From Gene to Function: In Silico Warfare on the West Nile Virus

Anne Marie Quinn, Luke Fisher and Dana Haley-Vicente

Accelrys Inc., San Diego, CA

Can we produce reliable structural models of proteins encoded by viral genomes even when the sequence identity among homologs is low? Such structural information about viral proteins is critical to further efforts to design drugs to fight the onslaught of disease. Here we show how Discovery Studio® (DS) Gene, DS Modeling, DS GeneAtlas®, and DS MODELER can be used to produce reliable structural and functional annotation of the proteins encoded by the West Nile Virus (WNV) genome.

Introduction

The WNV is a single-stranded RNA flavivirus. It has a single open reading frame encoding a polypeptide that is cleaved by co- and post-translational processing to produce protein products. No crystal or NMR structures have been produced for any of these proteins although a low-resolution cryo-electron microscopy structure was determined for the virus[1].

Structural information can be used to classify protein function and to identify the potential binding sites of new drug targets to combat the virus. The use of 3D homology models in the absence of an atomic resolution structure has been shown to be an effective alternative to study the structure and search for new lead candidates using structure-based drug design (SBDD). A recent publication employs the DS GeneAtlas pipeline to annotate the SARS virus in preparation for SBDD[2]. Here we apply DS GeneAtlas to the WNV proteome to functionally annotate the proteins and determine potential binding sites.

Methods

DS GeneAtlas is an automated, high-throughput, computational pipeline for structural and functional protein annotation[3]. The DS GeneAtlas pipeline predicts domains using TransMem and HMMer/Pfam[4], finds sequence similarity search using forward and reverse PSI-BLAST, predicts secondary structure using DSC[5], uses SeqFold for fold recognition, builds homology models using MODELER[6], and does 3D annotation using binding site searching and 3D-motif predictions[7].

Industry Sector

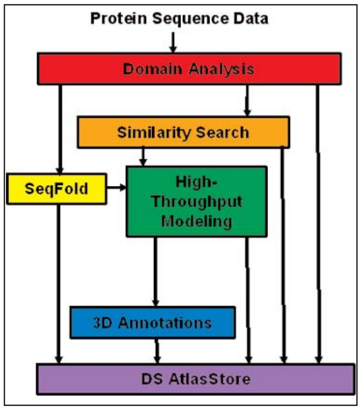
Pharmaceuticals

Organization

Accelrys

Key Products

DS Gene, DS Modeling, DS GeneAtlas, DS MODELER



▲ *Figure 1: The DS GeneAtlas High Throughput Pipeline*

Domain Analysis
- Low complexity Filtering
- TransMEM: Transmembrane prediction
- HMMer Pfam: Profile Analysis

Similarity Search
- PSI-BLAST nr and PDB (nr95) databases
- Forward and reverse search

SeqFold
- Fold Recognition for detecting low similarity templates

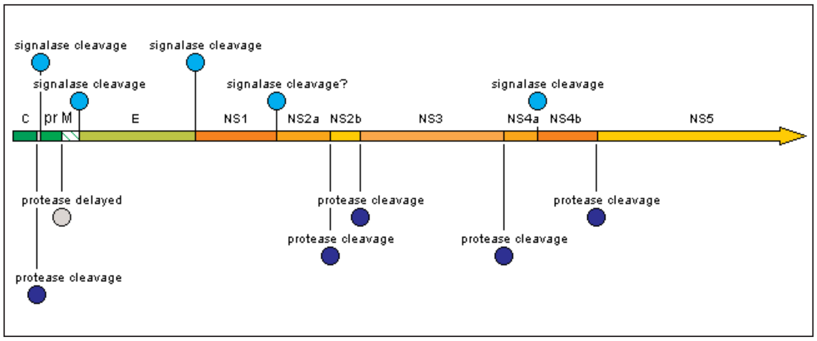
High throughput modeling (HTM)
- MODELER
- Create 3D models with single templates from PSI-BLAST or SeqFold hits
- Create 3D models from pre-aligned multiple templates from protein families (TFAMs)

3D Annotations
- 3D motif search (graph theory approach)
- PDB active site annotation
- Shape-based cavity finding

DS AtlasStore
- Manage, query and retrieve results

Alignments for homology modeling are selected from either PSI-BLAST or SeqFold hits reaching a predetermined cutoff value. Models are scored based on the Verify/Profiles-3D and Potential Mean Force (PMF) scores. The Verify/Profiles-3D score measures the compatibility of a protein model with its sequence using a 3D profile[8], while the PMF score uses statistical potentials of mean force to assess the accuracy of a model[9]. The final model score that is used to select the best models is used to measure the confidence of the structural annotation generated from the models. This score is the normalized combination of the model Verify/Profiles-3D and PMF scores. Models with values higher than 0 are considered to be sufficiently reliable.

