup>1,2</sup>

<sup>1</sup> Digital Biology Laboratory, Computer Science Department, University of Missouri-Columbia, Columbia, MO 65211, USA.

<sup>2</sup> UT-ORNL Graduate School of Genome Science and Technology, Oak Ridge, TN 37831, USA.

## ABSTRACT

Bioinformatics can play a special role in studying emerging infectious diseases, where a fast characterization of the diseases is often urgently needed before they are widespread. In this chapter, we described a computational framework for fast modeling of emerging infectious diseases, including five components: (1) gene survey of the biological sequences related to the emerging diseases; (2) biological problem definition; (3) computational analyses of each gene sequence, which may include molecular characterization, structure prediction, phylogenetic analysis, and regulatory motif prediction; (4) literature review; (5) inference of the biological mechanism and hypothesis for experimental testing. We selected Severe acute respiratory syndrome coronavirus (SARS-CoV) as an example to illustrate the computational framework. SARS-CoV is a most recently emerging coronavirus causing the worldwide SARS pandemic. It is a novel coronavirus, which is similar in genome organization but distantly related to previously characterized coronaviruses in gene sequences. Among the identified open ORFs, replicase ORF1ab, spike [S], envelope [E], membrane [M] and nucleopcapsid [N] are found in other known coronaviruses with a conserved genome organization. In addition to this common genome organization, this novel virus also has a number of nonstructural proteins with unknown functions. We first computationally surveyed all the open reading frames within SARS-CoV, and then predicted the cleavage sites between S1 and S2. According to the literature, M and S proteins play an important role in the processing of assembly, attachment, and fusion for coronavirus-host interaction. However, there has not been a systematically description about the mechanism of the M and S proteins. Although the cellular roles of the S and M proteins are known, it requires structural information to understand the molecular mechanisms of how these two proteins perform their functions. The protein structure of either the M protein or the S protein has not been solved. Questions remains include: (1) How S1 protein interacts with the associated human receptor? (2) How the S protein interacts with M protein? (3) What is the role of the SM complex in virus fusion? (4) What are the other possible roles of the M protein in the virus infection besides its role in cell assembly? In this chapter, we only focused on question 2. We first performed the molecular characterization of S and M proteins. Then we predicted the three-dimensional structures for the S and M proteins and the potential human receptor N-aminopeptidase. We further illustrated how the SM complex assembles based on correlated mutation analysis. In the end, hypotheses about the S-M complex was proposed based on our computational results and previous experimental results.

# INTRODUCTION

## Importance of studying the emerging disease, SARS

Throughout the human history, emerging infectious diseases have been shown to threaten our civilization. Two remote examples are the “Black Death” and “flu” epidemic. In the 14<sup>th</sup> century, the Black Death killed more than one third of the population in Europe. More than 20 million people were killed by the great influenza epidemic of 1918, which was caused by H<sub>1</sub>N<sub>1</sub> influenza virus A. Recently, a growing list of emerging infectious diseases were detected and reported. The diseases include Legionnaires’ disease, *Clostridium difficile* antibiotic-associated colitis, Toxic Shock Syndrome, Hemolytic Uremic Syndrome and Thrombotic Thrombocytopenic Purpura deriving from food borne infection, Human Immunodeficiency Syndrome (AIDS), Bovine Spongiform Encephalopathy, Chronic Wasting Disease of cervids, Hantavirus Pulmonary Syndrome, and West Nile virus encephalitis (Feldmann *et al.* 2002; Maki 2003).

In 2003, a new unusual pneumonia, Severe Acute Respiratory Syndrome (SARS), attacked us as “a new apocalyptic horseman” (Maki 2003), which was caused by a novel coronavirus, SARS-CoV (Ksiazek *et al.* 2003). After the first case of this disease was identified in November 2002 in Guangdong Province, China, this disease spread to more than 32 countries and areas around the world. Until August 7, 2003, 8422 persons worldwide have been infected by SARS, with the vast majority occurred in Mainland China (5327 infected; 349 deaths), Hong Kong (1755; 300), Taiwan (665; 180), Canada (251, 41), Singapore (238; 33), and Vietnam (63; 5). There were about 33 cases but no deaths in the United States (see the “WHO SARS case summary”, [http://www.who.int/csr/sars/country/en/country2003\\_08\\_15.pdf](http://www.who.int/csr/sars/country/en/country2003_08_15.pdf)). This disease resulted in mortality of 11% in general. Mortality in persons older than 60 years was reported to be more than 40% (Donnelly *et al.* 2003). In addition, the economic loss caused by SARS reached billions of dollars. Most recently, the SARS has been reported to occur again in Beijing, P. R. China and one death has been associated with this disease (<http://www.who.int>).

This new disease is unique among the numerous types of community-acquired pneumonia for the following reasons: (1) it is extremely contagious. Among the earlier cases, there are more than 50% infections occurred in the health care workers. (2) The early laboratory diagnosis is difficult. The early virus isolation was hard, RT-PCR is not effective for early detection, and the early seroreversion happened even about 3 weeks after infection (Sampathkumar *et al.* 2003). (3) SARS has a longer incubation period (an average of 6.4 days) than other respiratory viruses (generally 2 to 3 days) (Donnelly *et al.* 2003). (4) SARS-CoV can survive for hours on environmental surfaces, and it can spread fast by droplets spread in respiratory particles with diameters less than 10  $\mu$ m, which can reach more than 2 meter in space (Donnelly *et al.* 2003). (5) Similar to other RNA viruses, SARS-CoV showed a very rapid mutations. Genomic sequencing showed the strains isolated from different infections are significantly different, even when the infections in the same area were examined (Ruan *et al.* 2003).

The SARS has been gradually under control since July 2003 although occasional cases still occurred around the world. However, we still face the potential challenges from this disease that could be extraordinarily severe:

(1) SARS-CoV may be a genetic recombinant virus from human and/or animals (Maki 2003; Guan *et al.* 2003). The genetic evidence and phylogenetic studies showed SARS-CoV is a remote strain from other known coronaviruses (Rota *et al.* 2003; Marra *et al.* 2003; Drosten *et al.* 2003; Ruan *et al.* 2003). The molecular mechanisms of this virus are poorly understood. Even the functions of most individual open reading frames (ORFs) are not known. The routine laboratory work alone will take a long time to accomplish this task. Without understanding of gene functions and molecular mechanisms, the drug and vaccine development will be very difficult.

(2) The evidence showed SARS-associated coronaviruses may spread throughout the population of domestic and market animals (Guan *et al.* 2003). The roof rat was also hypothesized to be a transmission vector of SARS (Ng 2003). This provides a potential channel to create new species of coronaviruses besides the rapid mutants generated within SARS-CoV. The new species may be more harmful and more deadly. Even if they are not harmful to human, they may cause severe diseases among domestic animals, which could be a big loss to farming industry.

(3) SARS-CoV may have established a significant reservoir in the human population and it may come back from time to time as shown in the recent emergence of the SARS case in Beijing, P. R. China (<http://www.who.int/>). This virus reservoir can also serve to provide seeds for the generation of new more severe viruses strains as SARS-CoV evolves.

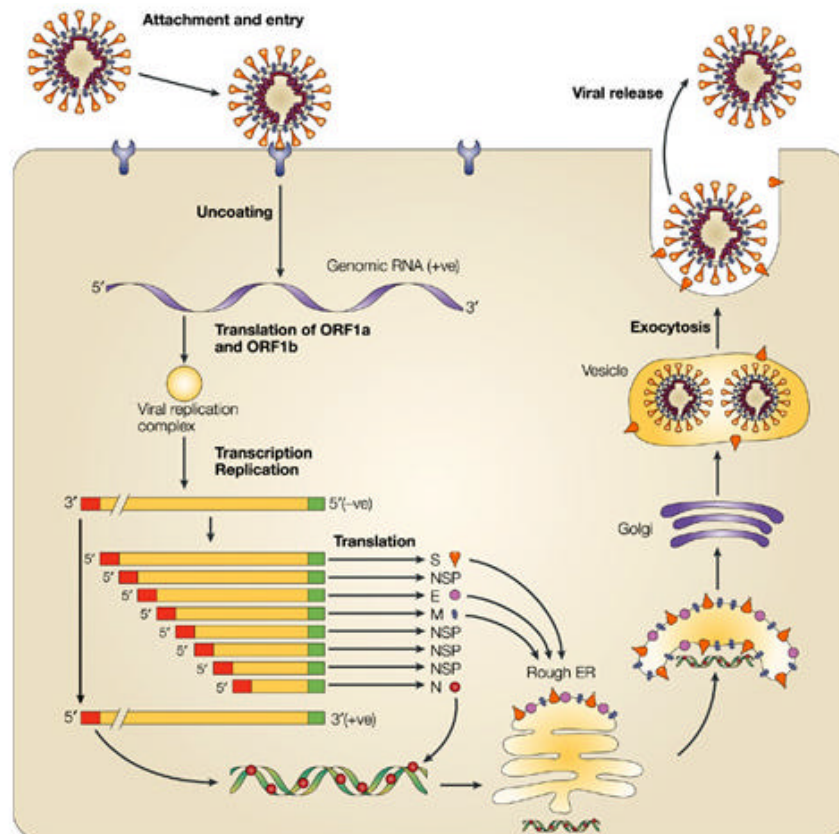
### **Current Understanding of SARS-CoV**

SARS-CoV is a novel coronavirus, which is similar in genome organization but distantly related to previously characterized coronaviruses in gene sequences (Rota *et al.* 2003; Marra *et al.* 2003; Yu *et al.* 2003). Among the identified ORFs, replicase ORF1ab, spike [S], envelope [E], membrane [M] and nucleocapsid [N] are found in other known coronaviruses with a conserved genome organization. In addition to this common genome organization, this novel virus also has a number of nonstructural proteins with unknown functions (Rota *et al.* 2003; Marra *et al.* 2003; Zeng *et al.* 2003).

