Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Available online at www.sciencedirect.com
SCIENCE DIRECT®
Biochemical and Biophysical Research Communications 307 (2003) 382–388
BBRC
www.elsevier.com/locate/ybbrc

ZCURVE_CoV: a new system to recognize protein coding genes in coronavirus genomes, and its applications in analyzing SARS-CoV genomes

Ling-Ling Chen,[a,b,1] Hong-Yu Ou,[a,1] Ren Zhang,[c] and Chun-Ting Zhang[a,*]

[a] *Department of Physics, Tianjin University, Tianjin 300072, PR China*
[b] *Department of Biology, Shandong University of Technology, Zibo 255049, PR China*
[c] *Department of Epidemiology and Biostatistics, Tianjin Cancer Institute and Hospital, Tianjin 300060, PR China*

Received 23 May 2003

Abstract

A new system to recognize protein coding genes in the coronavirus genomes, specially suitable for the SARS-CoV genomes, has been proposed in this paper. Compared with some existing systems, the new program package has the merits of simplicity, high accuracy, reliability, and quickness. The system ZCURVE_CoV has been run for each of the 11 newly sequenced SARS-CoV genomes. Consequently, six genomes not annotated previously have been annotated, and some problems of previous annotations in the remaining five genomes have been pointed out and discussed. In addition to the polyprotein chain ORFs 1a and 1b and the four genes coding for the major structural proteins, spike (S), small envelop (E), membrane (M), and nuleocaspid (N), respectively, ZCURVE_CoV also predicts 5–6 putative proteins in length between 39 and 274 amino acids with unknown functions. Some single nucleotide mutations within these putative coding sequences have been detected and their biological implications are discussed. A web service is provided, by which a user can obtain the annotated result immediately by pasting the SARS-CoV genome sequences into the input window on the web site (http://tubic.tju.edu.cn/sars/). The software ZCURVE_CoV can also be downloaded freely from the web address mentioned above and run in computers under the platforms of Windows or Linux.

Keywords: Coronavirus; Severe acute respiratory syndrome; SARS-CoV; Genome; Gene-finding; Mutation

An outbreak of a life-threatening disease, referred to as severe acute respiratory syndrome (SARS), has spread to many countries around the world [1–6]. By late May 2003, the World Health Organization (WHO) has recorded more than 7000 cases of SARS and more than 600 SARS-related deaths, and therefore a global alert for the illness was issued due to the severity of the disease (http://www.who.int/csr/sars/en/).

A growing body of evidence has convincingly shown that SARS is caused by a novel coronavirus, called SARS-coronavirus or SARS-CoV. Currently, the complete genome sequences of 11 strains of SARS-CoV isolated from some SARS patients have been sequenced [7–9], and more complete genome sequences of SARS-CoV are expected to come.

The SARS-CoV genomes are about 30 kb in length. For such short genome sequences, currently, there is no reliable software for the identification of protein-coding genes. Therefore, most sequenced genomes were annotated manually or not annotated. Among the 11 completed sequences, six were not annotated yet and the remaining were annotated manually.

Currently, most algorithms for gene identification in prokaryotic genomes, such as GeneMark.hmm [10] and Glimmer [11], are based either on the higher-order Markov chain model or the hidden Markov chain model in which thousands of parameters need to be trained. The large number of parameters may result in less adaptability, especially for small genomes. Meanwhile, ZCURVE [12] is a newly developed system for gene

* Corresponding author. Fax: +86-22-2740-2697.
E-mail address: ctzhang@tju.edu.cn (C.-T. Zhang).
[1] These authors contributed equally to this work.

0006-291X/03/$ - see front matter © 2003 Elsevier Inc. All rights reserved.
doi:10.1016/S0006-291X(03)01192-6

recognition in bacterial and archaeal genomes, in which only 33 parameters are used and the recognition accuracy is high. Therefore, the ZCURVE algorithm essentializes the coding properties of protein-coding genes with relatively small number of parameters. Thus, it is not only suitable for large but also especially suitable for small genomes.

In this paper, we describe a system, called ZCURVE_CoV, based on a coronavirus-specific ZCURVE algorithm, which is especially suitable for gene recognition in SARS-CoV genomes. The system has the advantages of simplicity, reliability, high accuracy, and quickness. The software system ZCURVE_CoV is freely available at http://tubic.tju.edu.cn/sars/.

