Evolutional insights on uncharacterized SARS coronavirus genes

Alex Inberg, Michal Linial*

Dept of Biological Chemistry, Life Science Institute, The Hebrew University, Jerusalem, 91904 Israel

Received 9 June 2004; revised 30 September 2004; accepted 30 September 2004

Available online 13 October 2004

Edited by Takashi Gojobori

Abstract The complete genome of the severe acute respiratory syndrome coronavirus (SARS-CoV) and many of its variants has been determined by several laboratories. The genome contains fourteen predicted open reading frames (ORFs). However, a function had been clearly assigned for only six of these ORFs, in the viral replication, transcription and structural constituents. The others are herein referred to as uncharacterized ORFs (UC-ORFs). Here, we try to provide a relational insight on those UC-ORFs, suggesting that a number of them are remotely related to structural proteins of coronaviruses and other viruses infecting mammalian hosts. Surprisingly, several of the UC-ORFs exhibit considerable similarity with other SARS-CoV ORFs. These observations may provide clues on the evolution and genome dynamics of the SARS-CoV.

Keywords: Viral evolution; Sequence homology; Remote homolog; Hypothetical protein; Coronavirus; Proteome

1. Introduction

A novel virus, discovered in April 2003, is responsible for the severe acute respiratory syndrome (SARS), a disease that was originally exposed in Guangdong Province, China in late 2002. The syndrome is a condition characterized by an atypical pneumonia, efficient transmission and high mortality rate. Presently, the SARS coronavirus (SARS-CoV) has spread to more than 30 countries all over the world. Over 8000 cases of SARS-CoV infected individuals were reported, with about 10% mortality. SARS-CoV may have originated in animals and its efficient human-to-human transmission is similar to that of the human immunodeficiency virus (HIV), and resulted in the global outbreak of the disease in the year 2003 [1,2].

The SARS-CoV genome is ~30 000 nucleotides long, with a few tens of nucleotides which vary among the different isolates [3–5]. All predicted open reading frames (ORFs) are divided into two groups: (i) those with a clear homology to other coronaviruses and for which viral functions are proposed; such ORFs include the replicase and the structural ORFs; (ii) those with no clear homology to any known genes and often referred to as non-structural ORFs [6]. Here, we define this latter set of ORFs with no proposed function or seemingly significant homology as uncharacterized ORFs (UC-ORFs).

All the UC-ORFs from ORF3a to ORF9b reside within only one-tenth of the virus's genomic length, clustered in ~3160 nucleotides region of the genome (Fig. 1, marked in gray). Following the complete sequencing of SARS-CoV genomes, homologies for UC-ORFs were sought. The results obtained from similarity search tools suggest some sporadic similarities for which supporting experimental evidence is still missing (a summary of all BLAST results is found at http://www.ncbi.nlm.nih.gov/genomes/SARS/sarsptt.html). SARS-CoV infected culture cells provided experimental evidence on the expression and translation levels for individual ORFs [7]. Yet, the participation of most of the UC-ORFs in the virulence and pathology of the virus is still questionable.

2. Materials and methods

2.1. Genomic information

All genomic information on SARS-CoV variants was taken from the non-redundant database of NCBI. Currently, a collection of more than 100 isolates is archived in the public database (http://www.ncbi.nlm.nih.gov/genomes/SARS/). The complete SARS genome NC_004718(29, 751 base pairs) was used as a reference with ORFs 1–9 terminology. The ORFs were divided into structurally and/or functionally known group (ORF1a, ORF1b, ORF2, ORF4, ORF5 and ORF9a) and the UC-ORFs group. The reference nucleotide coordinates for the UC-ORFs are as follows: ORF3a (25268–26092) ORF3b (25689–26153), ORF6 (27074–27265), ORF7a (27273–27641), ORF7b (27638–27772), ORF8a (27779–27898), ORF8b (27864–28118), and ORF9b (28130–28426). ORF9c (28583–29621) is considered false and was not included in our analysis.

