Assessment of putative protein targets derived from the SARS genome[1]

Lisa Yan, Mikhail Velikanov, Paul Flook, Wenjin Zheng, Sándor Szalma, Scott Kahn*

Accelrys Inc., 9685 Scranton Road, San Diego, CA 92121, USA

Accepted 24 September 2003

First published online 23 October 2003

Edited by Robert B. Russell

Abstract The ability to rapidly and reliably develop hypotheses on the function of newly discovered protein sequences requires systematic and comprehensive analysis. Such an analysis, embodied within the DS GeneAtlas™ pipeline, has been used to critically evaluate the severe acute respiratory syndrome (SARS) genome with the goal of identifying new potential targets for viral therapeutic intervention. This paper discusses several new functional hypotheses on the roles played by the constituent gene products of SARS, and will serve as an example of how such assignments can be developed or extended on other systems of interest.

Key words: Severe acute respiratory syndrome genome; GeneAtlas pipeline; Protein structure prediction; Protein function annotation; High-throughput pipeline; Homology modeling

1. Introduction

The ability to respond quickly to new contagions is made ever more necessary in the 21st century as global transmission vectors become more commonplace, and particularly so when said contagions pose a mortal threat to human health. This tenet is well supported by the current problems being presented by severe acute respiratory syndrome (SARS) on several continents [1,2], even more so as SARS is a viral infection whose genetic identity has only recently been reported [3,4]. Inspection of the SARS genome using a battery of bioinformatic techniques suggests some of the gene products that might be targeted as a site of intervention, albeit the majority of the genome escapes functional assignment [3,4]. The purpose of this note is to introduce the use of an automated structural proteomic pipeline to extend the assignment of putative function for the gene products of SARS, and in so doing offer several new targets for further assessment. Key to this approach is the development of robust functional assignments achieved by relying on multiple methods of analysis that are highly integrated.

Following previous analysis of the SARS genome [3,4], orf1a and orf1b are believed to be poly protein constructs that are cleaved to form the constituent proteins post transcription. When taken with the third long coding region for the S (Spike) protein, these three sequences are likely to be a fertile source or targets through which a therapeutic might be developed. While this study considered all of the open reading frames (orfs) within the genome, all hypotheses developed are isolated to these first three regions of the genome. Our strategy has been to employ the validated methods of protein function assignment embodied by the DS GeneAtlas™ pipeline, consisting of a cacophony of methods from both bioinformatic and structural biology fields, all of which is described elsewhere [5]. Such an automated approach is only enabled through the comprehensive integration of the respective methods, leveraging the strengths and weaknesses of each to develop robust assignments.

The genome data and the protein transcripts were extracted from GenBank for all the known isolates of SARS virus. Analysis of the variation of the different strains was performed to ensure that all non-synonymous polymorphisms were consistent with the assignments being proposed. Moreover, using predicted protein sequences from different strains and protein products based on putative cleavage sites has shown that the functional assignments discussed herein are robust and invariant to the specific protein sequences used as input to the DS GeneAtlas pipeline. In many cases the assignments of all resulting protein sequences contained inferences to structural homologies, albeit via sequence homologies that are below 30%. The robustness of such assignments follows from cooperating evidence obtained from protein threading and more traditional bioinformatics techniques, but also is afforded the automated and manual inspection of the structural alignment with respect to known active site residues, preservation of binding site residues, defined secondary structure, and conserved residues known to be involved in the function of the protein.

2. Materials and methods

DS GeneAtlas is an automated high-throughput pipeline for the prediction of protein structure and function [5]. The pipeline consists of transmembrane domain prediction using TransMem [6], domain prediction using HMMer and Pfam, sequence similarity search using PSI-BLAST, fold recognition using SeqFold [7], and homology modeling using MODELER [8]. The confidence of the annotations is often judged based on the consensus of several methods. For structural annotation, 'Consensus Score' is used to rank the confidence of the annotation. 'Consensus Score' is defined as 'Model Score' plus the sum of the active site percentage identity minus a threshold factor of 0.2 where the sum is over all the reported active site records in template PDB file. If there is no active site record in template, 'Consensus Score' is the same as 'Model Score'. The confidence is high if 'Consensus Score' is greater than 1.0, the confidence is medium if the 'Consensus Score' is between 0.0 and 1.0, and the confidence is low if 'Consensus Score' is less than 0.0. For Pfam (http://www.sanger.ac.uk/Software/Pfam/) annotations, if the *e*-value is less than 1e−5 the

*Corresponding author. Fax: (1)-858-799 5100.
E-mail address: skahn@accelrys.com (S. Kahn).