The current understanding about the life cycle of SARS virus is as follows (Figure 1):

1. The infection starts with the binding of the S protein of the SARS virus to a human cell surface receptor. The SARS virus diffuses into the human cell by receptor-mediated endocytosis or by membrane fusion with the plasma membrane (Blau and Holmes 2001; Bosch *et al.* 2003). Then the viral genome enters the cytoplasm.
2. The positive stranded genomic RNA of SARS virus translates RNA-dependent RNA polymerase. The replication complexes generate new genomic RNA and sub-genomic mRNAs, which encode the new viral proteins, including the three main structural proteins of SARS virus, i.e., S, M and N proteins (An *et al.* 1998;

- Sethna and Brian 1997; Baric and Yount 2000). The N protein is synthesized in the cytoplasm free ribosomes and assembled with the genomic RNA to form RNP (ribonucleoprotein). The M and S proteins are synthesized on the Rough Endoplasmic Reticulum (RER) and form the M-S complex, which then are transported into the Golgi complex.
3. The RNP attaches to the host cell membrane, where the S and M proteins have previously assembled (M-S complex). The RNP buds into the lumen of the vesicle. Finally, the membrane bound RNP and its radiating spikes detach and come to be free in the lumen as an immature virion (Zhang *et al.* 2003). These particles will be transported through Golgi apparatus, where they become mature in a denser and more icosahedral form (Risco *et al.* 1996 & 1998; Salanueva *et al.* 1999).
  4. Finally, the mature virions gather in large vesicles and released onto the cell surface to start the cycle again.



Nature Reviews | Microbiology

Figure 1. Life cycle of coronaviruses (**with permission**, adapted from Stadler *et al.* 2003, Nature Reviews Microbiology 1: 209-218)

Much but incomplete information has been obtained about coronaviruses. Since the first coronavirus was isolated from chickens in 1937, more than 15 species of coronaviruses have been identified in different hosts including human, cattle, pig, rodent,

cat, dog and bird. Some coronaviruses are very important veterinary pathogens. For example, the coronaviruses in chickens causes infectious bronchitis, which has always been a big problem for the poultry industry around the world. Generally the human coronavirus (human-CoV) infection is mild and common. Human-CoV infection occurs often in children. More than 30% of common cold in human is caused by Human-CoV mainly from strain 229E and OC43 (Holmes 2003). However, many human coronaviruses may have not been isolated because of the difficulty of their *in vitro* culture. There have been very rare reports about cross species infection for coronaviruses (Holmes 2003). The lack of understanding of SARS-CoV has shown that our knowledge about the coronavirus family is very limited. Since no good model animal to characterize SARS-CoV has been found, it will be very challenging and time-consuming to characterize this novel emerging coronavirus.

As a new coronavirus family member, SARS-CoVs forms a new (fourth) subgroup of coronaviruses. Until May 14, 2004, more than 150 SARS CoV genomes have been deposited in the NCBI database (<http://www.ncbi.nlm.nih.gov/genomes/SARS/SARS.html>). The genomes of two SARS CoV-like viruses isolated from animals have also been deposited into the database. Although S, E, M, N and 3CL protease of SARS-CoV may have similar structures and functions to the ones in other coronaviruses, the associated amino acid sequence identities are less than 40-50% (Rota *et al.* 2003). The genomic sequence analysis showed SARS-CoV differs substantially from other known coronaviruses, whose sequences are much more similar to each other (Drosten *et al.* 2003; Marra *et al.* 2003; Rota *et al.* 2003). Genetic evidence shows that SARS-CoV may be a recombinant virus from human and animal (Maki 2003; Guan *et al.* 2003). However, the parent sources are still unknown. This suggests that SARS-CoV has gone through a substantial evolution from other known coronaviruses. This novel coronavirus may possess many unique unknown features.

In this chapter, we first give a description about the computational framework for emerging disease and then give an example about how to apply this framework in modeling SARS-CoV, especially inference of the biological roles of the S-M complex.

## **COMPUTATIONAL FRAMEWORK**

The rapid characterization of an emerging disease will provide invaluable information in the prevention and control of the disease. Thus, computational studies may be particularly important for emerging diseases, where understanding of the diseases is often urgently needed before their widespread. Compared to the lab bench work, computational methods can provide a relatively fast and efficient approach to derive theoretical models based on experimental data, to simulate/predict biological processes and to provide working hypotheses for rational designs of new experiments. It usually takes years and months to characterized a single protein experimentally. However, it only takes about hours, minutes, or even seconds to infer the biological functions once the computational model is constructed for the specific requirement. Hence, a good computational framework can play important roles in the research of emerging diseases.

Figure 2 illustrates our computational framework for modeling an emerging disease. (1) We survey genes and get general pictures about an emerging disease. (2) We identify the significant questions to solve by combining current knowledge related to the emerging disease. (2) To target the biological questions, we perform different computational analyses, which may include: A. **Molecular Characterization:** Molecular characterization may include sequence comparison, transmembrane domain prediction, and coiled-coil region prediction. B. **Structure Prediction:** The structure prediction may provide useful information of understanding protein function and designing drugs. If the structure prediction is hard, we can predict the protein contact regions, which can help provide information about protein-protein contact information. C. **Phylogenetic Analysis:** The phylogenetic analysis may provide the evolutionary relationship of query genes. D. **Other Analyses.** The other related analyses can be regulatory motif prediction, codon usage bias analysis, and so on. (4) We need to do the literature review and summarize the computational analysis. (5) In the end, the biological functions will be inferred. At this stage, the hypothesis may be generated for biologists to test using experimental approaches. We can define new biology problems and initiate another similar computational inference procedure iteratively.

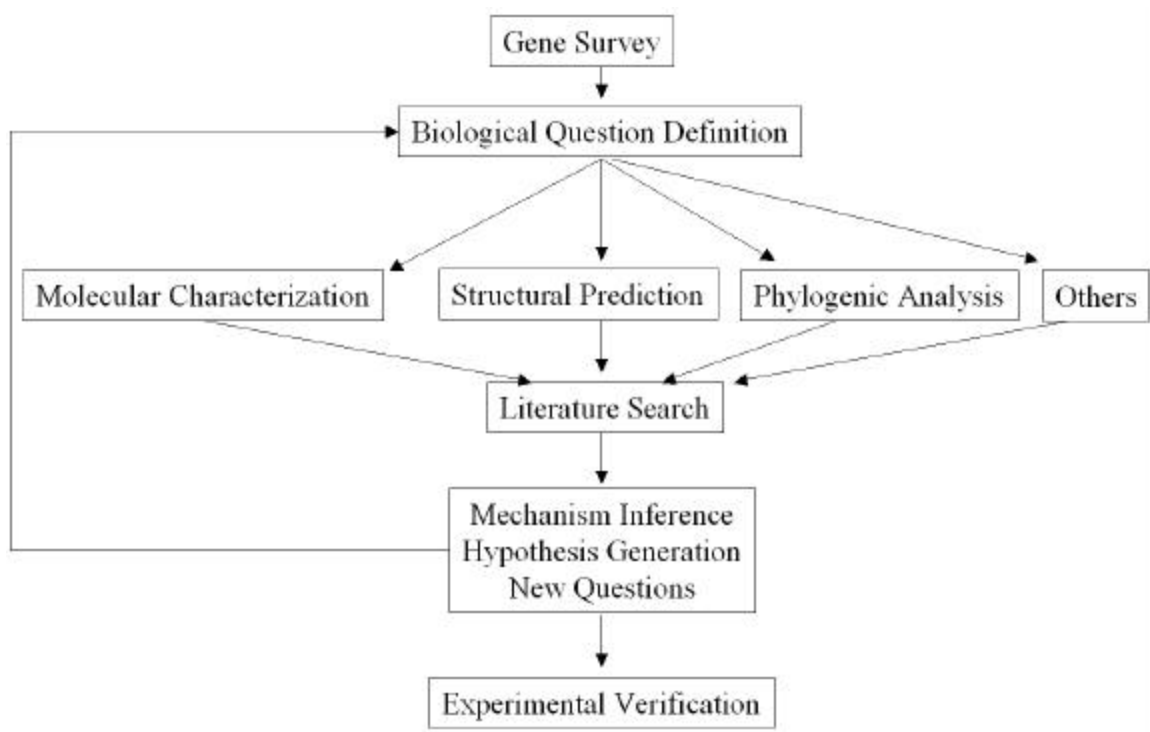


Figure 2. Computational framework for modeling an emerging disease

A central aspect of the computational framework is protein structure prediction. Protein structures provide the basis for functional studies of proteins on the molecular level. Traditionally, protein structures are solved using X-ray crystallography or NMR methods. It could take months or even years to solve one structure, which could cost

hundreds of thousands dollars. An alternative approach for protein structure solution is through computational modeling and prediction. Existing prediction methods fall into two main classes: (a) *ab initio* methods, and (b) comparative modeling methods – predictions based on identified structural relationship with known structures. A key advantage of computational methods is their efficiency – the solution time will be in days, hours or even minutes rather than months or years. Although computational techniques are yet to consistently produce structures of the same quality of experimental methods, the best comparative modeling techniques (including both threading and homology-based modeling) can often generate backbone structures with an accuracy of 4Å RMSD for a large class of proteins (Koehl & Levitt, 1999), as demonstrated in the CASP predictions. With this level of prediction accuracy, we expect that highly useful functional information can be derived. We will be mainly using threading-based methods for protein 3D structure predictions in this proposed study.

