

# High-throughput Structural Prediction and Functional Annotation of Proteomic Sequences



Dana Haley-Vicente, Luke Fisher & Lisa Yan  
Accelrys Inc., 10188 Telesis Court, Suite 100, San Diego, CA 92121, USA

**Abstract** Targets in structural genomics initiatives are often selected because they represent a family of proteins for which no 3D structure or fold is known. Such 3D information is crucial for structure-based drug design (SBDD), diagnosis and treatment of disease, and better understanding of basic biology. Once a representative structure for a protein family has been determined, structures of family members with similar in sequences can be inferred by comparative modeling techniques. Here we introduce a high-throughput *in-silico* strategy that uses sequence and fold comparison to obtaining protein structures of similar and distantly related proteins within a family. Using Discovery Studio® (DS) GeneAtlas®, an automated, high-throughput structure prediction and functional annotation pipeline, we are able to rapidly functionally annotate sequences and calculate homology models for thousands of proteins. Experimental templates for the model generation are identified using the combination of sequence comparison using PSI-BLAST and fold recognition using SeqFold. As an example, we will show functional annotation and comparative models of the Severe Acute Respiratory Syndrome (SARS) and West Nile virus (WNV) proteomes. Recent experimental structure data have validated the accuracy of the 3D models created. This data has also allowed us to perform SBDD studies to generate new lead candidates target these deadly viruses.

## Introduction

To accelerate the target discovery and validation step of the drug discovery process, by adding value to proteomic information, DS GeneAtlas, a high throughput, fully automated software environment was used to assign putative function to novel gene targets and to generate annotated homology models. In this study, the SARS and WNV viral proteomes were analyzed

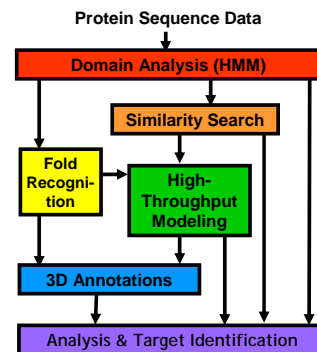
**SARS:** The outbreak of SARS was first identified in Guangdong Province, China in November 2002. This outbreak soon spread to several countries and has had significant health and economic impact. The SARS virus is a RNA virus and is a family member of the Corona viruses. The SARS virus genome is ~30k bases (Tor2 strain) and consists of five major open reading frames that encode ~15 proteins.

**WNV:** The WNV is widely distributed throughout the world, particularly in areas where the climate is warm and the mosquito population is high. The virus was first discovered in 1937 in West Nile district of Uganda in Africa. The virus spreads by the mosquito as the vector and infects human, birds and other mammals as the host. The majority of infected humans have symptoms that range from fever, headache, body aches, skin rash and swollen lymph glands to more severe symptoms such as meningitis or encephalitis. It is estimated that death occurs in 1 out of 1000 infections. The WNV is a single-stranded, RNA flavivirus that encodes ~12 proteins.

Structural information can be used to classify protein function and identify potential binding sites critical for developing new drug targets to combat these deadly viruses. In the absence of atomic resolution structures, 3D homology models are an effective alternative for studying the structure and searching for new lead candidates via SBDD. Here we use DS GeneAtlas to functionally annotate the SARS and WNV proteomes the proteins and determine potential binding sites in preparation for lead discovery.

## Methodology

DS GeneAtlas (figure to the right) is a unique automated protein annotation pipeline for analyzing protein sequences and identifying their biochemical function. DS GeneAtlas goes beyond traditional PSI-BLAST searching and fold recognition to provide complete 3D annotation and protein structure modeling, which can accelerate the target discovery and validation step of drug discovery. The viral protein products, 15 proteins from SARS and 12 proteins from WNV, were processed with DS GeneAtlas and analyzed in DS Modeling 1.1.