[1] Supplementary data associated with this article can be found at doi:10.1016/S0014-5793(03)01115-3

0014-5793/03/$22.00 © 2003 Federation of European Biochemical Societies. Published by Elsevier B.V. All rights reserved.
doi:10.1016/S0014-5793(03)01115-3

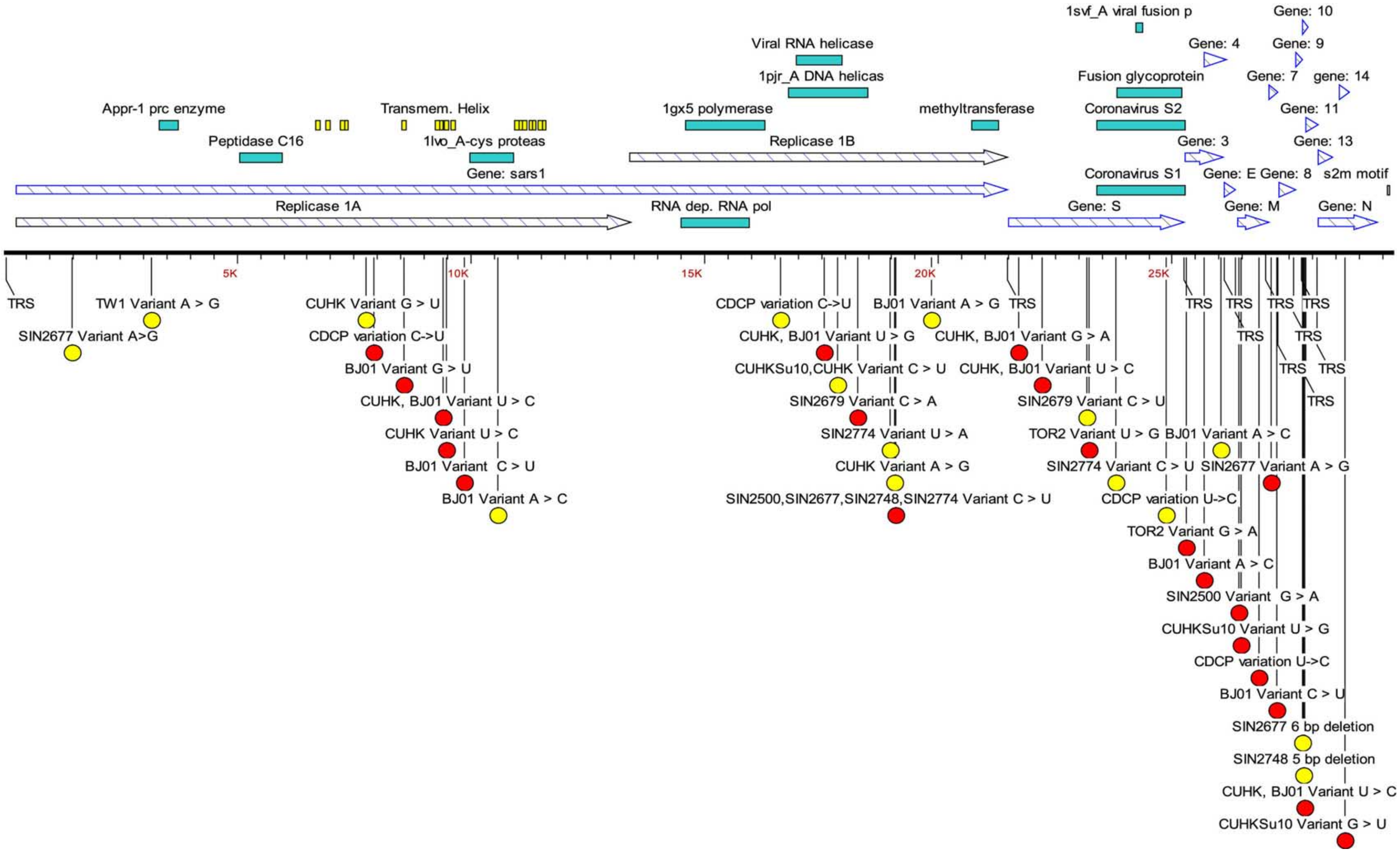


Fig. 1. DS GeneAtlas structural and functional annotations (in green), putative protein transcripts (in blue arrow), non-synonymous SNPs (in red dots) and synonymous SNPs (in yellow dots) mapped to the Tor2 genome sequence. The figure is created using DS Gene software from Accelrys, Inc.