Protein threading is a computational technique for protein backbone structure prediction (Bowie *et al.*, 1991, Bryant & Lawrence, 1993). It makes a structural fold prediction from the amino-acid sequence by recognizing a native-like fold of a query protein (if there is any) in a database of experimentally determined structures. The identified folds and predicted structures (even not highly accurate) could provide significant amount of information about the proteins' functions and give useful guidance to experimentalists to conduct further experiments for functional investigation. Recognition of native-like fold is typically achieved by finding an optimal placement (*alignment*) of residues of the query sequence (sequential order maintained; deletions and insertions allowed) onto residue positions of each structural fold in the database, and by identifying those sequence-fold alignments that are statistically significant. For each statistically significant alignment, the placed positions of the query sequence's residues are predicted to be the backbone positions in its structure. Generally speaking, core ( $\alpha$ -helix or  $\beta$ -strand) residues are expected to have more accurate predictions than loop residues, since the former tend to be conserved among structures of the same fold, while the latter may not be (Bryant & Lawrence, 1993). Statistics from the PDB Web site (<http://www.rcsb.org/pdb/holdings.html>) shows that about 90% of the proteins submitted to the PDB database (Westbrook *et al.*, 2000) share similar folds to structures already in PDB. This suggests that many of these protein structures are potentially solvable by computational techniques like protein threading. Though it may be an over-estimate that threading may apply to 90% of proteins because of the biased sampling from the protein space since people tend to solve the "easy" proteins first; and membrane proteins are clearly under-represented in PDB, it is generally believed that 60-70% of new proteins are potentially solvable using threading-based prediction methods (Montelione & Anderson, 1999).

We mainly used our in-house tool PROSPECT (PROtein Structure Prediction and Evaluation Computer Toolkit) (Xu & Xu, 2000) for structure prediction of SARS-CoV proteins. For a query SARS-CoV protein sequence, PROSPECT can effectively recognize native-like folds in the PDB database by finding an optimal sequence-structure alignment, and make a backbone structure prediction based on the recognized native-like folds and the calculated optimal threading alignment. Compared to other threading

programs, PROSPECT has two unique features: (1) it guarantees to find the globally optimal alignment between a query sequence and a template structure for an energy function consisting of the following additive terms: (1a) a mutation energy, (1b) a singleton energy, (1c) a pairwise contact energy, and (1d) a penalty for alignment gaps; and does so efficiently (Xu *et al.*, 1998; Xu & Xu, 2000); and (2) it allows a user to easily incorporate experimental data as constraints in the threading process, and guarantees to find the globally optimal threading under the specified constraints. The advantage of having a rigorous threading algorithm has been demonstrated in our CASP3 (Xu, *et al.*, 1999) and CASP4 (Xu, *et al.*, 2001a) predictions. PROSPECT has been extended to a pipeline (<http://compbio.ornl.gov/proteinpipeline/>) that includes other related computational analysis tools, such as (a) preprocessing of protein sequences, which includes signal peptide prediction, protein type prediction (membrane or soluble), protein domain partition, (b) secondary structure prediction, (c) fold recognition, and (d) atomic structural model generation. (Shah *et al.*, 2003; Guo *et al.*, in press).

During the following section, we are going to model SARS-CoV based on this computational framework: (1) computational survey of all the ORFs in SARS-CoV; (2) literature search to define the problems; (3) computational analysis: (A) molecular characterization of S and M proteins; (B) structural prediction of S, M, and the potential receptor N-aminopeptidase; (C) contact region prediction between S and M proteins using correlated mutation; (4) biological mechanisms inference for the S-M complex by combination of current knowledge about the S-M complex.

## **COMPUTATIONAL MODELING OF SARS**

### **Computational survey of all of ORFs in SARS-CoV**

A step towards characterizing the genes in SARS-CoV is an in-depth computational analysis of protein structure and function. We used the PROSPECT pipeline to survey the 11 Open Reading Frames (ORFs) in SARS-CoV strain Urbani, one of the first SARS-CoV genomes (Rota *et al.* 2003). Table 1 shows part of our survey results, which give the possible signal peptide and trans-membrane regions for each ORF. The different columns in Table 1 show the gene identification/gene name, number of amino acids, the number of homologs in SwissProt, the cleavage site of the predicted signal peptide (two boundary residues and prediction confidence), and predicted trans-membrane segments (primary helix means that the helix is stable in membrane by itself; the secondary helix requires interacting other trans-membrane helix/helices to keep it stable in membrane).

Table 1: A computational survey of all the ORFs in SARS-CoV

ORF	Protein length	Swiss-Prot entry	Signal P (start...end, <i>p</i> )	Trans-membrane
AAP13450 (X5)	84	-	-	Soluble protein
AAP13449 (X4)	122	-	15...16, 0.637	Membrane protein Secondary helix: 1-23; Primary helix: 96-117
AAP13448 (X3)	63	-	39...40, 0.106	Membrane protein Primary helix: 12-34
AAP13447 (X2)	154	-	-	Soluble protein
AAP13446 (X1)	274	1	61...62, 0.435	Membrane protein Primary helix: 40-62, 77-99; Secondary helix: 108-130
AAP13445 (N)	422	18	-	Soluble protein
AAP13444 (M)	221	12	39...40, 1.000	Membrane protein Primary helix: 46-68, 78-100; Secondary helix: 14-36
AAP13443 (E)	76	-	43...44, 0.880	Membrane protein Primary helix: 11-33, 37-59
AAP13441 (S)	1,255	18	13...14, 0.421	S1 domain: Soluble protein S2 domain: Membrane protein Primary helix: 531-553
AAP13440 (nonstructural polyprotein)	2,695	9	-	Soluble protein
AAP13439 (nonstructural polyprotein)	4,382	2	-	16 helices

### Biological Problem Definition: Significance of studying the biological roles of S-M complex

Previous studies demonstrated that the N-terminal subunit (S1) of the S protein forms the surface knob-like structure of the spike, and the C-terminal subunit (S2) forms the stem-like structure beneath the knob (de Groot *et al.* 1987). The cleavage of the S protein depends on individual coronaviruses. The cleavage process is essential for the infectivity and cell fusion in many group-II coronaviruses, whereas it may not be necessary for other coronaviruses. It has been shown that the S protein is involved in various functions. The S1 domain was reported to be involved in mediating the attachment to the host cell receptors (Kubo *et al.* 1994; Suzuki & Taguchi 1996). The S2 domain was found to be important for viral entry into cells and virus fusogenic activity (Taguchi & Shimazaki 2000), although S1 was shown to affect the efficiency of virus entry (Krueger *et al.* 2001). It is very interesting that deletions in the S1 region may change the tissue tropism of the virus while not changing the receptor specificity of the virus. Previous reports show that the S1-receptor binding can induce conformational changes of heptad repeat regions in S2, which is similar to the conformational changes in HIV and influenza viruses (Matsuyama & Taguchi 2002; Eckert & Kim 2001).

As shown by co-immunoprecipitation, co-sedimentation, and immunofluorescence analyses (Opstelten *et al.* 1995a & 1995b), the M protein directs the incorporation of S protein into the particles by forming the M-S complex (Opstelten *et al.* 1995). The S protein is first transported to plasma membrane and then binds the M protein in the Golgi complex. It has been shown that the structural requirement for the M protein in the

process of virus assembly is different from that in the M-S complex. Interruption of the amino-terminal or the carboxy-terminal tail did not interfere with the M-S complex formation. However, deletion of the first two or the last two trans-membrane segments of the triple-spanning protein in the M protein, which is known not to affect the structure of the globular part of the protein, interrupts in the M-S complex formation (de Haan *et al.* 1999).

The M protein is a component of the viral envelope that plays a central role in incorporation of the S protein into the viral envelope. It is involved in the assembly of virus particles by the interactions with other viral proteins such as the E and S proteins in the endoplasmic reticulum Golgi intermediate compartment of the host cell. The M-S complex interacts with RNPs and leads to the formation of virus particles as the RNP buds into the ER.

Although the cellular roles of the S and M proteins are known, it requires structural information to understand the molecular mechanisms of how these two proteins perform their functions. However, the protein structure of either the M protein or the S protein has not been solved. Protein structure predictions for the S and M proteins may provide some clues for the following questions: (1) How S1 interacts with the associated receptor? (2) How the S protein interacts with the M protein? (3) What is the role of the S-M complex in virus fusion? (4) What are the other possible roles of the M protein in the virus infection besides its role in cell assembly?