2.2. Sequence similarity search

Sequence similarity searches were performed using WU-Blast2 (Washington University Basic Local Alignment Search Tool, Ver. 2.0), based on the European Bioinformatics Institute website (http://www.ebi.ac.uk/blast2/index.html). Searches were performed against the SwissProt database (version 43.4 ~150,000 entries), as well as against the non-redundant UniProt protein database (1.4 million sequences). The BLAST search parameters were adjusted [8] and applied with no filtration for low complexity sequences [9]. The results from multiple searches were compared using closely related substitution matrices for which the gap penalties were properly adjusted [10]. All presented results are based on applying the BLOSUM 40 and the PAM 200 substitution matrices.

2.3. Analysis tools

Prediction of transmembrane helices in the UC-ORFs was performed using the implementation of the TMHMM system [11] (http://phobius.cgb.ki.se/), which discriminates membrane proteins from soluble ones with a very high accuracy [12]. Multiple sequence alignment was performed using ClustalW program [13] using the default parameters.

* Corresponding author. Fax: +97-226-586-448.
E-mail address: michall@cc.huji.ac.il (M. Linial).

Abbreviations: SARS, severe acute respiratory syndrome; ORF, open reading frame; UC-ORF, uncharacterized open reading frame; CoV, coronavirus; HIV, human immunodeficiency virus

0014-5793/$22.00 © 2004 Published by Elsevier B.V. on behalf of the Federation of European Biochemical Societies.
doi:10.1016/j.febslet.2004.09.076

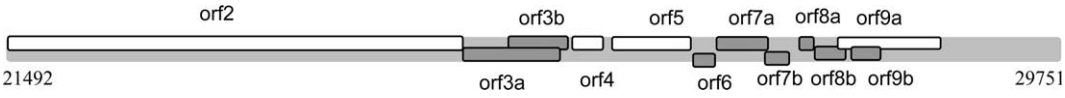

Fig. 1. A schematic illustration of nucleotides 21 492–29 751 from SARS-CoV (NC_004718). In white are ORFs encoding known structural genes and in gray are the eight putative uncharacterized ORFs (UC-ORFs). Note the overlap in sequences for most UC-ORFs.

3. Results

3.1. UC-ORFs' homology to viral proteins

We propose that even for UC-ORFs, whose translational potential is questionable, detection of similarity to other proteins by using customized and adapted search engines parameters is useful for gaining insight on the virus origin, homologies and genomic dynamics. We illustrate this notion by testing the similarity of some of the UC-ORFs against the current database of ~1.5 million protein sequences archived in UniProt (http://www.ebi.ac.uk/uniprot).

ORF3a is the longest predicted protein among the UC-ORFs (274 amino acids). This ORF was reported to exhibit no significant similarity to any known protein [4]. However, several viral E1 glycoprotein precursors of coronaviruses were detected as a result of a BLAST search against all proteins in SwissProt (applying non-default BLOSUM 40 or PAM 200 substitution matrices). Among the top hits (BLAST E-score ~1e-8), the matrix glycoprotein of coronaviruses from a wide range of hosts is prevalent. An example for the similarity of ORF3a to an E1 glycoprotein from Canine enteric coronavirus is shown (Fig. 2).

A similarity between ORF3a and other Matrix proteins of coronaviruses was illustrated by analyzing their multiple sequence alignments. In the SwissProt database, there are over 30 such proteins from different sources: avian (9), bovine (6) porcine transmissible gastroenteritis (4), murine (3), human (3), porcine respiratory (2), porcine epidemic diarrhea (2), rat (2), turkey (1), canine enteric (1) and feline (1). A multiple sequence alignment of ORF3a and other 17 proteins is shown. We included two viral proteins (VE6_CRPVK and ENV_SRV1) that showed a significant similarity to ORF3a and a subset of 15 Matrix proteins ranging from a broad taxonomical spectrum (Fig. 3A).