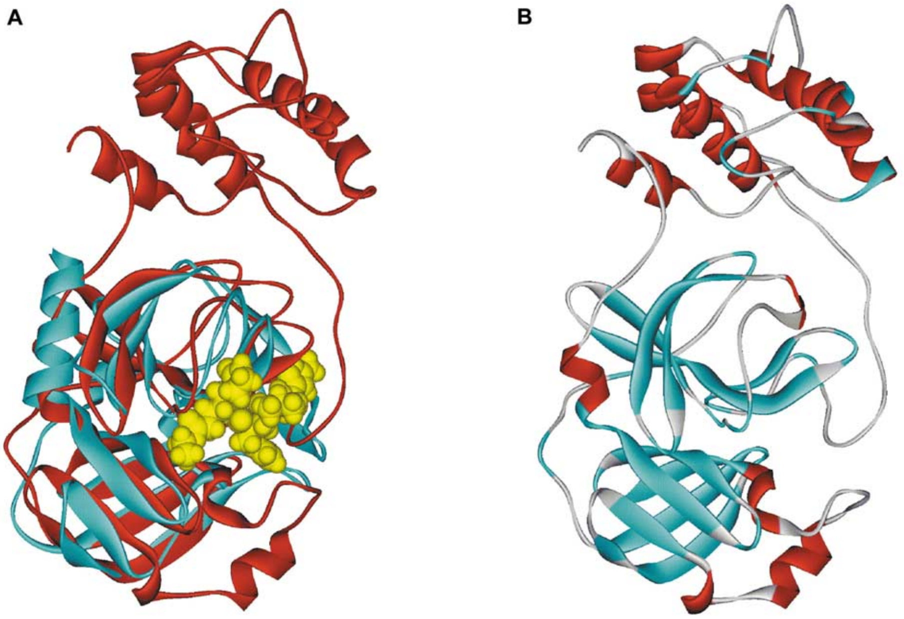


Fig. 2. A: Superimposed crystal structures of M^{pro} (1lvo_A in red) and 3C^{pro} (1cqq_A in cyan) with ligand AG7088 (in yellow). B: Model structure of SARS M^{pro} (residues 3241–3543) predicted using DS GeneAtlas. The figure is created using DS Modeling 1.1 software from Accelrys, Inc.

confidence is high; if *e*-value is between 0.1 and 1e−5 and the bit score is better than the noise cutoff [9], the confidence is medium; otherwise, the confidence is low.

Polyproteins pp1a and pp1ab are cleaved by a 3CL main protease and the papain-like accessory proteinase. We used the polyproteins pp1a and pp1ab as well as the cleaved products defined by the putative cleavage site as input to DS GeneAtlas. M^{pro} cleavage sites were identified using the [GAVSTP]XLQ[SAGN] motif, with cleavage occurring immediately after the Gln residue [10,11] yielding 11 cleavage sites. The papain-like protease (PLP) cleavage sites were identified only in the first (N-terminal) product of M^{pro} cleavage (residues 1–3240). Using the motif [RK]XXXG[GA], which is most consistent with all known data about PLP cleavage sites in coronaviruses [12,13], six cleavage sites were found.

3. Results

Several isolates of SARS virus have been reported [3,4]. The current analysis is based on the 11 protein sequences of BJ01 strain as well as the 15 protein sequences of the Tor2 strain (Table 1) from NCBI protein database (http://www.ncbi.nlm.nih.gov/). Structural and functional domains have been annotated for three sequences, polyprotein 1a (pp1a), polyprotein 1b (pp1b), and S protein. For other sequences, very little significant homology has been found with any known structural templates and the sequence analysis results are similar to previous reports by others in the analysis of the Tor2 [3] and CDCP [4] strains. The analysis based on the BJ01 strain and the Tor2 strain using the DS GeneAtlas pipeline resulted in the same annotation for all the SARS protein sequences. All predicted structural domains and functional domains are mapped to the genome sequence of Tor2 (Fig. 1) for subsequent discussion.

3.1. *pp1a*

A domain in pp1a from residues 3241 to 3543 is found to be homologous to coronavirus main cysteine proteinase (M^{pro}) of another coronavirus, porcine transmissible gastroenteritis virus (TGEV) [14] and the three-dimensional (3D) model is built based on template 1lvo_A. Model scores and confidence are

Table 1
Protein transcripts of SARS isolates and their genomic coordinates

BJ01	Tor2	CDCP	Start	Stop	Frame	Actual start	Actual stop	AA length
1a	1a	1a	265	13 398	1	265	13 413	4 382
1b	1b	1b	13 398	21 485	3	13 398	21 485	2 695
S	S	S	21 492	25 259	3	21 492	25 259	1 255
1	3	X1	25 268	26 092	2	25 268	26 092	274
2	4	X2	25 689	26153	3	25 689	26 153	154
E	E	E	26 117	26347	2	26 117	26 347	76
M	M	M	26 398	27 063	1	26 398	27 063	221
3	7	X3	27 074	27 265	2	27 074	27 265	63
4	8	X4	27 273	27 641	3	27 273	27 641	122
N/A	9	N/A	27 638	27 772	2	27 638	27 772	44
N/A	10	N/A	27 779	27 898	2	27 779	27 898	39
N/A	11	X5	27 864	28 118	3	27 864	28 118	84
N	N	N	28 120	29 388	1	28 120	29 388	422
5	13	N/A	28 130	28 426	2	28 130	28 426	98
N/A	14	N/A	28 583	28 795	2	28 583	28 795	70
N/A	s2m motif	N/A	29 590	29 621	N/A	29 590	29 621	N/A

Table 2
Structural and functional annotations using DS GeneAtlas[a]