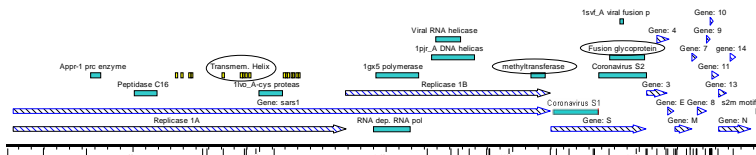


A schematic representation of the DS GeneAtlas high throughput pipeline for functional annotation of protein sequences

## Results and Discussion

### SARS VIRUS

DS Gene was used to map all predicted structural and functional domains (based on 15 open reading frames) to the SARS genome sequence of Tor2 strain (see figure below). The DS GeneAtlas structural and functional annotations are shown as green rectangles and the putative protein transcripts are shown as black and blue arrows. A detailed analysis of each protein functional annotation is available (see SARS references below).

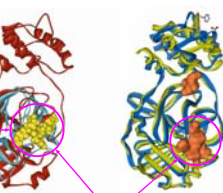


DS GeneAtlas was used to annotate structural and functional domains for three protein sequences, polyprotein1a (pp1a), polyprotein 1b (pp1b), and S protein. For the other nine protein sequences, very little significant homology was found with the known structural templates and the sequence analysis results are similar to previous reports.

### SARS Example Annotation (PP1a protein):

Porcine Transmissible Gastroenteritis main cysteine proteinase (Mpro) crystal structure (1lvo chain A, red) with an extra helical domain.

Human Rhinovirus 3C cysteine proteinase (3Cpro) crystal structure (1cqq chain A, cyan) with ligand AG7088 drug candidate (yellow)



Ligand binding site from the crystal structure of 3Cpro coincide with the predicted ligand binding site of SARS virus Mpro

Crystal structure of SARS virus Mpro (1q2w, blue) superimposed with DS GeneAtlas model (yellow) based on the 1lvoA as template with 3D annotation.

Backbone atom RMSD:  
2.45 Å over 295 residues  
1.32 Å over 220 residues at the predicted ligand binding site

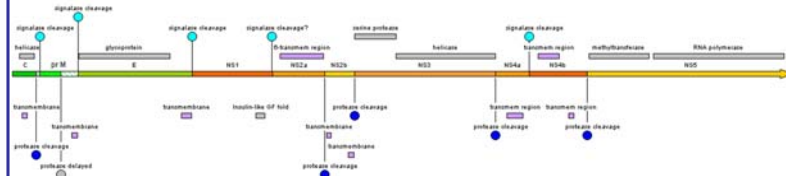
Model binding pocket predictions displayed as (orange) surface displays

A domain in the pp1a protein (residues 3241 to 3543) is homologous to the coronavirus main cysteine proteinase (Mpro, red) of the porcine transmissible gastroenteritis coronavirus. The 3D model generated from DS GeneAtlas (yellow) is built based on this template, 1lvo.pdb, chain-A (see figure above). This template shares a common fold with 1cqq.pdb, chain-A from the human rhinoviral protease with a bound inhibitor (yellow, CPK representation), a 3C cysteine protease (3Cpro, cyan). Note Mpro (1lvo.pdb) has an additional helical domain at the C-terminus compared to 3Cpro. Mpro and 3Cpro have very low sequence similarity given the common fold; they have less than 10% sequence identity and their ligand binding pockets are located at the same position in the cleft between the two beta domains and may bind to similar ligand. After our calculation was completed, the crystal structure (1q2w.pdb, blue) of the SARS virus Mpro was determined, which showed high structural conservation especially in the core folding region and ligand binding site.

Overall, DS GeneAtlas confirmed the existing SARS annotations, as well as three additional novel annotations that are potentially new targets for drug discovery (7 Transmembrane protein, FtsJ methyltransferase, and Fusion protein which are circled in the above figure).