### Molecular Characterization of S and M proteins

**S protein:** Through the transmembrane prediction of SOSUI (Hirokawa *et al.* 1998), the S protein has only a single primary helix with 23 amino acids (1201-1223, GFIAGLIAIVMVTILLCCMTSC C), which is located in the S2 region. The coiled-coil-like regions in S2 protein predicted by LearnCoil-VMF (Wolf *et al.* 1997) are shown in Figure 3.

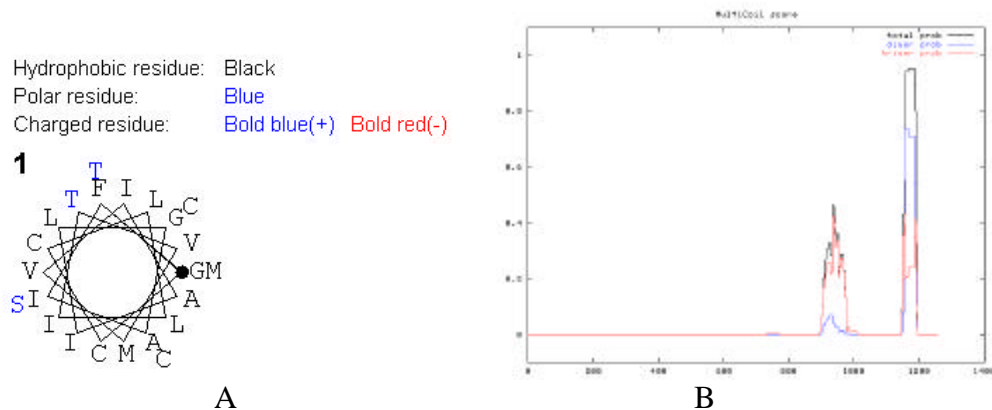


Figure 3. Molecular analysis of the S protein. A. The helical wheel diagram of the predicted transmembrane helix in the S protein. B. Coiled-coil-like region in S protein.

Our computational analysis confirmed the early suggestion (Rota *et al.* 2003) that the S protein of SARS-CoV is not cleaved due to the lack of a cleavage signal. However, structurally the S protein of SARS-CoV probably still forms two loosely connected domains that have similar structures to S1 and S2 in other coronaviruses' S proteins with cleavage sites. To identify the boundary between the two domains, we searched homologs of the S protein of SARS-CoV in Swiss-Prot. If the homolog has a known cleavage site, we align the homolog with the S protein of SARS-CoV. We defined the aligned position of the S protein of SARS-CoV from the cleaved site of the homolog as the predicted cleavage site of the S protein of SARS-CoV. It turns out that the predicted cleavage sites based on different homologs are very similar, i.e., between 670<sup>th</sup> and 671<sup>st</sup> residues. This suggests that the first 670 residues will be in the S1 domain and the residues 671-1255 in the S2 domain.

**M protein:** We predicted the trans-membrane domains using SOSUI (Hirokawa *et al.* 1998). The average hydrophobicity of M protein is 0.41 (Figure 4). Three trans membrane regions with 23 amino acids were predicted as Helix I (14-36, QLLEQWNLVIGFLFLAWIMLLQF), Helix II (46-68, YIIKLVFLWLLWPVTLACFVLAV), and Helix III (78-100, GIAIAMACIVGLMWLSYFVASFR) (Figure 5). The Helix II and III are the primary helices (which mean that each helix can be stable in the membrane by itself) whereas the Helix I is the secondary helix (i.e., the helix is not stable by itself in the membrane and it can only be stable through interacting with other transmembrane helix). We did not find the coiled-coil-like regions in M protein through LearnCoil-VMF (Wolf *et al.* 1997).

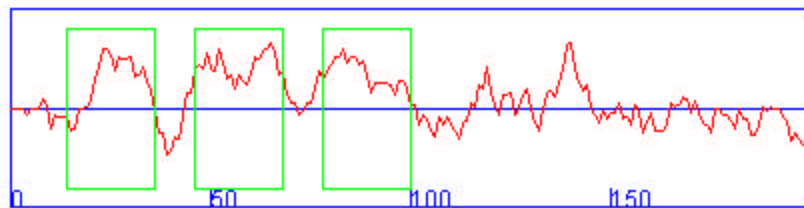


Figure 4. The hydrophobicity plot of the M protein.

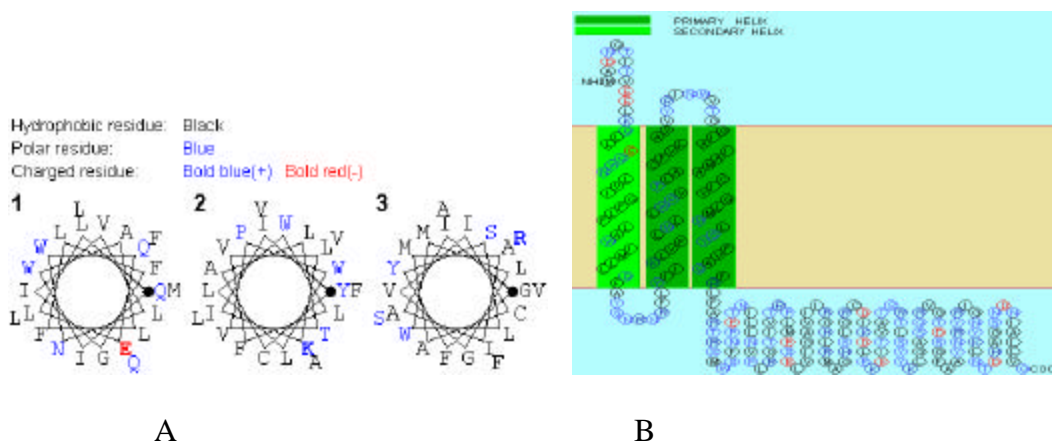


Figure 5. The transmembrane regions of the SARS-CoV M protein. A. Wheel diagram of the M protein; B. Transmembrane regions of the M protein predicted by SOSUI (Hirokawa et al. 1998).

## Structural prediction of S, M proteins and the potential human receptor N-aminopeptidase

### Structural prediction of S and M protein

We submitted the sequences of the M protein as well as the S1 and S2 domains to the PROSPECT pipeline (<http://compbio.ornl.gov/proteinpipeline/>). Table 2 summarizes the structure prediction results. The prediction results can be evaluated through Z-scores. Based on our previous experience, if the Zscore is above 10, the prediction is highly confident, and typically no manual assessment is needed. This is the case for the S2 domain. If the Z-score is less than 6, the prediction confidence level is low and manual analyses are necessary. For the M protein, our manual analyses indicated a good confidence level that the M protein adopts structural templates of 1boy or 1edha, both of which share the same structural fold, i.e., immunoglobulin (Ig)-like beta sandwich. Initial manual analyses did not yield a confident structure prediction for the S1 domain, and further studies are needed.

Interestingly, both the S2 domain and the M protein adopt the fold of Ig-like beta sandwich. The structural similarity suggests that the S2 domain and the M protein may be evolutionarily related through gene fusion and duplication, although their sequences do not have significant similarity anymore after the long period of evolution. Such a phenomenon often occurs among the proteins related to the same biological pathway (Murzin *et al.* 1995). Our results might explain how the M protein interacts with the S2 domain, for virus assembly: since the S2 domain with the fold of Ig-like beta sandwich can interact with the S1 domain, the M protein with the same fold can probably interact with the S1 domain in the mode. This suggests that the S1 domain may act as an on-off switch between the S2 domain and the M protein. Such a mechanism may suggest that the M protein could be also involved in the virus-host cell interaction. After we made the prediction, we found that our suggestion was supported by a recent study in the murine

hepatitis coronavirus study, which showed the glycosylation of the M protein affected the interferogenic capacity of the virus (de Haan *et al.*2004).

Table 2: Structural analyses of the M and S proteins, and the N-aminopeptidase

PDB template (SARS-CoV ORF)	Z score	Class	Fold	Family	Function
1vfaa (S2)	15.45	All beta protein	Ig-like beta sandwich	V set domains (antibody variable domain-like)	It acts as mouse monoclonal antibody.
1boy (M)	<6	All beta protein	Ig-like beta sandwich	Fibronectin type III	It plays a role in initiating the cell- surface assembly and propagation of the coagulation protease cascade.
1edha (M)	<6	All beta protein	Ig-like beta sandwich	Cadherin	Cadherins are cell adhesion proteins interacting with themselves in a homophilic manner in order to connect cells.
1has6 (aminopeptidase)	73.62	Alpha/beta Protein	Zincin-like	Leukotriene A4 hydrolase catalytic domain	It hydrolyzes an epoxide moiety of leukotriene A4 to leukotriene B4. The enzyme also has some peptidase activity.

Table 2 shows the PDB code (and chain name in the fifth letter if available), the Z-score estimated from the PROSPECT pipeline, the structural class of the predicted fold, the category of fold and family, and the function of predicted structural template.