Domain	Methods	Template and function	Scores	Confidence[b]
pp1a 3241–3543	Structure	1lvo_A cysteine-like protease (TGEV)	PSI-BLAST e-value = 0 Model score = 0.96 Seq-ID% = 43.9%	Consensus score = 1.76 high
pp1a 1026–1154	HMMer/Pfam	Appr-1″-p processing enzyme family	e-value = 1.1e−20 Bit score = 78.3 Noise cutoff = −21.5	high
pp1a 1598–1893	HMMer/Pfam	Peptidase C16 family	e-value = 0.043 Bit score = −87.5 Noise cutoff = −95.7	medium
pp1b 4780–5334[c]	Structure	1gx5 RNA dependent RNA polymerase (HCV)	PSI-BLAST e-value = 9.2e−25 Model score = −0.10 Seq-ID = 10.3%	Consensus score = 0.5 medium
pp1b 4770–5249[d]	Structure	1gx5 RNA dependent RNA polymerase (HCV)	PSI-BLAST e-value = 1.4e−49 Model score = −0.21 Seq-ID% = 10.4%	Consensus score = −0.41 low
pp1b 5512–6066	Structure	1pjr_A DNA helicase	PSI-BLAST e-value = 3.4e−78 Model score = −0.23 Seq-ID% = 8.3%	Consensus score = −0.23 low
pp1b 4747–5219	HMMer/Pfam	RNA dependent RNA polymerase	e-value = 0.093 Bit score = −194.6 Noise cutoff = −130.1	low
pp1b 5569–5887	HMMer/Pfam	Viral (Superfamily 1) RNA helicase	e-value = 0.0058 Bit score = −49.6 Noise cutoff = −30.1	low
pp1b 6815–6998	HMMer/Pfam	Fts-J-like methyltransferase	e-value = 0.0044 Bit score = −51.6 Noise cutoff = −53.1	medium
S protein 910–949	Structure	1svf_A viral fusion protein core	PSI-BLAST e-value = 8e−05 Model score = −0.29 Seq-ID% = 17.5%	Consensus score = −0.29 low
S protein 631–1255	HMMer/Pfam	Coronavirus S1 glycoprotein	e-value = 0.87 Bit score = −283.4 Noise cutoff = −273.1	low
S protein 631–1255	HMMer/Pfam	Coronavirus S2 glycoprotein	e-value = 1.3e−132 Bit score = 450.2 Noise cutoff = −469.8	high
S protein 777–1231	HMMer/Pfam	Fusion glycoprotein F0	e-value = 0.031 Bit score = −276.3 Noise cutoff = −236.1	low

[a] The PDB files of the models listed in this table are provided at the following link with full DS GeneAtlas output of the SARS genome: http://www.accelrys.com/references/supplemental/.

[b] See Section 2 for the definition of the confidence.

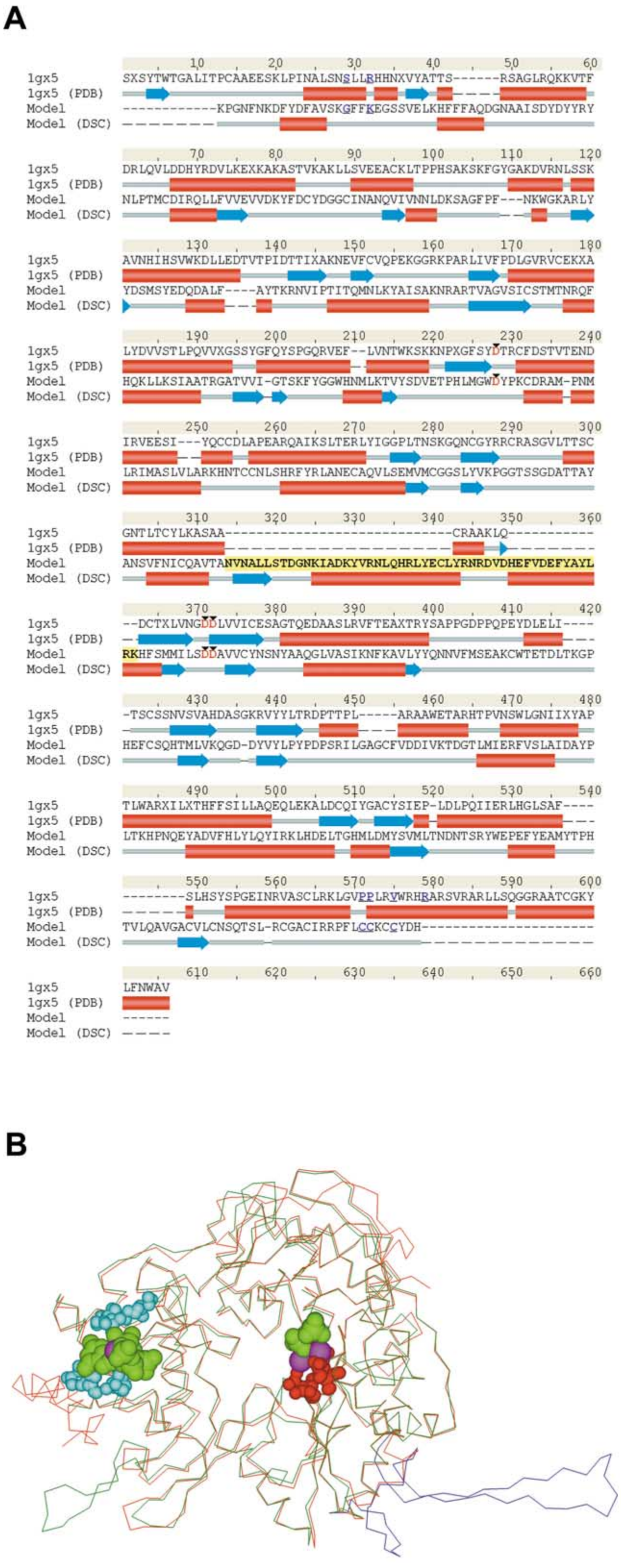
[c] Annotation using full sequence of pp1ab as input.