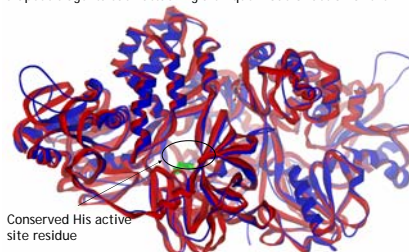
### WEST NILE VIRUS

All 12 proteins encoded by the WNV genome are displayed below using DS Gene (signalase and protease enzymatic cleavage sites are shown in cyan and blue circles). DS GeneAtlas was used to assign functional and structural annotations to these proteins (grey and purple rectangles). Homology models and active site annotation for the structural proteins, capsid (C) and envelope (E), and non-structural proteins, NS1, NS3 and NS5 were produced. A detailed analysis of each protein functional annotation is available (see WNV references below).

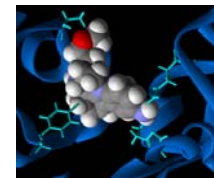


### WNV Example Annotation (NS3 Protein):

After reviewing the literature and the accuracy of the homology models, the NS3 protease was selected as the target for lead generation. The NS3 protein plays a key role in viral replication and has been cited as a promising therapeutic target. In addition, at the time of this study, there were no known inhibitors to this protein target. The NS3 protein (619 residues) model was generated first from the A-chain of 1cu1.pdb by DS GeneAtlas and then the protein was remodeled with MODELER (final model in blue in the figure below) using the complete (A and B-chains) 1cu1.pdb template (red) (superimposition = 1.38 Å over 1173 residues). The model has only 17% sequence identity to the template. This template (1cu1.pdb) and several others determined through DS GeneAtlas have excellent model scores (0.8 to 0.9 model scores) and indicate a bifunctional protein with both serine protease and RNA helicase functions. The model below shows a conserved Histidine (His) residue (green) in the active site of the serine protease domain of the N-terminus. The NS3 model, with the conserved His in the active site, was the selected protein target for lead generation using LUDI and AUTOLUDI. Using the scoring function LUDI3 as a measure, results of a virtual high throughput screening experiment were focused on a small set of diverse molecules (the image below, in CPK, shows the best hit) that have the potential to inhibit the NS3 protein and minimize the chance of ADME/Tox failures in future development. However, this is the first of many possible steps. Future plans include pursuing a variety of new targets including the helicase functionality of NS3, NS5 or the envelope protein as well as a collaboration to verify our findings for the NS3 protein. This provides an even greater benefit to halting the effect of the WNV through a 'drug-cocktail' approach - multiple therapeutic agents each attacking a unique mode of action of the virus.



Conserved His active site residue



The top molecule hit from AUTOLUDI ranking using the LUDI3 Score makes several contacts with the NS3 protein at key residues, including the active site His residue.

## References

- DS GeneAtlas (Accelrys Inc. San Diego)
- Kitson, D.H., et al. *Briefings in Bioinform.*, 3 (2003) 32-44.
- DS Gene, MODELER, LUDI and AUTOLUDI software (Accelrys Inc. San Diego)
- SARS:
  - Yan, L., et al. *FEBS Letters*, 554 (2004) 257-263.
  - Case Study:  
[http://www.accelrys.com/reference/cases/studies/sars\\_genome\\_annot.pdf](http://www.accelrys.com/reference/cases/studies/sars_genome_annot.pdf)
- WNV:
  - Fisher F., Quinn AM. & Haley-Vicente, D. submitted to *J. Proteome Res.* (2005)
  - Case Study:  
[http://www.accelrys.com/reference/cases/studies/west\\_nile\\_virus.pdf](http://www.accelrys.com/reference/cases/studies/west_nile_virus.pdf)

**Conclusion** This study shows how high throughput functional annotation using a multitude of methods from PSI-BLAST to 3D annotation prediction allows one to determine the function and 3D structure of proteins despite low sequence identity. The 3D structures can then be used for SBDD. We also supported these methods by functionally annotating a SARS and WNV protein. A detailed analysis of the entire genomes is available (see references).

Email Dana Haley-Vicente at [dhw@accelrys.com](mailto:dhw@accelrys.com) for more information.