Using multiple sequence alignment based on ClustalW (Fig. 3A), a cladogram indicating the proposed evolutionary ancestral relationships among the proteins analyzed is shown (Fig. 3B). Note that the closest relationship of ORF3a among the listed proteins is to a nuclear matrix-associated protein (VE6_CRPVK) from papillomavirus and the Env polyprotein (ENV_SRV1) from simian retrovirus. Among the coronaviruses matrix proteins, the avian, feline and canine proteins show maximal similarity to ORF3a. The significance of the similarity of ORF3a to matrix E1 proteins of coronaviruses is further supported indirectly as (i) these proteins are of a similar length of ~240–260 amino acids; (ii) these proteins share three transmembrane domains with similar organization (Fig. 2), the three transmembrane domains are a hallmark for all coronaviruses matrix proteins; (iii) the number of hits in a BLAST search for ORF3a among the coronaviruses is significant considering the low abundance of these proteins in the database searched.

Very recently, two works dealing with the ORF3a characterization have been published. The first one [14] shows that the product of the ORF3a is a membrane protein that is ex-

```
VME1_CVCAI P36299 E1 glycoprotein precursor (Matrix), Length = 262

Score = 161 (54.7 bits), E-Score = 2.8e-07
Identities = 48/241 (19%), Positives = 116/241 (48%)

ORF3a:     6 RFFTLGSITAQPVKIDNASPASTVHATATIPLQ-ASLPFGWLVIGVAFLAVFQ------- 57
             R+ ++   ++ + +   A+  ++   T+ +  + A+  F+W VI + F++V+Q
VME1:     19 RYCAMTE-SSTSCRNSTAGNCASCFETGDLIWHLANWNFSWSVILIIFITVLQYGRPQFS 77

ORF3a:    58 ---SATKIALNKRWQLALYKGFQFICNLLLLF-VTIYSHLLLVAAGMEAQFLYLYALIY 113
                + K++ +   W + L  ++ ++ N  L + V+ Y  + +  AG    F+ L  ++Y
VME1:     78 WFVCGIKMLIMWLLWPIVL--ALTIF-NAYLEYRVSRYVMFGFSVAGATVTFI-LW-IMY 132

ORF3a:    114 FLQCINACRIIMRCWLCWKCKSKNPLLYDANYFVCWHTHNYDYCIPYNSVTDTIVVTEGD 173
              F++ I+  R   + W  W   S NP   +++ ++C  +    Y +P ++V   + +T
VME1:     133 FVRSIQLYRRT-KSW--W---SFNP---ETSAILCVSALGRSYVLPLEGVPTGVTLT--- 180

ORF3a:    174 GISTPKL-KEDYQI-GGYSEDRHSGVKDYVVVHGYFTEVYYQLESTQITTDTGIENATFF 231
               + + +L  E ++I GG + D    +  YV+V      + Y L + ++ + ++   A ++
VME1:     181 -LLSGNLCAEGFKIAGGMNIDN---LPKYVMVALPVRTIVYTLVGKKLKASSATGWA-YY 235

ORF3a:    232 I 232
              +
VME1:     236 V 236
```

Fig. 2. Result of a BLAST search for ORF3a and E1 glycoprotein precursor – Matrix glycoprotein from Canine enteric coronavirus (UniProt: VME1_CVCAI, 262 aa). The level of similarity is 116/241 amino acids (48%). A similar degree of similarity was detected for an M-protein from Feline CoV (TrEMBL: Q8JVQ9), Porcine respiratory CoV (strain RM4, UniProt: VME1_CVPRM), a Canine CoV isolated from liver of giant panda (EMBL: AAR11075.1) and others. The results of prediction of transmembrane domains using the TMHMM are shown in gray. For both proteins, three transmembrane domains in the amino-terminal half of the proteins were predicted with high confidence.