[d] Annotation using cleaved sequence from residues 4231 to 5301 of pp1ab as input.

listed with other annotations in Table 2. The 3D structure of M$^{\text{pro}}$ shares a common fold with the human rhinoviral protease, a 3C cysteine protease (3C$^{\text{pro}}$), except that M$^{\text{pro}}$ has an additional helical domain at the C-terminus. Although M$^{\text{pro}}$ and 3C$^{\text{pro}}$ have very low sequence similarity (less than 10% sequence identity), given the common fold, their ligand binding pockets are located at the same position in the cleft between the two β domains (Fig. 2) and may bind to a similar ligand. Several cysteine protease inhibitors are studied including the rhinoviral protease inhibitor developed by Pfizer's La Jolla unit, Agouron, for treatment of common cold. The catalytic dyad (residue His41 and Cys144) is conserved among TGEV, SARS virus, and HCoV.

After our calculation was completed, two new crystal structures of M$^{\text{pro}}$, one from human coronavirus 229E (HCoV) and another one from TGEV were released [14]. The latter has the same sequence as the previously published TGEV structure, except with a bound ligand. The TGEV M$^{\text{pro}}$ structure of the apo protein and the ligand bound protein are similar with backbone rmsd less than 1.0 Å without major conformational changes around the ligand binding site. The percentage sequence identity between SARS M$^{\text{pro}}$ and TGEV M$^{\text{pro}}$ is 44% and is 41% between SARS and HCoV M$^{\text{pro}}$. Therefore, the original template that we used is still one of the best templates for creating the homology model of SARS M$^{\text{pro}}$. Subsequently, the crystal structure of the M$^{\text{pro}}$ of SARS genome was determined and deposited into the PDB databank (1q2w). The backbone rmsd between our predicted model and the X-ray structure is 2.45 Å over 295 residues. The structure is more conserved around the ligand binding site, the backbone atom rmsd is 1.32 Å over 220 residues.

Pfam [9] searching using HMMer [15] found two additional domains in pp1a, Appr-1 domain and peptidase C16 family. The Appr-1 domain is identified with high confidence and is found in a number of unrelated proteins, e.g. in the C-terminus of the macro-H2A histone protein, in the non-structural proteins of several types of ssRNA viruses such as NSP3 from alphaviruses, in a family of proteins from bacteria, archaebacteria and eukaryotes, suggesting that it is involved in an important and ubiquitous cellular process. Peptidase C16 is a cysteine protease and is often referred to as PLP. Coronavirus has one or two PLP proteins and only one is found in SARS virus by Rota, et al. [4] from residues 1632 to 1847.