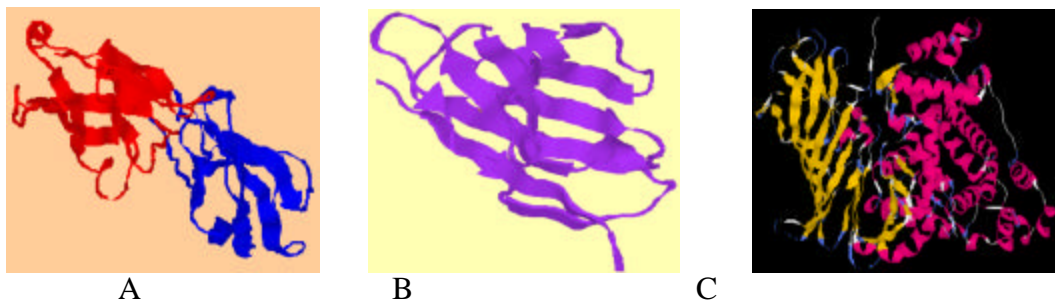


Figure 6: A. Structural template 1vfa for the S2 domain. B. Structural template 1edha for the M protein. C. Structural template 1hs6a for the N-aminopeptidase.

Independent of our study, (Spiga *et al.* 2003) also predicted the structural of Spike protein of SARS-CoV. They predicted the S2 is an Ig-like beta sandwich structure with the template of *C. botulinum* neurotoxin B (pdb ID 1G9D), which is similar to our result. They also match S1 structure of SARS-CoV spike protein into the same template. Figure 7 is the template of *C. botulinum* neurotoxin B.

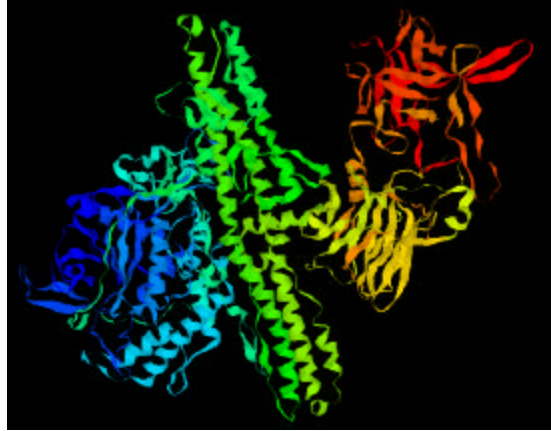


Figure 7. The 3D structure of *C. botulinum* neurotoxin B (pdb ID 1g9d).

### Structural prediction of N-aminopeptidase

N-aminopeptidases in many organisms act as a cell surface receptor for coronaviruses including TGEV, FIPV, FeCV, Human-CoV, and PRCV (Delmas *et al.* 1992; Tresnan & Homles 1998). It has been shown that the N-aminopeptidase interacts with the TGEV S protein in a specific manner (Delmas *et al.* 1992). However, the receptors may be different even for the coronaviruses of the same hosts. Human-CoV -229E uses hAPN as the virus receptor whereas Human-CoV-OC43 uses MHC I as its receptor (Yeager *et al.* 1992; Collins 1993). Bovine coronavirus (BCV) uses 9-O-acetylneuraminic acid as its receptor (Schultze & Herrler 1993), and murine coronavirus use CEACAM as its receptor (Tan *et al.* 2002).

The aminopeptidases are a group of universal peptidase with various functions (Riemann *et al.* 1999; Wan *et al.* 2004a&b). For example, besides functioning in the cell adhesion and amino acid scavenging, this enzyme can serve as the receptor of Human CoV 229E as stated above. Although it is possible that SARS-CoV may utilize other receptors, the N-aminopeptidase or a similar structural fold of N-aminopeptidase might also be the receptor of SARS-CoV (Yu *et al.* 2003). Before predicting the human receptors for SARS-CoV systematically, we first predicted the structure of N-aminopeptidase as a basis for understanding its interaction with the S protein. As shown in Table 2, the structure prediction for N-aminopeptidase has a high Z score, which indicates a good confidence of the structure prediction. The graphic view of the predicted structural fold is shown in Figure 6.

### Contact region prediction between S and M proteins

Prediction of the contact residues between S and M proteins will generate the picture about how the two proteins form the S-M complex. Thus, we can generate a drug analog to block the interaction between S and M proteins thus intercept the life cycle of SARS-CoV. A common method for protein-protein interaction measurement is docking. Docking technique is a physical approach by minimizing the energy between two protein domains. It assumes that the physical principles underlying protein-protein interaction is

the same as protein folding. It characterizes the protein-protein interface by rigorous scanning (Xu et al. 2001, Vakser & Aflalo 1994). However, it required high-quality of 3D structures of the query proteins. This will not be feasible for our problems, since both the structures of S and M protein are not available. Although we were able to model the M and S structures based on threading techniques as shown before, the structural quality for the M and S2 is probably not good enough for docking. An alternative method addressing the contact region correlated mutation approach. Here we will first describe correlated mutation and then illustrate the examples about application of correlated mutation in contact region prediction between S and M proteins.

### Correlation mutation

The rationale for correlated mutation is based on the same physical principle as docking. When one residue mutates in one protein, the contact residue in the other protein often mutates accordingly to adapt the fitness for the interaction between the two residues. Changes of amino acids within a single protein but not the interacting protein may result in the changes of the physical energy thus loosen contacts between two proteins. Especially, when the residue mutates from a negatively charged residue to a positively charged residue or vice versa, it may dramatically change the energetics of the protein-protein interaction. To maintain the functions of protein complex, the interacting proteins may mutate in a correlated manner to minimize physical energy. It is possible that more than one residue may complement the mutations within another protein sequence.

Different research efforts have been put to address the contact problems using correlation mutations. The prediction was based on the information derived from the occurrence of correlated mutations in similar proteins (Gobel *et al.* 1994; Shindyalov *et al.* 1994; Taylor & Hatrick 1994). The other features such as sequence conservation, sequence separation along the chain, alignment stability and residue-specific contact occupancy were incorporated into the contact prediction (Olmea & Valencia 1997). The neural network was employed to execute the prediction (Lund *et al.* 1997; Fariselli & Casadio 1999; Fariselli *et al.* 2001).

The correlated mutation score was calculated as the following (Gobel *et al.* 1994; Pazos *et al.* 1997). Given datasets  $S = \{S_1, S_2, \dots, S_k, \dots, S_N\}$ , and  $S^c = \{S^c_1, S^c_2, \dots, S^c_k, \dots, S^c_N\}$ , the correlated mutation score is:

$$r_{ij} = \frac{1}{N^2} \sum_{kl} \frac{W_{kl} (S_{ikl} - u_i)(S_{jkl} - u_j)}{s_i s_j} \dots\dots\dots(1)$$

where  $N$  is the number of sequences,  $s_i$  and  $s_j$  is the standard deviation of similarities at alignment position  $i$ ,  $W_{kl}$  is weight for sequence  $k$  and  $l$ ,  $S_{ikl}$  is similarity for alignment position  $i$  between sequences  $k$  and  $l$ .  $u_i$  is the average similarity at alignment position  $i$ .

$$W_{kl} = 1 - \frac{I_{kl}}{L} \dots\dots\dots(2)$$

Where  $I_{kl}$  is the sequence identity between sequence  $k$  and  $l$ , and  $L$  is the length of the sequences. We can predict the contact region within a single protein in the same way by deem  $S\mathcal{C}=S$ .

### Contact residue prediction between S protein and M protein

As we mentioned in the section 1, there are four families of coronaviruses. SARS-CoV is a novel member of coronaviruses belonging to family IV. We retrieved the coronaviruses with both S and M proteins in the databases, and table 3 lists the virus strains in our analysis. These viruses included 8 strains in family I, 7 strains in family II, 6 strains in family III, and 28 strains in family IV of SARS-CoV. Both the S protein and the corresponding M protein sequences in the same strain were extracted in the database. The S-M complex in each virus particle was considered as a unit.

The multiple sequence alignments were performed based on ClustalW (Thompson *et al.*, 1994). By writing a C++ programming, we calculate the correlated mutation scores between the S and M sequences based on equation 1 and 2. PAM250 (Jones *et al.* 1992) was used for calculation of the alignment score between the S and M proteins. We used the strain sequence SARS-CoV Frankfurt 1strain (Thiel *et al.* 2003) as the query sequence. The discussion below is based on the position of the S and M proteins in this strain. The positions with gaps were ignored.

We first performed the contact mapping between the positions between S proteins and between M proteins, respectively. Figure 8A is the correlated mutation profiling between the S proteins with a cutoff of 0.55, and Figure 8B is the correlated mutation profiling between M protein with a cutoff of 0.55. The regions of S2 seem to fold by itself and many regions are contact with itself. The residues with correlated score over 0.58 include 270, 272, 369, 525, 527, 531, 579, 700, 715, 716, 723, 726, 734, 749, 807, 808, 825, 845, 859, 885, 917, 929, 944, 947, 965, 980, 990, 994, 1006, 1036, 1049, 1075, 1087, 1097, 1129, 1172, 1175, 1185, 1213, 1217, and 1223. Most of these positions are located within the S2. Instead, in the M proteins (Figure 8B), the scores above its cutoff is 19, 53, 55, 70, 73, 81, 91, 96, 100, 107, 111, 131, 191, 193, 195, 201, 203, 219.

Figure 9 shows the correlated mapping profiling with a cutoff of 0.55 between S and M proteins. From this figure, we can see the contact region is located between positions 40-100 of M protein (major positions 46, 53, 62, 73, 81, 100) and positions 800-1200 of S protein (major positions 16, 822, 869, 1036, 1075, and 1134).