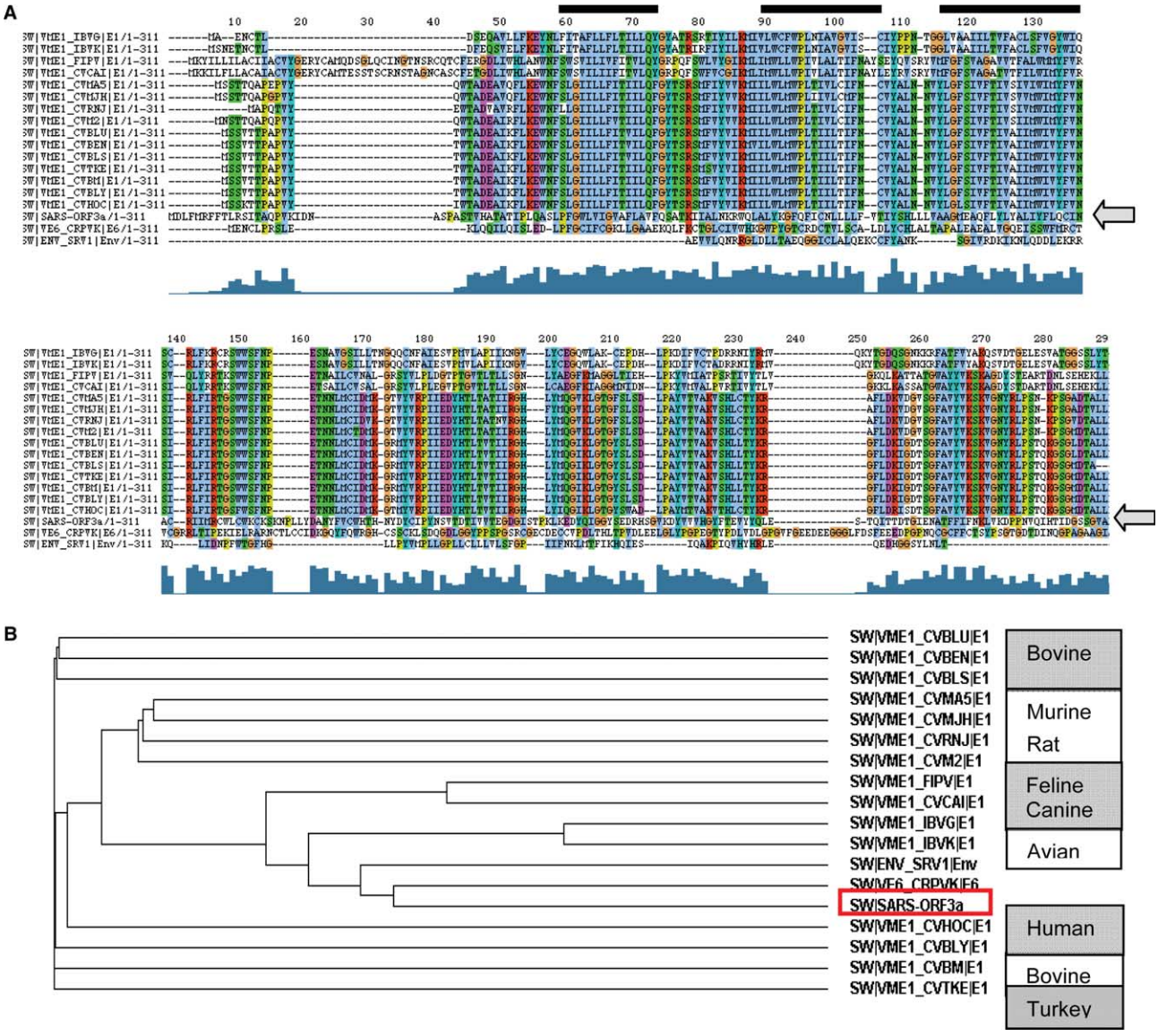

Fig. 3. (A) Multiple sequence alignment of ORF3a and 15 representatives of Matrix E1 glycoproteins from coronaviruses. Two additional viral proteins that were detected as top hits in BLAST search for ORF3a are included. Those proteins are nuclear matrix-associated (VE6_CRPVK) from papillomavirus and the Env polyprotein precursor (ENV_SRV1) from simian retrovirus. Note that the alignment of ORF3a (marked by an arrow) is most significant in the region covering the putative transmembrane domains, marked by a bold line above the alignment. (B) A cladogram indicates a common ancestry based on the alignments shown in A. The origin of the coronaviruses proteins is listed next to the protein name. ORF3a is in the red box.

pressed at high levels in SARS infected cells, transported to the cell surface and undergoes endocytosis. Another work, using proteomics methods, shows that the ORF3a protein product might have a structural role and probably interacts with the virus spike protein [15]. These reports support our assumption that ORF3a is indeed an evolutionary modified variant of one of the structural viral proteins.