Materials and methods

Six genome sequences of coronaviruses and the annotation information were downloaded from the web site of NCBI RefSeq project (http://www.ncbi.nih.gov/RefSeq/). These coronaviruses include avian infectious bronchitis virus (NC_001451), bovine coronavirus (NC_003045), human coronavirus 229E (NC_002645), murine hepatitis virus (NC_001846), porcine epidemic diarrhea virus (NC_003436), and transmissible gastroenteritis virus (NC_002306). A total of 48 genes were extracted from the above six genomes and used to train the gene-finding algorithm. Currently, 15 genome sequences of SARS coronavirus (SARS-CoV) strains are available in the GenBank database, of which there are 11 complete and four partial genomes, respectively. The former includes SARS-CoV TOR2 (Accession No. AY274119), Urbani (AY278741), HKU-39849 (AY278491), CUHK-W1 (AY278554), BJ01 (AY278488), CUHK-Su10 (AY282752), SIN2500 (AY283794), SIN2748 (AY283797), SIN2679 (AY283796), SIN2774 (AY283798), and SIN2677 (AY283795), whereas the latter includes SARS-CoV BJ02 (AY278487), BJ03 (AY278490), BJ04 (AY279354), and GZ01 (AY278489), respectively.

The gene-finding algorithm presented in this paper is based on the Z curve [13], which is a graphic representation of DNA sequences. The Z curve method has been used to recognize protein coding genes in the budding yeast genome [14]. A new ab initio gene-finding system for bacterial and archaeal genomes has been developed recently, based on the Z curve method [12]. Here the method with some modifications is used to recognize protein coding genes in coronavirus genomes, which is presented briefly as follows. Suppose that the occurrence frequencies of the bases A, C, G, and T (U) at the first, second, and third codon positions in an ORF are denoted by a_i, c_i, g_i, and t_i, respectively, where $i = 1, 2, 3$. The four numbers, a_i, c_i, g_i, and t_i, are mapped onto a point in a 3-dimensional space V_i with the coordinates

$$\begin{cases} x_i = (a_i + g_i) - (c_i + t_i), \\ y_i = (a_i + c_i) - (g_i + t_i), \quad i = 1, 2, 3, \\ z_i = (a_i + t_i) - (g_i + c_i). \end{cases} \tag{1}$$

Then, each ORF may be represented by a point or a vector in a 9-dimensional space V, where $\mathrm{V} = \mathrm{V}_1 \oplus \mathrm{V}_2 \oplus \mathrm{V}_3$, where the symbol $\oplus$ denotes the direct-sum of two subspaces. The nine components u_1–u_9 of the space V are defined as follows:

$$\begin{cases} u_1 = x_1, \quad u_2 = y_1, \quad u_3 = z_1, \\ u_4 = x_2, \quad u_5 = y_2, \quad u_6 = z_2, \\ u_7 = x_3, \quad u_8 = y_3, \quad u_9 = z_3. \end{cases} \tag{2}$$

To train the system, two sets of samples are needed, which are positive samples corresponding to protein-coding genes (seed ORFs) and negative samples corresponding to non-coding sequences. In the Z curve method, essentially, the gene recognition is based on the compositional asymmetry of three codon positions in coding sequences. It was shown that the overall extent of codon usage bias in RNA viruses is low and there is little variation in bias between genes [15]. Coronaviruses belong to the coronaviridae and the G + C content of the published coronavirus genomes ranges from 37% to 42% [7]. Therefore, it is reasonable to deduce that the published coronavirus genomes have similar codon usage. Based on this consideration, it is possible that gene-finding parameters derived from some published coronavirus genomes may be applied to recognize genes in other coronavirus genomes. Because the SARS-CoV genomes are relatively small (≈30 kb), it is difficult to obtain enough seed ORFs from its own genome. Therefore, we used some other published coronavirus genomes to train gene-finding parameters. Consequently, the genomes of avian infectious bronchitis virus, bovine coronavirus, human coronavirus 229E, murine hepatitis virus, porcine epidemic diarrhea virus, and transmissible gastroenteritis virus, respectively, were used, in which 48 seed ORFs were selected. The detailed information about the 48 seed ORFs is described in Table 1 of the supplementary materials (see: http://tubic.tju.edu.cn/sars/).