DS Gene software for molecular sequence analysis was used to identify the signalase and dibasic viral protease cleavage sites in the WNV genome sequence. Twelve protein products derived from the polyprotein sequence of the genome were then analyzed with DS GeneAtlas. These protein sequences include the capsid (C), membrane (M), and envelope (E) structural proteins, as well as the non-structural proteins NS1, NS2a, NS2B, NS3, NS4A, NS4B, and NS5 (Figure 2). The precapsid includes the C-terminal signal sequence for the M-protein. PreM is the mature M protein with the signal sequence cleaved, and M is the fully processed protein.



▲ *Figure 2: The West Nile Polyprotein.*
DS Gene was used to annotate the WNV polyprotein sequence with the co-translational and post-translational cleavage sites reported by Chambers et. al.[10].

Results and Discussion

Reliable structural and functional annotations were generated for the capsid, envelope, NS1, NS3, and NS5 proteins (see Table 1). Note that despite low sequence similarity and high (worse) PSI-BLAST values, DS GeneAtlas can still predict good models (high model scores) and thus associate homologous function with the protein databank (PDB) template protein(s). The membrane, NS2a, NS2b, NS4a, and NS4b proteins had low similarity to PDB template proteins and bad model scores and thus are less reliable, but useful information was still revealed by DS GeneAtlas through other methods such as HMMer Pfam. Only the capsid, envelope, NS1, NS3, and NS5 proteins will be discussed below; however, more details are available upon request.

Protein	Description of Template Function	DS GeneAtlas Methods*	PDB Template	Scores**
Capsid Protein	Post-translationally modified by c-terminal cleavage	HTM	1bvs DNA BINDING PROTEIN (RUVA)	PSI-BLAST E =0.04 % Identity = 25% Model Score = 0.11
Envelope Protein	Envelope protein 291-791	HTM and PB90	1oam DENGUE 2 VIRUS ENV PROTEIN	PSI-BLAST E = 0 % Identity = 46% Model Score = 0.99
NS1	Non-structural protein 1	HTM and PB90	1gzy INSULIN-LIKE GROWTH FACTOR	PSI-BLAST E = 0.001 % Identity = 40% Model Score =0 .63
NS2a	Non-structural (multiple transmembrane domain) protein 2a	HTM	1jbi COCHLIN CHAIN A; LCCL MODULE.	PSI-BLAST E = 0.45 % Identity = 37% Model Score = 0.17
NS3	Nonstructural protease complex	HTM and PB90	1cu1 HYDROLASE	Psi_BLAST E = 2.6e-98% Identity = 19%Model Score = 0.88
NS5 6-269	Uncleaved, non-structural conserved cytoplasmic protein	HTM and PB90	1l9k DENGUE METHYLTRANSFERASE	PSI-BLAST E = 0 % Identity = 63% Model Score = 1.0
NS5 370-777	Uncleaved, non-structural conserved cytoplasmic protein	HTMM And PB90	1c2p-1gx5-[1nb4] (Template family) Viral RNA polymerase	PSI-BLAST E = 0 % Identity = 20% Model Score = 0.80

▲ *Table 1: Highlights of the DS GeneAtlas results for some of the West Nile proteins. Only one hit for each protein sequence with the best PSI-BLAST E-Value is shown.*

**HTM = high throughput modeling, HTMM = High throughput multiple template modeling, PB90 = PSI-BLAST against NR database (GenBank) at 90% sequence identity or less*

***Model score corresponds to linear combination score.*

The capsid protein was assigned a single transmembrane domain by TransMEM from residue position 44 to 66. Approximately the same region, position 44 to 58, was predicted to be helical by secondary structure analysis. This region may correspond to the residues that anchor to the WNV membrane. HTM indicates nucleotide binding functionality when evaluating the PDB template functionality. One binding site (178 Å3) was also predicted and may correspond to the RNA binding site. Note that the WNV is an RNA virus, thus differences in the binding site residues may distinguish RNA binding from the template DNA binding.