A similarity search applied for the other UC-ORFs revealed relatedness to viral sequences outside of the coronaviruses, albeit with a very low statistical significance. For example, ORF6 (63 aa) is most resembled to Gene 6 protein from Spiroplasma virus (VG6_SPV1R, 113 aa), a single-stranded circular DNA virus [16] and to Vpu protein [17]. The latter is used for sub-typing lentiviruses and was detected in African immunodeficiency virus type 1 (HIV-1) (not shown). Unexpectedly, ORF9b (98 aa) shows also a similarity to lentivirus RNA viruses such as the HIV for their Gag proteins. The significance of the similarity for those very short ORFs cannot be confirmed without experimental support.

3.2. Internal homology among ORFs

For most UC-ORFs, expression from a nested set of subgenomic mRNAs was confirmed experimentally following transfection of the virus to cultured cells [6]. Whether functional proteins are produced is not yet known and should be experimentally validated. Irrespective of the level of protein expression, we observed that many UC-ORFs (i.e., ORF3b, ORF6 and ORF9b) exhibit a surprising pairwise similarity among themselves spanning a large part of their sequences

(Fig. 4A). Such internal similarity may result from duplication, fusion or shuffling events that may have occurred along the evolution of SARS-CoV. An example of such internal pairwise similarity is seen in Fig. 4A. Examples of internal pairwise similarity among some of the structural ORFs and the UC-ORFs are shown in Fig. 4B.

In an attempt to trace the evolutionary history of the short UC-ORFs, we tested the extent of the internal pairwise similarities and their position on the protein sequence. A graphic view for the overlapping alignments presented in Fig. 4B is shown (Fig. 5). It is possible that ORF3a is an authentic duplicated variant of a coronavirus Matrix glycoprotein that gave rise to short ORFs of ORF7b and ORF8a, whose similarity to ORF3a is evident. Interestingly, ORF7a most likely reflects an internal duplication with traces of similarity to ORF8a and ORF7b at both termini. The possibility that ORF7a served as an intermediate for those ORFs cannot be excluded. Some of the similarities among the discussed ORFs are also traceable at the nucleotide level (not shown).

The main features of the ORFs of SARS-CoV related to our results are summarized in Table 1. Most notable is the observation that almost all UC-ORFs are homologous to viral proteins within the coronaviruses lineage (ORF3a, ORF8a, and ORF9a) and other viral groups. We favor the idea that most UC-ORFs are reminiscent of genetic dynamics in the evolution of the virus. This is indirectly supported by inspecting the tens of SARS-CoV variants whose sequences were archived in the public database. We evaluated the average