Below we describe the strategy to produce the negative samples. It is a rather difficult problem to produce an appropriate set of non-coding sequences in coronavirus genomes, because the amount of non-coding DNA sequences in these genomes is too few to be used. A method to produce negative samples has been developed previously and it has been shown to be an effective way to solve the problem [12]. The same method is still used in the current study. In this method, a negative sample is just derived from a seed ORF. Generally speaking, if the regular structure of a coding sequence is completely destroyed, it is transformed into a non-coding one. Therefore, the negative sample may be simply obtained by shuffling the corresponding coding sequence sufficiently (20,000 times in current study). The resulting random sequences from all 48 seed ORFs were used as non-coding sequences. The major difference is that the former has some regular structures, whereas the latter is a random sequence. In fact, a random sequence is not a non-coding sequence, but it is a good approximation. As shown below, this approximation generally results in good gene-finding results.

The Fisher linear equation for discriminating the positive and negative samples in the 9-dimensional space V represents a superplane, described by a vector $\mathbf{c}$ which has nine components $c_1, c_2, \ldots$, and c_9. For more details about Fisher discrminant algorithm, refer to, for example [14]. Based on the data in the training set (including the positive and negative samples), the vector $\mathbf{c}$ and the threshold c_0 are obtained. The decision of coding/non-coding for each ORF and negative sample is simply made by the criterion of $\mathbf{c} \cdot \mathbf{u} > c_0 / \mathbf{c} \cdot \mathbf{u} < c_0$, where $\mathbf{c} = (c_1, c_2, \ldots, c_9)^{\mathrm{T}}$, $\mathbf{u} = (u_1, u_2, \ldots, u_9)^{\mathrm{T}}$, and "T" indicates the transpose of a matrix. The criterion of $\mathbf{c} \cdot \mathbf{u} > c_0 / \mathbf{c} \cdot \mathbf{u} < c_0$ for making the decision of coding/non-coding can be rewritten as $Z(\mathbf{u}) > 0 / Z(\mathbf{u}) < 0$, where $Z(\mathbf{u}) = \mathbf{c} \cdot \mathbf{u} - c_0$. $Z(\mathbf{u})$ is called the Z score or Z index for an ORF or a fragment of DNA sequence. Finally, the strategy to deal with overlapping ORFs used here is similar to that described in the previous paper [12].

Results and discussions

Comparison with the existing system—GeneMark.hmm

No coronavirus-specific annotation systems have been available so far. Currently, GeneMark.hmm is commonly used for gene-finding in virus genomes [10]. We submitted the SARS-CoV TOR2 genome to GeneMark.hmm website using default settings and the prediction result is listed in Table 1. It can be seen that the

Table 1
The genes predicted by GeneMark.hmm for the SARS-CoV, TOR2 strain

Gene	Start	Stop	Gene length (bp)
1	<3	53	51
2	265	13,413	13,149
3	13,599	21,485	7887
4	21,492	25,259	3768
5	25,268	26,092	825
6	26,398	27,063	666
7	27,074	27,265	192
8	27,273	27,641	369
9*	27,864	28,118	255
10*	28,130	28,426	297
11*	28,423	29,388	966

* The same genome was submitted several times to the website, however, the prediction results were not identical at all times, indicating that the system is unstable. An important structural protein gene (N protein), which is located from 28120 to 29388, was predicted as 'gene 10' and 'gene 11' in some predicted results. Sometimes, 'gene 9,' a quite conserved ORF in all of the 11 SARS-CoV genomes mentioned above, was not predicted. In addition, the gene coding for a structural protein (small envelope protein E) was also missed by the prediction. For more details, see the text.

predicted 'gene 1' is questionable, because of its short length and the lack of a start codon. An important structural protein gene (small envelope protein E), which is located from 26117 to 26347, was not predicted by GeneMark.hmm. Moreover, we submitted the same genome sequence several times to the website, however, the prediction results were not identical at all times, indicating that the system is unstable. An important structural protein gene (N protein), which is located from 28120 to 29388, was predicted as 'gene 10' and 'gene 11' (marked with * in Table 1) in some predicted results. Sometimes, 'gene 9' (marked with * in Table 1), a quite conserved ORF in all of the 11 SARS-CoV genomes mentioned above, was not predicted. Compared with GeneMark.hmm for gene-finding in the SARS-CoV genomes, the performance of ZCURVE_CoV is better (see Table 3 in the supplementary materials).