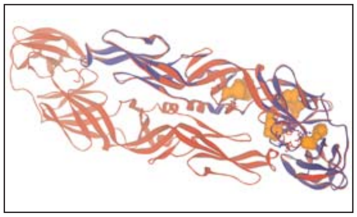
The best PDB template based on model score (0.99) for the E (Envelope) protein is the crystalline structure of the Dengue E protein (Figure 3). This hit was determined by a combination of methods in DS GeneAtlas (SeqFold, HTM, and PB90). At the sequence level, the Dengue and West Nile proteins are 46% identical and 64% similar.

This model spans the length of the West Nile protein, from position 3 to 400, excluding the last 101 amino acids of the C-terminus. Transmembrane analysis indicates that the C-terminal region anchors this protein in the viral membrane. The template structure is in complex with N-Octyl-Beta-D-Glucoside, lending the potential for homology-based docking experiments on the West Nile E protein, as many of the residues in one of the four binding pockets are conserved. Further annotation by HMMer indicates an immunoglobulin-like (Ig-like) domain from position 291 to 409 in the protein sequence. This Ig-like fold is also seen in the homology modeling. The active site, predicted by CSC[7], shows a conserved 3D motif (Thr-Asp-His) in a cavity. To better represent the biological assembly of the E protein, the entire dimer must be remodeled (i.e. using MODELER).

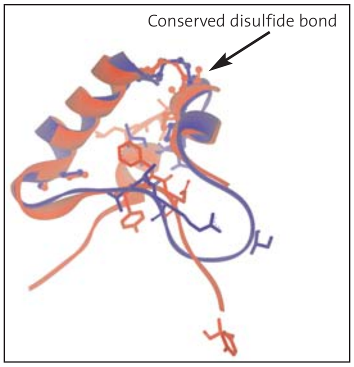
Seven homology models for NS1 with good model scores ranging from 0.2 to 0.6 correspond to 40 residues in the C terminus from position 279 to 318 in the primary sequence. All of the templates for this domain are in the same SCOP family (Structural Classification of Proteins) and indicate an insulin-like growth factor fold with a conserved disulfide bond. The best of model with the highest model score is based on the PDB entry 1gzy (B chain), the crystalline structure for insulin-like growth factor I (Figure 4).

The DS GeneAtlas pipeline initially identified a good model for the NS3 protein from the alpha chain of the PDB entry 1cu1 (A-chain), a hydrolase protein. This was remodeled in DS Modeling (DS MODELER) using both chains (A and B) of the 1cu1 template (Figure 5). The dimer template protein is from the Hepatitis C virus. The results indicate a bifunctional protein with both serine protease and RNA helicase properties. The active site of the protease features a conserved histidine residue. This template as well as several others determined by DS GeneAtlas produced excellent model scores ranging from 0.8 to 0.9 despite low sequence similarity between the NS3 target and templates. Fold recognition using SeqFold reports compatible secondary structure with good alignment and larger coverage across the length of the sequence.

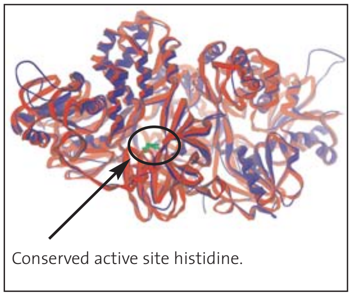
HMMer Pfam analysis in DS GeneAtlas identified two distinct domains for NS5-an N terminal FtsJ-like methyltransferase domain and a C-terminal RNA polymerase domain-thus indicating a bifunctional protein (Figure 6). These two domains cover the majority of this 905 residue protein, but a complete model was not available that covered both domains. The 119KDa structure from Dengue virus methyltransferase serves as a template to build a model of the first domain. At the sequence level, these proteins are 63% identical and yield a model with a predictably high reliability rating. The polymerase domain was created from a multiple template model using the alignment of the template structures 1gx5, 1c2p, and 1nb4, all from Hepatitis C virus. All of the templates are part of the RNA-dependent RNA polymerase SCOP family.