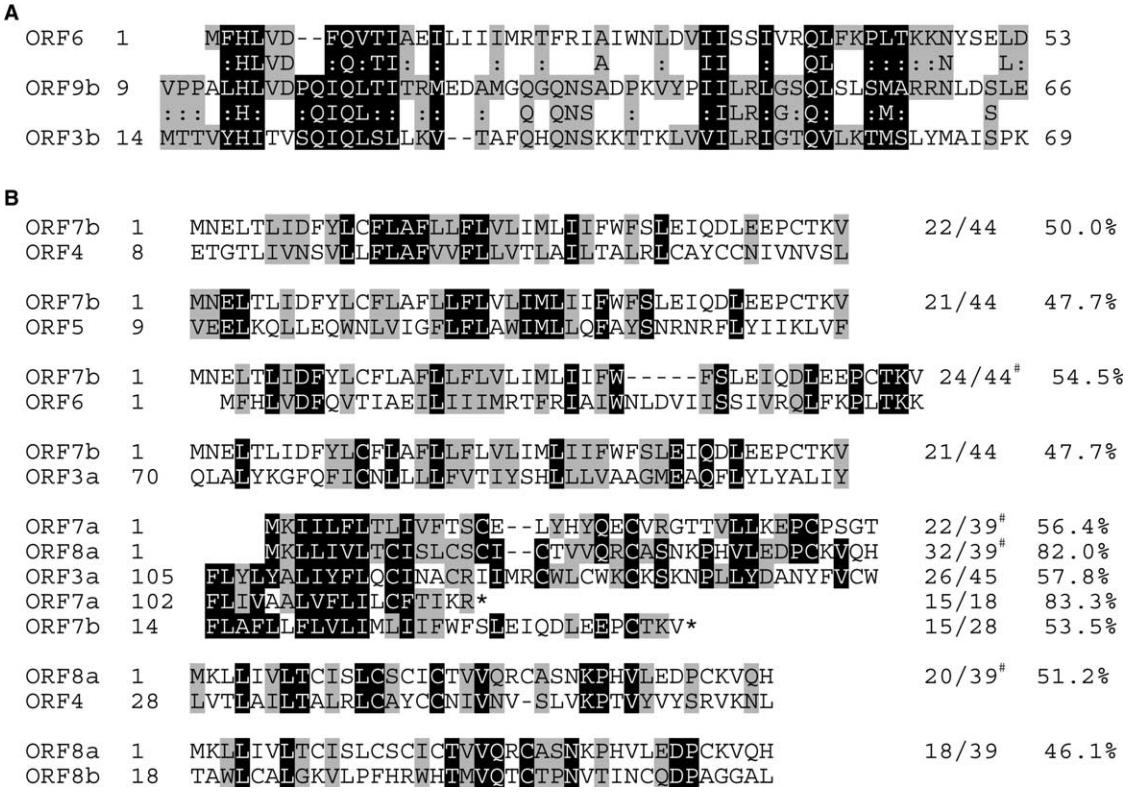

Fig. 4. Internal similarity among SARS-CoV ORFs. (A) The similarity between ORF6, ORF9b and ORF3b. Black color marks identical residues and gray color marks conserved residues. (B) The alignment between ORF7b and ORF8a to other ORFs is shown. The identical residues and similar residues are marked as above. A global alignment for the entire length of ORF7b or ORF8a (44 and 39 aa, respectively) was applied. The level of similarity is indicated. Recall that both ORF4 and ORF5 have a defined function in the structure of the virus, while the function of the other ORFs is still unknown. (–) marks a gap in the alignments; (*) marks the C-terminal of the ORF.

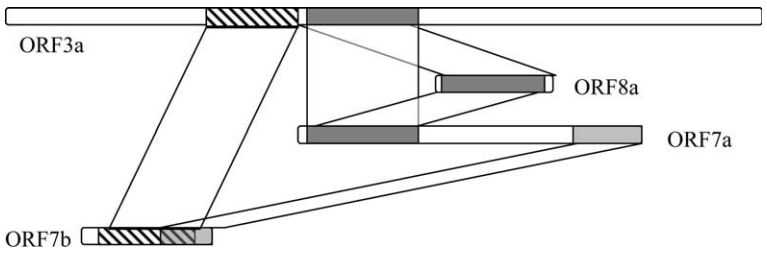

Fig. 5. Pairwise similarity among ORF3a, ORF7a, ORF7b and ORF8b. The similarity is based on alignment in Fig. 4B. Only the most significant similarities are marked. ORFs are drawn to scale.