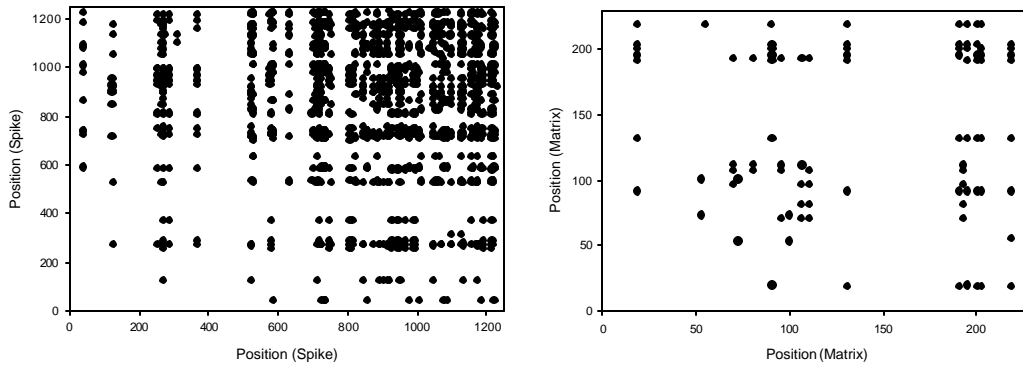


Figure 8. The correlated mutation profiling between the S and M proteins. A. S protein with cutoff 0.55. B. The M protein with cutoff 0.55.

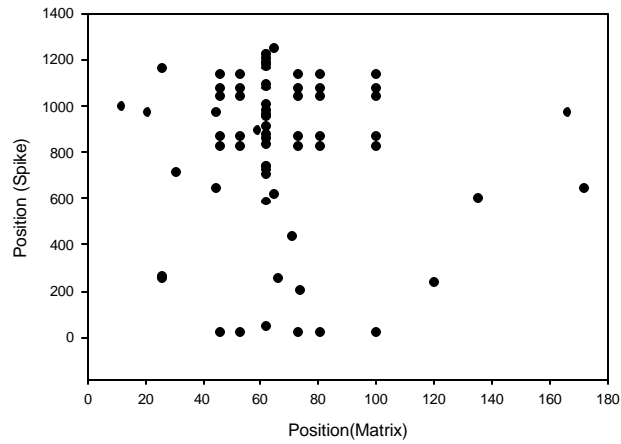


Figure 9. The correlated mutation profiling between S and M proteins with a cutoff of 0.55.

Table 3. Coronavirus sequences used for the contact residue study

Family	Virus Name	S Protein	M protein	Host
1	Feline Infectious Peritonitis Virus	138184	138771	Feline
1	Canine Coronavirus	549301	549376	Canine
1	Porcine epidemic diarrhea virus	19387577	19387580	Porcine
1	Porcine epidemic diarrhea virus	138181	465381	Porcine
1	Porcine epidemic diarrhea virus	138182	138769	Porcine
1	Porcine epidemic diarrhea virus	267339	138767	Porcine
1	Porcine epidemic diarrhea virus	31340517	138768	Porcine
1	Human coronavirus 229E	12175748	12175752	Human
2	Human coronavirus OC43	38018026	38018029	Human
2	Bovine coronavirus	138171	32699786	Bovine
2	Bovine coronavirus	138172	138763	Bovine
2	Equine coronavirus NC99	34398255	34398259	Equine
2	Murine hepatitis virus strain 2	6625763	32699770	Murine
2	Murine hepatitis virus strain ML-11	7739598	7739602	Murine
2	Rat sialodacryoadenitis coronavirus	9719318	9719322	Murine
3	Infectious Bronchitis Virus	138185	138772	Avian
3	Infectious Bronchitis Virus	14149035	14149039	Avian
3	Infectious Bronchitis Virus	14253130	14253134	Avian
3	Infectious Bronchitis Virus	138186	138773	Avian
3	Infectious Bronchitis Virus	138189	138775	Avian
3	Infectious Bronchitis Virus	138190	32699769	Avian
4	SARS-CoV	32187357	32187349	Human
4	SARS-CoV	38231929	38231930	Human
4	SARS-CoV	38231934	38231935	Human
4	SARS-CoV	38231939	38231940	Human
4	SARS-CoV	40457434	40457436	Human
4	SARS-CoV	40457450	40457454	Human
4	SARS-CoV	38505490	38505488	Human
4	SARS-CoV	38505499	38505497	Human
4	SARS-CoV	33285970	33285974	Human
4	SARS-CoV	33285982	33285986	Human
4	SARS-CoV	37576848	37576852	Human
4	SARS-CoV	33188328	33188332	Human
4	SARS-CoV	30795145	30795149	Human
4	SARS-CoV	30698329	30698333	Human
4	SARS-CoV	33411402	33411406	Human
4	SARS-CoV	33411417	33411421	Human
4	SARS-CoV	33411432	33411436	Human
4	SARS-CoV	33411447	33411451	Human
4	SARS-CoV	33411462	33411466	Human
4	SARS-CoV	30027620	30027623	Human
4	SARS-CoV	30275669	30275673	Human
4	SARS-CoV	33114193	33114197	Human
4	SARS-CoV	33114205	33114209	Human
4	SARS-CoV	33114217	33114221	Human
4	SARS-CoV	30023954	30023958	Human
4	SARS-CoV	33578018	33578022	Human
4	SARS-CoV	31581505	31581509	Human
4	SARS-CoV	31416295	31416299	Human

### Inference of Biological Mechanisms of S-M Complex

**S protein:** As mentioned in previous section, the N-terminal subunit (S1) of S protein forms the surface knob-like structure of the spike, and the C-terminal subunit (S2) forms the stem-like structure beneath the knob (de Groot *et al.* 1987). The S1 domain was reported to be involved in mediating the attachment to the host cell receptors (Kubo *et al.* 1994; Suzuki & Taguchi 1996). Recently, (Babcock *et al.* 2004) confirmed that the position 270 to 510 is required for binding human receptor. The S2 domain was found to be important for viral entry into cells and virus fusogenic activity (Taguchi & Shimazaki 2000), although S1 was shown to affect the efficiency of virus entry (Krueger *et al.* 2001). Previous reports show that the S1-receptor binding can induce conformational changes of HRs in S2, which is similar to the conformational changes in HIV and influenza viruses (Matsuyama & Taguchi 2002; Eckert & Kim 2001). The correlated mutation analysis showed the S2 may fold with each other (Figure 8A). The 3D structure modeling and the coiled-coil region prediction support this hypothesis.

**M proteins:** The above analysis showed that M protein has two primary helices and one secondary helix. Based on current understanding, only C tails of M protein is exposed to the interior face of the virion and interacts with the nucleocapsid and form the core (Escors *et al.*, 2001). The M-M protein interaction analysis shows the tails of C may be able to fold by itself (Figure 8B). It shows that M proteins without part of the amphipathic domain or the hydrophilic C-terminal tail were still able to associate with assembly-competent M proteins, resulting in their incorporation into VLPs. However, a mutant M protein in which all three transmembrane domains had been replaced lost this ability. Thus, the interaction between M proteins or M-S complex will have multiple contact sites, particularly at the transmembrane helices (de Haan *et al.*, 2000). This is supported by there is a cluster of positions with high correlated mutation scores around Helix II and Helix III (Figure 8B), which may be involved in the contact between M molecules. It might have some fold at the region close to the Helix III which are not in inner side of the membrane.

**S-M complex:** It was also the N-terminal or the carboxyl-terminal tail of M protein did not interfere with the M-S complex formation. Instead, deletion of the first two or the last two trans-membrane segments of the triple spanning protein in M protein, which is known not to affect the structure of the globular part of the protein, interrupts in the M-S complex formation (de Haan *et al.* 1999). The above prediction shows the Helix II and Helix III of M protein may bind to the areas spanning both of the two predicted coiled-coil regions. Multiple residues are involved in forming the S-M complex. Interesting, these positions are not overlapped M-M contact proteins.

Based on the above analysis, we can pursue a hypothesis: The interaction between multiple residues between region 40-100 of M protein (major positions 46, 53, 62, 73, 81, 100) and region 800-1200 of the S protein (major positions 16, 822, 869, 1036, 1075, and 1134) may form the S-M complex. By using positional mutation strategy, the laboratory experiments may verification this hypothesis. By design specific drug target one or more of these positions, the formation of S-M complex can be prohibited thus kill the SARS-CoV.

New computational questions may also be generated by the computational analysis at this stage. For example, are the contacted residues in the S-M complex still contacted when they form the membrane? If not, how they dissociate during virus assembly?

## SUMMARY

In this chapter, we described a computational framework for modeling emerging disease. This framework can be composed of five components: (1) Gene survey of the sequences related to the emerging diseases; (2) biological problem definition; (3) computational analyses of the query sequences, which may include molecular characterization, structure prediction, phylogenetic analysis, and other analysis such as regulatory motif prediction; (4) literature review and summarize the solution to the problems; (5) inference the biological mechanism and possible generate further question and hypothesis for experimental testing.

We gave an example to apply this computational framework in modeling SARS-CoV. By surveying the ORFs in SARS-CoV, we chose to predict the interaction between S and M complex as our computational problem. Through transmembrane domain prediction, three-dimensional structural prediction, the interaction residues between S and M complex, we were able to define the putative contact regions between S and M proteins. The region 40-100 of M protein (major positions 46, 53, 62, 73, 81, 100) and region 800-1200 of the S protein (major positions 16, 822, 869, 1036, 1075, and 1134) may form the S-M complex. Further biological experiments could design to test these hypotheses. Based on the contact positions, we can design some protein homologs (small protein peptides) to compete with S or M proteins and block the S-M complex formation thus the virus assembly.