A
10 20 30 40 50 60
1gx5 SXSYTWTGALITPCAAEESKLPINALSNSLLRHHNXVYATTS------RSAGLRQKKVTF
1gx5 (PDB)
Model ------------KPGNFNKDFYDFAVSKGFFKEGSSVELKHFFFAQDGNAAISDYDYYRY
Model (DSC)
70 80 90 100 110 120
1gx5 DRLQVLDDHYRDVLKEXKAKASTVKAKLLSVEEACKLTPPHSAKSKFGYGAKDVRNLSSK
1gx5 (PDB)
Model NLPTMCDIRQLLFVVEVVDKYFDCYDGGCINANQVIVNNLDKSAGFPF---NKWGKARLY
Model (DSC)
130 140 150 160 170 180
1gx5 AVNHIHSVWKDLLEDTVTPIDTTIXAKNEVFCVQPEKGGRKPARLIVFPDLGVRVCEKXA
1gx5 (PDB)
Model YDSMSYEDQDALF----AYTKRNVIPTITQMNLKYAISAKNRARTVAGVSICSTMTNRQF
Model (DSC)
190 200 210 220 230 240
1gx5 LYDVVSTLPQVVXGSSYGFQYSPGQRVEF--LVNTWKSKKNPXGFSYDTRCFDSTVTEND
1gx5 (PDB)
Model HQKLLKSIAATRGATVVI-GTSKFYGGWHNMLKTVYSDVETPHLMGWDYPKCDRAM-PNM
Model (DSC)
250 260 270 280 290 300
1gx5 IRVEESI---YQCCDLAPEARQAIKSLTERLYIGGPLTNSKGQNCGYRRCRASGVLTTSC
1gx5 (PDB)
Model LRIMASLVLARKHNTCCNLSHRFYRLANECAQVLSEMVMCGGSLYVKPGGTSSGDATTAY
Model (DSC)
310 320 330 340 350 360
1gx5 GNTLTCYLKASAA-----------------------------CRAAKLQ-----------
1gx5 (PDB)
Model ANSVFNICQAVTANVNALLSTDGNKIADKYVRNLQHRLYECLYRNRDVDHEFVDEFYAYL
Model (DSC)
370 380 390 400 410 420
1gx5 --DCTXLVNGDDLVVICESAGTQEDAASLRVFTEAXTRYSAPPGDPPQPEYDLELI----
1gx5 (PDB)
Model RKHFSMMILSDDAVVCYNSNYAAQGLVASIKNFKAVLYYQNNVFMSEAKCWTETDLTKGP
Model (DSC)
430 440 450 460 470 480
1gx5 -TSCSSNVSVAHDASGKRVYYLTRDPTTPL-----ARAAWETARHTPVNSWLGNIIXYAP
1gx5 (PDB)
Model HEFCSQHTMLVKQGD-DYVYLPYPDPSRILGAGCFVDDIVKTDGTLMIERFVSLAIDAYP
Model (DSC)
490 500 510 520 530 540
1gx5 TLWARXILXTHFFSILLAQEQLEKALDCQIYGACYSIEP-LDLPQIIERLHGLSAF----
1gx5 (PDB)
Model LTKHPNQEYADVFHLYLQYIRKLHDELTGHMLDMYSVMLTNDNTSRYWEPEFYEAMYTPH
Model (DSC)
550 560 570 580 590 600
1gx5 --------SLHSYSPGEINRVASCLRKLGVPPLRVWRHRARSVRARLLSQGGRAATCGKY
1gx5 (PDB)
Model TVLQAVGACVLCNSQTSL-RCGACIRRPFLCCKCCYDH----------------------
Model (DSC)
610 620 630 640 650 660
1gx5 LFNWAV
1gx5 (PDB)
Model ------
Model (DSC)
B

←
Fig. 3. A: The sequence alignment between template (1gx5) and the model (Model) of RNA dependent RNA polymerase. The catalytic residues are annotated with carot (in red) and the residues in surface pocket are annotated with underline (in blue). The predicted secondary structure for the model sequence using the DSC method and the secondary structure of the template are also displayed in the alignment (helix in red and strand in blue). The long insertion in the model sequence where the structure is uncertain is colored with yellow background. B: Model structure (in green) of RNA dependent RNA polymerase domain superimposed with template structure 1gx5 (in red). The rGTP ligands from the template are shown in green and the Mn^{2+} ions are in purple. The catalytic triad is in red and the other ligand binding site is in cyan. The long insertion in the model where the structure is uncertain is colored blue.

Several transmembrane helices are predicted for pp1a using the TransMem [6] program (Fig. 1) and a large transmembrane domain from residues 3561 to 3774 is found. Based on the predicted cleavage site of the main proteases, this TM domain is part of the cleaved product from residues 3547 to 3919, and may represent a 7TM protein motif if the seven transmembrane helices are indeed distinct as currently assigned.

3.2. pp1b

Two structural domains, RNA dependent RNA polymerase and helicase are predicted (see Table 2) for pp1b by DS GeneAtlas pipeline. The polymerase and helicase domains are predicted by others based on sequence analysis methods [3,4], however, these are the first model structures created using homology modeling. Although the model scores are poor, the assignment is confirmed by Pfam analysis used in GeneAtlas pipeline. It should be noted that the predicted structure of the large insertion region for about 40 residues (colored with yellow background in Fig. 3A and blue in Fig. 3B) is tenuous. Homology models created automatically by the DS GeneAtlas pipeline are mainly used to confirm the low scoring regions in matches found by PSI-BLAST or SeqFold searches and to reduce false positives inherent in these methods [5]. Model regions with high sequence identity to a template with small insertions are generally more reliable than model regions with large insertions. Model regions with conserved functional motifs such as active sites or ligand binding sites are in general more reliably assigned than assignments absent these features. This is demonstrated by the model of M^{pro} where the ligand binding site has a low rmsd with the X-ray structure. These experiences support the conclusion that homology models created via an automated pipeline can provide useful insights on the fold of the protein, the mapped active sites and functional motifs from the template, and the location(s) of putative binding pockets. For subsequent structure-based studies such as ligand docking the models should be used with some caution as they often require further manual refinement to improve the sequence alignment and to regenerate an improved model.

A new functional assignment is made to residues 6815–6998 as Fts-J-like methyltransferase by Pfam analysis. There is no transmembrane domain predicted for pp1b.