Table 1
A summary on SARS-CoV ORFs

Similarity	Known function	Length (aa)	Variants[a] aa	TMD[b]	Related viral[c]	Pairwise ORFs[d]
ORF 1a/b	pp1ab, Polyprotein1a/b	7073	76	+(16)	+	–
ORF2	S, Spike – E2	1255	21	+(2)	+	–
ORF3A	–	274	8	+(3)	+	+
ORF3B	–	154	4	–	–	+
ORF4	E, Small envelope	76	1	+(2)	+	+
ORF5	M, Matrix protein	221	6	+(3)	+	+
ORF6	–	63	2	+(1)	++	+
ORF7A	–	122	1	+(2)	++	–
ORF7B	–	44	–	+(1)	–	+
ORF8A	–	39	–	+(1)	+	+
ORF8B	–	84	1	–	–	+
ORF9A	N, Nucleocapsid	422	5	–	+	–
ORF9B	–	98	1	–	++	+

For detailed information on ORFs with known viral function, see [3,4].
[a] Variants, indicating the number of amino acid changes that were collected from all currently known SARS-CoV variants based on UniProt annotations (www.ebi.ac.uk/uniprot).
[b] TMD – the transmembrane domain and the number of appearances (in parentheses) as predicted by TMHMM.
[c] Similarity outside the CoV groups is marked by ++.
[d] Internal similarity with at least one other SARS CoV-ORF, based on Fig. 4.

number of amino-acids that were changed in all SARS-CoV variants for the known 6 ORFs, as well as for the other 8 UC-ORFs. A reasonable assumption is that the evolutionary pressure for amino acid synonym conservation might be weaker for redundant and non-functional genomic regions. We tested whether the UN-ORFs (recall their genomic organization at the vicinity of structural genes, Fig. 1) are more prone to mutations that lead to amino acids substitutions. Indeed, we counted 109 such amino acid changes for all known ORFs (total of 9047 amino acids, Table 1) and 17 changes for all UC-ORFs (total length of 878 amino acids). The amounts of amino acid changes (normalized per amino acid) are in favor of the UC-ORFs by a ratio of 1.6:1.0, reflecting the tendency of non-synonymous mutations in UC-ORFs to accumulate. Interestingly, by taking into account only non-conserved amino acid changes (based on aa similarity groups, as in Fig. 4), the accumulation of amino acid changes is almost double for the UC-ORFs relative to all other ORFs.

4. Discussion

A proteomic approach to detect traces of the SARS-CoV proteins in blood samples of SARS infected individuals was reported [9]. However, only major proteins were detected by this method, leaving the question of UC-ORFs translation still open. Only recently, the presence of the ORF3a protein product was shown [14,15]. In coronaviruses, transcription efficiency was correlated with the presence of transcription-regulating sequences (TRS). As well as the characterized ORFs, the S, E and N proteins (Spike, Envelope and Nucleocapsid, respectively) that contain the minimal core sequence of the TRS in the vicinity of their initiation AUG, the ORF3a, ORF7a and ORF8a also contain such a sequence. None of the 'nested' ORFs (ORF3b, ORF7b, ORF8b and ORF9b) have a genuine TRS and their transcription suggests a TRS independent mode (discussed in [6]). While translation efficiency is strongly dependent on nucleotide context next to the initial AUG [18], the extent by which the transcribed mRNA of UC-ORFs are translated to properly folded proteins is not yet known.

It was proposed that the SARS genome is unique in view of the mode of divergence within the CoV group 2 [9]. Based on inspection of ORF1a,b, an event of recombination between mammalian and avian viral origin was proposed [19]. A recent survey of the mutational patterns suggests that the accumulation of mutations (and their nature) is reflected by the genomic organization of the various coronaviruses, among which the SARS-CoV showed a unique rate of mutational stabilization [20]. Our observations suggest that some of the UC-ORFs have originated from a duplication of structural proteins of the coronaviruses lineage (Figs. 2,3 and 5). Although at that stage it is very speculative, the similarity of the UC-ORFs with human viral proteins such as lentivirus, arterivirus and megalovirus (not shown and Table 1) is intriguing. It is plausible that some of the information exchange mechanisms from and to the SARS-CoV had occurred through a multi-infected host cells (probably in mammals). At present, the candidate host for SARS-CoV is a civet cat [21], though this has not been fully confirmed. Our observations provide additional support for an active mode for internal exchanging of genetic information during the evolution of the SARS-CoV (Fig. 5).