## ACKNOWLEDGMENTS

This work was supported by the Laboratory Directed Research and Development Program of Oak Ridge National Laboratory, under Contract DE-AC05-00OR22725, managed by UT-Battelle, LLC. It was also supported by the US Department of Energy's Genomes to Life program (<http://www.doegenomestolife.org>) under project, "Carbon Sequestration in *Synechococcus* Sp.: From Molecular Machines to Hierarchical Modeling" ([www.genomes-to-life.org](http://www.genomes-to-life.org)). We would like to thank Dr. Nagiza Samatova and her group members for helpful discussions.

## REFERENCES

- 1 **An, S., Maeda, A., and Makino, S.** (1998) Coronavirus transcription early in infection. *J. Virol.* **72**:8517-8524.

- 2 **Babcock, G. J., Esshaki, D. J., Thomas, W. D. Jr, and Ambrosino, D. M.,** (2004) Amino acids 270 to 510 of the severe acute respiratory syndrome coronavirus spike protein are required for interaction with receptor. *J Virol.* **9**:4552-60.
- 3 **Baric, R. S., and Yount, B.** (2000) Subgenomic negative-strand RNA function during mouse hepatitis virus infection. *J. Virol.* **74**:4039-4046.
- 4 **Blau, D. M., and Holmes, K. V.** (2001) Human coronavirus HCoV-229E enters susceptible cells via the endocytic pathway. *Adv. Exp. Med. Biol.* **494**:193-198.
- 5 **Bosch, B. J., van der Zee, R., de Haan, C. A., and Rottier, P. J.** (2003) The coronavirus spike protein is a class I virus fusion protein: structural and functional characterization of the fusion core complex. *J. Virol.* **77**:8801-8811.
- 6 **Bowie, J. U., Luthy, R., and Eisenberg, D.** (1991) A method to identify protein sequences that fold into a known three-dimensional structure. *Science* **253**:164-170.
- 7 **Bryant, S. H. and Lawrence, C. E.** (1993) An empirical energy function for threading protein sequence through the folding motif. *Proteins: Struct. Funct. Genet.* **16**:92-112.
- 8 **Collins, A. R.** (1993) HLA class I antigen serves as a receptor for human coronavirus OC43. *Immunol. Invest.* **22**:95-103.
- 9 **de Groot, R. J., Luytjes, W., Horzinek, M. C., van der Zeijst, B. A. M., Spaan, W. J. M., and Lenstra, J. A.** (1987) Evidence for a coiled-coil structure in the spike of coronaviruses. *J. Mol. Biol.* **196**:963-966.
- 10 **de Haan, C. A., Smeets, M., Vernooij, F., Vennema, H., and Rottier, P. J.** (1999) Mapping of the coronavirus membrane protein domains involved in interaction with the spike protein. *J. Virol.* **73**:7441-7452.
- 11 **de Haan, C. A., Vennema, H., and Rottier, P. J.** (2000) Assembly of the coronavirus envelope: homotypic interactions between the M proteins. *J Virol.* **74**:4967-4978.
- 12 **de Haan C. A., Stadler, K., Godeke, G. J., Bosch, B. J., Rottier, P. J.** (2004) Cleavage inhibition of the murine coronavirus spike protein by a furin-like enzyme affects cell-cell but not virus-cell fusion. *J Virol.* **78**:6048-6054.
- 13 **Delmas, B., Gelfi, J., L'Haridon, R., Vogel, L. K., Sjostrom, H., Noren, O., and Laude, H.** (1992) Aminopeptidase N is a major receptor for the entero-pathogenic coronavirus TGEV. *Nature* **357**:417-420.
- 14 **Donnelly, C. A., Ghani, A. C., Leung, G. M., Hedley, A. J., Fraser, C., Riley, S., Abu-Raddad, L. J., Ho, L. M., Thach, T. Q., Chau, P., et al.** (2003) Epidemiological determinants of spread of causal agent of severe acute respiratory syndrome in Hong Kong. *Lancet* **361**:1761-1766.
- 15 **Drosten, C., Preiser, W., Gunther, S., Schmitz, H., and Doerr, H. W.** (2003) Severe acute respiratory syndrome: identification of the etiological agent. *Trends Mol. Med.* **9**:325-327.
- 16 **Eckert, D. M. and Kim, P. S.** (2001) Mechanisms of viral membrane fusion and its inhibition. *Annu Rev Biochem.* **70**:777-810.
- 17 **Escors, D., Ortego, J., and Enjuanes, L.** (2001) The membrane M protein of the transmissible gastroenteritis coronavirus binds to the internal core through the carboxy-terminus. *Adv. Exp. Med. Biol.* **494**:589-593.
- 18 **Fariselli, P. and Casadio, R.** (1999) A neural network based predictor of residue contacts in proteins. *Protein Eng.* **12**:15-21.

- 19 **Fariselli, P., Olmea, O., Valencia, A. and Casadio, R.** (2001) Prediction of contact maps with neural networks and correlated mutations. *Protein Eng.* **14**:835-843.
- 20 **Feldmann, H., Czub, M., Jones, S., Dick, D., Garbutt, M., Grolla, A., and Artsob, H.** (2002) Emerging and re-emerging infectious diseases. *Med. Microbiol. Immunol. (Berl)*. **191**:63-74.
- 21 **Gobel, U., Sander, C., Schneider, R. and Valencia A.** (1994) Correlated mutations and residue contacts in proteins. *Proteins* **18**:309-317.
- 22 **Guan, Y., Zheng, B. J., He, Y. Q., Liu, X. L., Zhuang, Z. X., Cheung, C. L., Luo, S. W., Li, P. H., Zhang, L. J., Guan, Y. J., et al.** (2003) Isolation and Characterization of Viruses Related to the SARS Coronavirus from Animals in Southern China. *Science* **302**:276-278.
- 23 **Guo, J. T., Ellrott, K., Chung, W. J., Xu, D., Passovets, S., and Xu, Y.** PROSPECT-PSPP: An Automatic Computational Pipeline for Protein Structure Prediction. *Nucleic Acid Research*. In press.
- 24 **Hirokawa T, Boon-Chieng S, and Mitaku S.** (1998) SOSUI: classification and secondary structure prediction system for membrane proteins. *Bioinformatics* **14**:378-379.
- 25 **Holmes, K. V.** (2003) SARS-associated coronavirus. *N. Engl. J. Med.* **348**:1948-1951.
- 26 **Montelione, G. T. and Anderson, S.** (1999) Structural genomics: keynote for a human proteome project. *Nature Struct. Biol.* **6**:11-12.
- 27 **Jones, D. T., Taylor, W. R., Thornton, J. M.** (1992) The rapid generation of mutation data matrices from protein sequences. *Comput Appl Biosci.* **8**:275-82.
- 28 **Koehl, P. and Levitt, M.** (1999). A brighter future for protein structure prediction. *Nature Struct. Biol.* **6**:108-111.
- 29 **Krueger, D. K., S. M. Kelly, D. N. Lewicki, R. Ruffolo, and T. M. Gallagher.** (2001) Variations in disparate regions of the murine coronavirus spike protein impact the initiation of membrane fusion. *J. Virol.* **75**:2792-2802
- 30 **Ksiazek, T. G., Erdman, D., Goldsmith, C. S., Zaki, S. R., Peret, T., Emery, S., Tong, S., Urbani, C., Comer, J. A., Lim, W., et al. SARS Working Group.** (2003) A novel coronavirus associated with severe acute respiratory syndrome. *N. Engl. J. Med.* **348**:1953-1966.
- 31 **Kubo, H., Yamada, Y. K. and Taguchi, F.** (1994) Localization of neutralizing epitopes and the receptor-binding site within the amino-terminal 330 amino acids of the murine coronavirus spike protein. *J. Virol.* **68**:5403-5410
- 32 **Lund, O., Frimand, K., Gorodkin, J., Bohr, H., Bohr, J., Hansen, J., and Brunak, S.** (1997) Protein distance constraints predicted by neural networks and probability density functions. *Protein Eng.* **10**:1241-1248.
- 33 **Maki, D. G.** (2003) SARS: 1918 revisited? The urgent need for global collaboration in public health. *Mayo Clin. Proc.* **78**:813-816.
- 34 **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., et al.** (2003) The Genome sequence of the SARS-associated coronavirus. *Science* **300**:1399-1404.
- 35 **Matsuyama, S., and Taguchi, F.** (2002) Receptor-induced conformational changes of murine coronavirus spike protein. *J Virol.* **76**:11819-11826.