DS GeneAtlas identified several known RNA dependent RNA polymerases from hepatitis C virus (HCV) as templates for domain 4780–5334. The model created based on the 1gx5

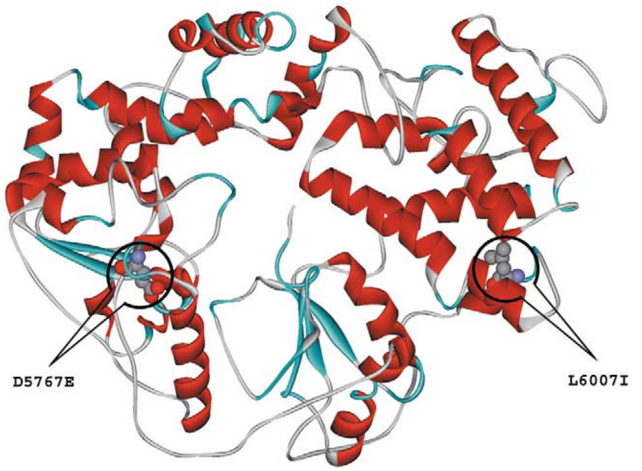


Fig. 4. Non-synonymous SNPs mapped to helicase structure.

template is the longest model with most complete structure and the catalytic triad Asp220, Asp318, Asp319 in the template is conserved in the model-template alignment (Fig. 3A). A crystal structure of the HCV polymerase has two ligand binding sites, one in the catalytic reaction center and one on the surface of the protein (Fig. 3B) about 30 Å from the catalytic center. For HCV, there is no need for a primer for the RNA replication and the riboguanosine triphosphate (rGTP) binding to the surface pocket is believed to be the initiation step of the RNA replication [16]. The residues in the surface binding site are not conserved between template and model which suggests that the SARS virus replication is likely to be activated by other means or by a different type of ligand. The structures are more conserved around the catalytic reaction center and there are no large gaps between model sequence and template sequence. On the other hand, several parts of the model structure on the surface are modeled with lower quality due to a few large insertions in the model sequence. Another model structure is created based on template 1c2p albeit more than 100 residues in the N-terminus are not modeled by this template. Notwithstanding, the model has a slightly better model score (0.18), and the sequence deletion is on the surface of the protein. For ligand docking to the catalytic center, this second model is likely to be a better structure to use for this protein target.

The sequence homology of SARS RNA helicase is very low to any RNA or DNA helicases with known structure and the helicase domain is modeled based on a DNA helicase.

A domain from residues 6815 to 6998 is annotated as Fts-J methyltransferase by Pfam using HMMer with medium confidence. Fts-J methyltransferase protein is also found in other viruses, such as flaviviral NS5 protein, and is involved in viral RNA capping which leads to stabilization of the RNA sequence [17].

3.3. S protein

The domain from residues 910–949 of S protein is found to have medium similarity to the viral fusion protein. A model is built based on chain A of template 1svf. The template adopts a coiled-coil fold with four helices from four monomers and is a repeat of two different monomers, chain A and chain B. We did not find any domain that is homologous to chain B and this can be attributed to the short length of chain B with only 38 residues. This is consistent with the knowledge that it is difficult to find sequence matches to short segments using the PSI-BLAST program. The similarity to viral fusion protein is confirmed by a Pfam search, even though the confidence level of the HMMer score is low. It is known that the C-terminal domain of coronavirus glycoprotein forms a coiled-coil structure during the cell membrane fusion process [4].

One transmembrane helix is predicted by the TransMem program at residues 1195–1217, which differs notably from the prediction by Rota at al. [4] of a longer 37 residue transmembrane domain. Our predicted TM domain is within the TM domain reported by Rota with a few polar residues from the N-terminal side and the cysteine rich segment from the C-terminal side removed.

3.4. Single nucleotide polymorphism (SNP) analysis

We found 35 SNPs for SARS virus by comparing the genome sequences from different SARS strains in the GenBank database (http://www.ncbi.nlm.nih.gov/Genbank/index.html). Twenty of them are non-synonymous mutations, i.e. result in amino acid residue changes (Fig. 1). Three non-synonymous SNPs V2770L, V3047A, and V3072A are inside the predicted transmembrane helices or near the end of the TM helix of pp1a. They are all conserved mutations which are unlikely to change the property of the transmembrane domain. Two additional SNPs, D5767E and L6007I, are found in the predicted helicase domain of pp1b (Fig. 4). Other SNPs are not mapped to any structural or functional domains that are predicted herein.

4. Discussion

Polyproteins pp1a and pp1ab are cleaved by a 3CL main protease and the papain-like accessory proteinase. Using directly the polyprotein or the cleaved products defined by the putative cleavage site as input to DS GeneAtlas pipeline resulted in overall similar levels of annotation, however, with slightly different quality scores for the predicted models. An example of the robustness of assignment is given for the RNA dependent RNA polymerase (see Table 2). The model predicted using the complete pp1ab sequence as input has a better overall quality compared to the model predicted using the cleaved sequence from residues 4231 to 5301 of pp1ab. The aspartic catalytic triad is conserved from template to model if the full pp1ab is the input sequence, whereas the triad is misaligned when the cleaved sequence is used. The alignment used to create the homology model is generated by PSI-BLAST which can yield slightly different profiles depending if a sequence is used as input or a domain of that sequence is used as input. The difference is only significant when the homology between template and model is extremely low. In the case of the main protease model, the cleaved sequence and the complete sequence resulted in exactly the same model.