▲ *Figure 3: Model of the West Nile envelope protein shown in blue, superimposed on Dengue virus envelope protein template shown in red. 501 residues overlap for good coverage, with high sequence identity and predicted binding sites shown in orange.*

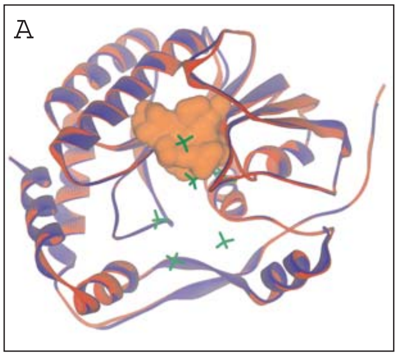


▲ *Figure 4: A model of the region near the C terminus of the West Nile NS1 protein shown in blue, based on the superimposed template from PDB entry 1gzy, shown in red. Sequence identity with the template is 40% over 40 residues shown here.*

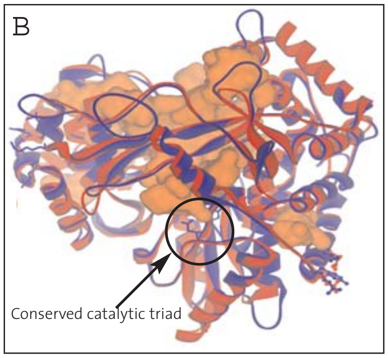


▲ *Figure 5: A model of the West Nile NS3 protein (blue) superimposed with the 1cu1 dimerized template protein from Hepatitis C virus (red). Similarity between the template and NS3 sequences is only 19%. The superimposition = 1.38 Å RMSD over 1173 residues.*

The alignment is based on structural properties and sequence conservation of all the templates. The percent sequence identity between WNV NS5 polymerase domain and any of the template sequences is less than 19%. The active site annotation in the template sequences is perfectly conserved in the WNV sequence, residues D536, D668, and D669 and the model score is high (see Table 1), illustrating the advantage of the multiple template approach to generate reliable homology models, especially when sequence similarity is low.



▲ *Figure 6A:*
A model of the N-terminal domain of the West Nile NS5 protein is shown in blue, with the Dengue methyltransferase template shown in red (superimposition = 0 .22 Å over 260 residues). The dengue protein was crystallized in the presence of S-adenosyl-l-homocysteine and sulfate ions (green). The homocysteine active site is indicated in orange. The green atoms correspond to sulfate ions.



▲ *Figure 6B:*
A model of the C-terminal domain of the West Nile nonstructural NS5 protein shown in blue, superimposed with the alpha chain of the 1nb4 template, the crystal structure of the Hepatitis C virus RNA polymerase in red (superimposition = 1.06 Å RMSD over 495 residues).

Conclusions

This study shows how the DS GeneAtlas high-throughput structural annotation pipeline can help to characterize proteins of unknown function. 3D Models generated by the pipeline can then be used for structure-based drug design. The predicted structure of the protein, combined with active site annotation, can reveal the residues involved in substrate recognition and can provide a structural basis for exploring mutational effects on activity. DS GeneAtlas was used to predicte reliable 3D models for the WNV proteins NS3 and NS5, despite very low sequence similarity with the available templates. Further analysis is underway now using the Ludi program for rational drug design to generate a specific serine-protease inhibitor against the NS3 protein.

A poster presentation of the WNV analysis presented at the ISMB/ECCB 2004 conference in Scotland (August 2004) is available upon request.

References

1. Mukhopadhyay, S. *et al.*, Science **2003**, *302*, 248.
2. Yan, L., *et.al., FEBS Letters* **2003**, *554*, 257-263.
3. Kitson, D.H., *et al., Briefings in Bioinformatics* **2002**, *3*, 32-44.
4. Bateman, A., et al., *Nucleic Acids Research* **2002**, *30*, 276-280.
5. King, R.D. and Sternberg, M.J.E., *Protein Science* **1996**, *5*, 2298.
6.Sali, A. and Blundell T.L., *J. Mol. Biol.* **1993**, *234*, 779-815.
7. Milik, M. *et al., Protein Engineering* **2003**, *16*, 543-552.
8. Fischer D. and Eisenberg D., *Protein Science* **1996**, *5*, 947-955.
9. Melo, F. Sanchez, and R. Sali, A. *Protein Science* **2002**, *11*, 430-448.
10. Chambers, T.J., *et al., Annual Review of Microbiology* **1990**, *44*, 649-688.