It is logical to assume that, along the evolutionary process, viruses tend to eliminate non-functional ORFs and minimize the non-essential genetic replication load. In the case of SARS-CoV, we may be witnessing an early stage of the virus evolution, in which a small part of the genome is still occupied by very short overlapping non-functional ORFs. Tens of new SARS-CoV variants that were recently sequenced possibly indicate rapid accumulation of genetic variations in UC-ORFs, much above the extent of this phenomenon in functional ORFs (replicase and structural ORFs). For example, the amino acid substitutions detected in ORF4 and ORF5 (E and M protein, respectively) are mostly conserved, while changes that occur in the coding regions of the UC-ORFs (in ORF3a, ORF6, ORF8b and ORF9b) are by large non-conserved. In one variant of the virus, a single nucleotide deletion resulted in a frame shift change in two of the overlapping UC-ORFs (not shown). In general, UC-ORFs tend to accumulate almost twice as many non-conserved amino acids compared to all other functional ORFs (Table 1). This observation is consistent with

the notion that most UC-ORFs are redundant for the infectivity and the life cycle of the virus.

In summary, we hypothesize that the most UC-ORFs are reminiscent of recent events of genetic information exchange. From this perspective, the UC-ORFs resemble the origin and dynamics of pseudogenes in high eukaryotes [22]. Ranging from yeast to human, many pseudogenes resulted from an early duplication, that followed by mutation accumulation, leading to inactivation of their coding region. In a complementary process, alteration in transcriptional and translational signals may cause the gene to become non-functional and consequently to reduce the mutational selective pressure, causing the expansion of pseudogenes.

Acknowledgements: We are grateful to Noam Kaplan and Menachem Fromer for critical reading of the manuscript. This study is supported by The Sudarsky Center for Computational Biology in the Hebrew University and the NoE BioSapience consortium grant.

References

[1] Oxford, J.S., Bossuyt, S. and Lambkin, R. (2003) Immunology 109, 326–328.
[2] Guan, Y. et al. (2003) Science 302, 276–278.
[3] Rota, P.A. et al. (2003) Science 300, 1394–1399.
[4] Marra, M.A. et al. (2003) Science 300, 1399–1404.
[5] Ruan, Y.J. et al. (2003) Lancet 361, 1779–1785.
[6] Snijder, E.J. et al. (2003) J. Mol. Biol. 331, 991–1004.
[7] Thiel, V. et al. (2003) J. Gen. Virol. 84, 2305–2315.
[8] Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. (1997) Nucleic Acids Res. 25, 3389–3402.
[9] Krokhin, O. et al. (2003) Mol. Cell Proteom. 2, 346–356.
[10] Reese, J.T. and Pearson, W.R. (2002) Bioinformatics 18, 1500–1507.
[11] Krogh, A., Larsson, B., von Heijne, G. and Sonnhammer, E.L. (2001) J. Mol. Biol. 305, 567–580.
[12] Kall, L., Krogh, A. and Sonnhammer, E.L. (2004) J. Mol. Biol. 338, 1027–1036.
[13] Aiyar, A. (2000) Methods Mol. Biol. 132, 221–241.
[14] Tan, Y.J. et al. (2004) J. Virol. 78, 6723–6734.
[15] Zeng, R. et al. (2004) J. Mol. Biol. 341, 271–279.
[16] Sha, Y., Melcher, U., Davis, R.E. and Fletcher, J. (2000) Virus Genes 20, 47–56.
[17] Scriba, T.J., Treurnicht, F.K., Zeier, M., Engelbrecht, S. and van Rensburg, E.J. (2001) AIDS Res. Hum. Retroviruses 17, 775–781.
[18] Kozak, M. (1996) Mamm. Genome 7, 563–574.
[19] Stavrinides, J. and Guttman, D.S. (2004) J. Virol. 78, 76–82.
[20] Grigoriev, A. (2004) Trends Genet. 20, 131–135.
[21] Enserink, M. (2003) Science 300, 1351.
[22] Echols, N., Harrison, P., Balasubramanian, S., Luscombe, N.M., Bertone, P., Zhang, Z. and Gerstein, M. (2002) Nucleic Acids Res. 30, 2515–2523.