Apply ZCURVE_CoV to analyze the SARS-CoV genomes

Currently, the genome sequences of 15 SARS-CoV strains are available in GenBank/EMBL databases, of which there are 11 complete and four partially complete genomes. The gene-finding software ZCURVE_CoV Version 1.0 has been run for each of the 11 complete SARS-CoV genomes. To save space, the detailed results are listed in Table 3 of the supplementary materials (see also the discussion below). In addition to the polyprotein chain ORFs 1a and 1b, the program predicts four structural genes coding for the four major structural proteins, i.e., spike (S), small envelop (E), membrane (M), and nuleocaspid (N), respectively, in all the 11 SARS-CoV genomes. Additionally, ZCURVE_CoV 1.0 also predicts 5–6 putative proteins with lengths between 39 and 274 amino acids for the 11 genomes. These putative genes might code for non-structural proteins in the SARS-CoV genomes.

To compare the gene-finding result of the system ZCURVE_CoV 1.0 with that of known annotation, the SARS-CoV TOR2 strain is used as an example. The genome of TOR2 strain was annotated manually [8] and the annotated result is listed on the left part of Table 2, whereas the annotated result of ZCURVE_CoV 1.0 is listed on the right part of Table 2. As we can see both annotations are in good agreement with each other, except three ORFs. The three ORFs, i.e., ORF4, ORF13, and ORF14 annotated by Marra et al. [8] are not predicted by ZCURVE_CoV 1.0. These ORFs are completely embedded, with a frameshift, within the genes coding for some structural proteins. The absence of the transcription regulating sequences (TRSs) at the 5′ end of these ORFs [8] suggests that they are unlikely to be the protein-coding genes. The principal component analysis performed below further confirms the above conjecture. As mentioned in the Materials and methods section, each ORF is represented by a point in a 9-dimensional (9-D) space. Consequently, the positive samples (genes) and negative samples (non-coding sequences) are represented by two groups of points in the 9-D space, respectively. For the TOR2 strain, the 12 putative genes predicted by ZCURVE_CoV and ORF 4, ORF 13, and ORF 14 are represented by the corresponding points in the 9-D space, respectively. We project the points in the 9-D space onto the 3-D space spanned by the first, second, and third principal axes based on the principal component analysis. The fraction of the first three principal components accounts for about 70% of the total inertia of the 9-D space. Fig. 1 shows the distribution of the corresponding points in the 3-D space, where green and orange balls represent the positive samples (genes) and negative samples (non-coding sequences), respectively. Blue balls correspond to the genes predicted by ZCURVE_CoV for the TOR2 strain, while red balls correspond to ORF 4, ORF 13, and ORF 14 annotated by Marra et al. [8]. It is clear that the three red balls are located at the side of non-coding sequences, indicating that ORF 4, ORF 13, and ORF 14 are very unlikely to code for proteins.

Similar analysis was performed to the Urbani strain [7]. The result is listed in Table 3, in which the putative gene X2 annotated by Rota et al. [7], corresponding to ORF 4 in Marra et al. [8], is not predicted by ZCURVE_CoV. Based on the above analysis, X2 is also very unlikely to code for a protein. Of the 11 complete SARS-CoV genomes, six have not yet been annotated. We have run the program ZCURVE_CoV for each of the 11 genomes. Consequently, those already annotated have been re-annotated and those not annotated yet

Table 2
Comparison of the genes annotated and those predicted by ZCURVE_CoV 1.0, for the SARS-CoV, TOR2 strain