- 36 **Murzin, A. G., Brenner, S. E., Hubbard, T., and Chothia, C.** (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **247**:536-540.
- 37 **Ng, S. K.** (2003) Possible role of an animal vector in the SARS outbreak at Amoy Gardens. *Lancet* **362**:570-572.
- 38 **Olmea, O. and Valencia, A.** (1997) Improving contact predictions by the combination of correlated mutations and other sources of sequence information. *Fold Des.* **2**:S25-32.
- 39 **Opstelten, D. J., Raamsman, M. J., Wolfs, K., Horzinek, M. C., and Rottier, P. J.** (1995a) Coexpression and association of the spike protein and the membrane protein of mouse hepatitis virus. *Adv. Exp. Med. Biol.* **380**:291-297.
- 40 **Opstelten, D. J., Raamsman, M. J., Wolfs, K., Horzinek, M.C., and Rottier, P. J.** (1995b) Envelope glycoprotein interactions in coronavirus assembly. *J. Cell Biol.* **131**:339-349.
- 41 **Pazos, F., Olmea, O., and Valencia, A.** (1997) A graphical interface for correlated mutations and other protein structure prediction methods. *Comput Appl Biosci.* **13**:319-321.
- 42 **Riemann, D., Kehlen, A., and Langner, J.** (1999) CD13--not just a marker in leukemia typing. *Immunol. Today.* **20**:83-88.
- 43 **Risco, C., Anton, I. M., Enjuanes, L., and Carrascosa, J. L.** (1996) The Transmissible gastroenteritis coronavirus contains a spherical core shell consisting of M and N proteins. *J. Virol.* **70**:4773-4777.
- 44 **Risco, C., Muntion, M., Enjuanes, L., and Carrascosa, J. L.** (1998) Two types of virus-related particles are found during transmissible gastroenteritis virus morphogenesis. *J. Virol.* **72**:4022-4031.
- 45 **Rota, P. A., Oberste, M. S., Monroe, S. S., Nix, W. A., Campagnoli, R., Icenogle, J. P., Penaranda, S., Bankamp, B., Maher, K., Chen, M. H., et al.** (2003) Characterization of a novel coronavirus associated with severe acute respiratory syndrome. *Science* **300**:1394-1399.
- 46 **Ruan, Y. J., Wei, C. L., Ee, A. L., Vega, V. B., Thoreau, H., Su, S. T., Chia, J. M., Ng, P., Chiu, K. P., Lim, L., et al.** (2003) Comparative full-length genome sequence analysis of 14 SARS coronavirus isolates and common mutations associated with putative origins of infection. *Lancet* **361**:1779-1785.
- 47 **Salanueva, I. J., Carrascosa, J. L., and Risco, C.** (1999) Structural maturation of the transmissible gastroenteritis coronavirus. *J. Virol.* **73**:7952-7964.
- 48 **Sampathkumar, P., Temesgen, Z., Smith, T. F., and Thompson, R. L.** (2003) SARS: epidemiology, clinical presentation, management, and infection control measures. *Mayo Clin. Proc.* **78**:882-890.
- 49 **Schultze, B. and Herrler, G.** (1993) Recognition of N-acetyl-9-O-acetylneuraminic acid by bovine coronavirus and hemagglutinating encephalomyelitis virus. *Adv. Exp. Med. Biol.* **342**:299-304.
- 50 **Sethna, P. B., and Brian, D. A.** (1997) Coronavirus genomic and subgenomic minus-strand RNAs copartition in membrane-protected replication complexes. *J. Virol.* **71**:7744-7749.

- 51 **Shindyalov, I. N., Kolchanov, N. A., and Sander, C.** (1994) Can three-dimensional contacts in protein structures be predicted by analysis of correlated mutations? *Protein Eng.* **7**:349-358.
- 52 **Spiga, O., Bernini, A., Ciutti, A., Chiellini, S., Menciassi, N., Finetti, F., Causarano, V., Anselmi, F., Prischi, F., Niccolai, N.** (2003) Molecular modelling of S1 and S2 subunits of SARS coronavirus spike glycoprotein. *Biochem Biophys Res Commun.* **310**:78-83
- 53 **Shah, M., Passovets, S., Kim, D., Ellrott, K., Wang, L., Vokler, I., LoCascio, P., Xu, D., and Xu, Y.** (2003) A computational pipeline for protein structure prediction and analysis at genome scale. *Bioinformatics*, **19**:1985-1996.
- 54 **Stadler, K., Masignani V., Eickmann M., Becker S., Abrignani S., Klenk H., Rappuoli R.,** (2003) SARS — beginning to understand a new virus, *Nature Reviews Microbiology* **1**:209-218
- 55 **Suzuki, H. and Taguchi, F.** (1996) Analysis of the receptor binding site of murine coronavirus spike glycoprotein. *J. Virol.* **70**:2632-2636
- 56 **Taguchi, F. and Shimazaki, Y. K.** (2000) Functional analysis of an epitope in the S2 subunit of the murine coronavirus spike protein: involvement in fusion activity. *J. Gen. Virol.* **81**:2867-2871.
- 57 **Tan, K., Zelus, B. D., Meijers, R., Liu, J. H., Bergelson, J. M., Duke, N., Zhang, R., Joachimiak, A., Holmes, K. V., and Wang, J. H.** (2002) Crystal structure of murine sCEACAM1a[1,4]: a coronavirus receptor in the CEA family. *EMBO J.* **21**:2076-2086.
- 58 **Taylor, W. R. and Hatrick, K.** (1994) Compensating changes in protein multiple sequence alignments. *Protein Eng.* **7**:341-348.
- 59 **Thiel, V., Ivanov, K. A., Putics, A., Hertzog, T., Schelle, B., Bayer, S., Weissbrich, B., Snijder, E. J., Rabenau, H., Doerr, H. W., Gorbalenya, A. E., and Ziebuhr, J.** (2003) Mechanisms and enzymes involved in SARS coronavirus genome expression. *J. Gen. Virol.* **84**:2305-2315.
- 60 **Thompson, J. D., Higgins, D. G., and Gibson, T. J.** (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22**:4673-4680.
- 61 **Tresnan, D. B. and Holmes, K. V.** (1998) Feline aminopeptidase N is a receptor for all group I coronaviruses. *Adv. Exp. Med. Biol.* **440**:69-75.
- 62 **Vakser, I. A. and Aflalo, C.** (1994) Hydrophobic docking: A proposed enhancement to molecular recognition technique. *Proteins: Struct. Funct. Genet.* **20**:320-329.
- 63 **Wan, X., Branton, S. L., Hughlett, M. B., Hanson, L.A., and G. T. Pharr.** (2004a) Expression and subcellular location of a leucine aminopeptidase of *Mycoplasma gallinarum*. *International Journal of Poultry Science* **3**:70-74.
- 64 **Wan, X., Branton, S. L., Hanson, L. A., and Pharr, G. T.** (2004b) Identification and initial characterization of an aminopeptidase gene in *M. gallinarum*. *Current Microbiology*, **48**:32-38.
- 65 **Westbrook, J., et al.** (2000) The Protein Data Bank: unifying the archive. *Nucleic Acids Res.* **30**:245-248.

- 66 **Wolf, E., Kim, P. S., and Berger, B.** (1997). MultiCoil: A Program for Predicting Two- and Three-Stranded Coiled Coils, *Protein Science* **6**:1179-1189.
- 67 **Xu, D., Baburaj, K., Peterson, C. B., and Xu, Y.** (2001) A Model for the Three-Dimensional Structure of Vitronectin: Predictions for the Multi-Domain Protein from Threading and Docking. *Proteins: Structure, Function, and Genetics*. **44**: 312-320.
- 68 **Xu, D., Kim, D., Dam, P., Shah, M., Uberbacher, E. C., and Xu, Y.** (2003) Characterization of protein structure and function at genome scale using a computational prediction pipeline. In "*Genetic Engineering, Principles and Methods*", edited by Setlow, J. K. 269-293.
- 69 **Xu, Y., and Xu, D.** (2000) Protein threading using PROSPECT: Design and evaluation. *Proteins: Struct Funct Genet* **40**:343-354.
- 70 **Xu, Y., Xu, D., Crawford, O. H., Einstein, J. R., Larimer, F., Uberbacher, E. C., Unseren, M. A., and Zhang, G.** (1999) Protein threading by PROSPECT: A prediction experiment in CASP3. *Protein Eng.* **12**:899-907.
- 71 **Xu, Y., Xu, D., and Uberbacher, E. C.** (1998) An efficient computational method for globally optimal threading. *J. Comp. Biol.* **5**:597-614.
- 72 **Yeager, C. L., Ashmun, R. A., Williams, R. K., Cardellicchio, C. B., Shapiro, L. H., Look, A. T., and Holmes, K. V.** (1992) Human aminopeptidase N is a receptor for human coronavirus 229E. *Nature* **357**:420-422.
- 73 **Yu, X. J., Luo, C., Lin, J. C., Hao, P., He, Y. Y., Guo, Z. M., Qin, L., Su, J., Liu, B. S., Huang, Y., et al.** (2003) Putative hAPN receptor binding sites in SARS-CoV spike protein. *Acta Pharmacol. Sin.* **24**:481-488.
- 74 **Zeng, F. Y., Chan, C. W., Chan, M. N., Chen, J. D., Chow, K. Y., Hon, C. C., Hui, K. H., Li, J., Li, V. Y., et al.** (2003) The complete genome sequence of severe acute respiratory syndrome coronavirus strain HKU-39849 (HK-39). *Exp. Biol. Med. (Maywood)* **228**:866-873.
- 75 **Zhang, Q. F., Cui, J. M., Huang, X. J., Lin, W., Tan, D. Y., Xu, J. W., Yang, Y. F., Zhang, J. Q., Zhang, X., Li, H., et al.** (2003) Morphology and morphogenesis of severe acute respiratory syndrome (SARS)-associated virus. *Acta. Biochimica. Et. Biophysica. Sinica.* **35**:587-591.