3C-like main protease and RNA polymerase are crucial proteins for the survival of (+)sense ssRNA viruses. 3C-like main protease of coronavirus is well characterized and several known structures of type I coronavirus are determined by experimental methods with bound ligand. They are close homologs to the SARS coronavirus M^{pro}. The model structure of the M^{pro} based on TGEV has good quality and can be used in the de novo inhibitor design. The RNA polymerase of SARS virus exhibits remote homology to HCV polymerase which is an important drug target for treating hepatitis C infection. Also predicted are several other novel structural or functional domains that are not currently known, such as the Fts-J-like methyltransferase and the fusion peptide. Each of these represents a new hypothesis for subsequent verification and potential therapeutic intervention.

Recently, a meta server, 3D-Jury system, has been used to assign mRNA Cap-1 methyltransferase function to nsp13 [18]. This is consistent with our assignment of Fts-J-like methyltransferase to a domain in pp1b. Meta servers such as the 3D-Jury system [19] are based on the consensus score of many fold recognition programs and have been shown to be quite effective at the identification of correct folds for unknown proteins. However, fold recognition programs are often time consuming and the meta servers are not typically designed for high-throughput (whole genome) analysis as described herein. Moreover, DS GeneAtlas annotates non-structural regions of the available sequence with transmembrane domain predictions, and structural annotations that are important for protein function are automatically identified and mapped onto the model structures, for example, active site annotations from a template protein. While existing meta servers [19–21] focus on the use of structural modeling to annotate individual novel proteins, the present study combines protein modeling with mutation analysis and membrane spanning region prediction to provide a comprehensive functional view of the SARS genome in which assignments are correlated across a variety of analysis methods.

In summary, this paper reports comprehensive analysis of the SARS genome and proposes several novel hypotheses regarding the function of the gene products of the SARS genome using an automated assignment pipeline. Most of the hypotheses are supported by multiple methods of assignment in DS GeneAtlas. We propose that such automated development of functional hypotheses successfully extends what is known about SARS, and in most cases suggests follow-on experiments to refine these hypotheses further.

Acknowledgements: The authors would like to acknowledge the support and encouragement of our colleagues at Pharmacopeia Labs, and in particular Dr. Maria Webb.

References

[1] Drosten, C. et al. (2003) N. Engl. J. Med. 348, 1967–1976.
[2] Ksiazek, T.G. et al. (2003) N. Engl. J. Med. 348, 1953–1966.
[3] Marra, M. et al. (2003) Science 300, 1399–1404.
[4] Rota, P.A. et al. (2003) Science 300, 1394–1399.
[5] Kitson, D.H. et al. (2002) Brief. Bioinform. 3, 32–44.
[6] TransMem user manual, Accelrys, Inc., 2002.
[7] (a) Olszewski, K.A., Yan, L. and Edwards, D.J. (1999) Theor. Chem. Acc. 101, 57–61; (b) Olszewski, K.A. (2000) Proc. Pacific Symp. Biocomputing, pp. 143–154.
[8] Šali, A. and Blundell, T.L. (1993) J. Mol. Biol. 234, 779–815.
[9] Bateman, A. et al. (2002) Nucleic Acids Res. 30, 276–280.
[10] Ziebuhr, J., Snijder, E.J. and Gorbalenya, A.E. (2000) J. Gen. Virol. 81, 853.
[11] Hegyi, A. and Ziebuhr, J. (2002) J. Gen. Virol. 83, 595.
[12] Hughes, S.A., Bonilla, P.J. and Weiss, S.R. (1995) J. Virol. 69, 809.
[13] Ziebuhr, J., Thiel, V. and Gorbalenya, A.E. (2001) J. Biol. Chem. 276, 33220.
[14] Anand, K. et al. (2003) Science 300, 1763–1767.
[15] Durbin, R., Eddy, S., Krogh, A. and Mitchison, G. (1998), Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids, Cambridge University Press.
[16] Bressanelli, S., Tomei, L., Rey, F.A. and De Francesco, R. (2002) J. Virol. 76, 3482.
[17] Koonin, E.V. (1993) J. Gen. Virol. 74, 733–740.
[18] von Grotthuss, M., Wyrwicz, L.S. and Rychlewski, L. (2003) Cell 113, 701–702.
[19] Ginalski, K. and Rychlewski, L. (2003) Nucleic Acids Res. 31, 3291–3292.
[20] Kurowski, M.A. and Bujnicki, J.M. (2003) Nucleic Acids Res. 31, 3305–3307.
[21] Eyrich, V.A. and Rost, B. (2003) Nucleic Acids Res. 31, 3308–3310.