Genes annotated					Genes predicted by ZCURVE_CoV 1.0				
Start	Stop	bp	a.a.	Feature	Start	Stop	bp	a.a.	Feature
265	13,398	13,134	4377	ORF 1a	265	13,398[a]	13,134	4377	ORF 1a
13,398	21,485	8088	2695	ORF 1b	13,398[a]	21,485	8088	2695	ORF 1b
21,492	25,259	3768	1255	S protein	21,492	25,259	3768	1255	S protein
25,268	26,092	825	274	ORF 3	25,268	26,092	825	274	Sars274
25,689	26,153	465	154	ORF 4					
26,117	26,347	231	76	E protein	26,117	26,347	231	76	E protein
26,398	27,063	666	221	M protein	26,398	27,063	666	221	M protein
27,074	27,265	192	63	ORF 7	27,074	27,265	192	63	Sars63
27,273	27,641	369	122	ORF 8	27,273	27,641	369	122	Sars122
27,638	27,772	135	44	ORF 9	27,638	27,772	135	44	Sars44
27,779	27,898	120	39	ORF 10	27,779	27,898	120	39	Sars39
27,864	28,118	255	84	ORF 11	27,864	28118	255	84	Sars84
28,120	29,388	1269	422	N protein	28,120	29,388	1269	422	N protein
28,130	28,426	297	98	ORF 13					
28,583	28,795	213	70	ORF 14					

[a] The program ZCURVE_CoV 1.0 has two options. The default option is to use the heptamer UUUAAAC as the conservative 'slippery sequence' to find the coronavirus −1 frameshift site [16]. Once the heptamer is found in the upstream sequence near the ending site of ORF 1a originally predicted, the ending site of ORF 1a and starting site of ORF 1b are both corrected to the frameshift site (13398 in this genome) according to this 'slippery sequence.' Otherwise, if this heptamer cannot be found, only the original sites predicted for ORF 1a and ORF 1b are displayed in the output file. The second option is to ignore the −1 frameshift, and the original sites predicted for ORF 1a and ORF 1b are always displayed, regardless of the existence of the heptamer UUUAAAC.

Table 3
Comparison of the genes annotated and those predicted by ZCURVE_CoV1.0, for the SARS-CoV, Urbani strain

Genes annotated					Genes predicted by ZCURVE_CoV 1.0				
Start	Stop	bp	a.a.	Feature	Start	Stop	bp	a.a.	Feature
265	13,398	13,134	4377	ORF 1a	265	13,398[a]	13,134	4377	ORF 1a
13,398	21,485	8088	2695	ORF 1b	13,398[a]	21,485	8088	2695	ORF 1b
21,492	25,259	3768	1255	S protein	21,492	25,259	3768	1255	S protein
25,268	26,092	825	274	X1	25,268	26,092	825	274	Sars274
25,689	26,153	465	154	X2					
26,117	26,347	231	76	E protein	26,117	26,347	231	76	E protein
26,398	27,063	666	221	M protein	26,398	27,063	666	221	M protein
27,074	27,265	192	63	X3	27,074	27,265	192	63	Sars63
27,273	27,641	369	122	X4	27,273	27,641	369	122	Sars122
					27,638	27,772	135	44	Sars44
					27,779	27,898	120	39	Sars39
27,864	28,118	255	84	X5	27,864	28,118	255	84	Sars84
28,120	29,388	1269	422	N protein	28,120	29,388	1269	422	N protein

[a] See the footnote in Table 2.

have been annotated. All of the annotated results are listed in Table 3 of the supplementary materials.

Analyze the mutations of the six putative non-structural genes by sequence alignment

To test the nucleotide mutations of the predicted genes coding for non-structural proteins, we aligned the coding sequences of Sars274, Sars63, Sars122, Sars44, Sars39, and Sars84, respectively, for the 11 complete SARS-CoV genomes using ClustalW 1.8 [17]. The results of multiple sequence alignment for the above six predicted genes coding for non-structural proteins are listed in Fig. 1 of the supplementary materials. For the three ORFs, Sars122, Sars44, and Sars84, the nucleotide sequences are all conserved in the 11 SARS-CoV genomes, indicating that the three ORFs might have crucial biological functions. Mutations in these gene sequences would result in loss of important functions. Therefore, these coding sequences might serve as the candidate targets for designing drugs against SARS. On the contrary, Sars39 is not found in the strains SIN2677 and SIN2748, and a nucleotide mutation occurs at nucleotide position 49, leading to the mutation of Cys → Arg in the strains BJ01 and CUHK-W1. The rapid mutations occurring in Sars39 imply that it is probably not a key

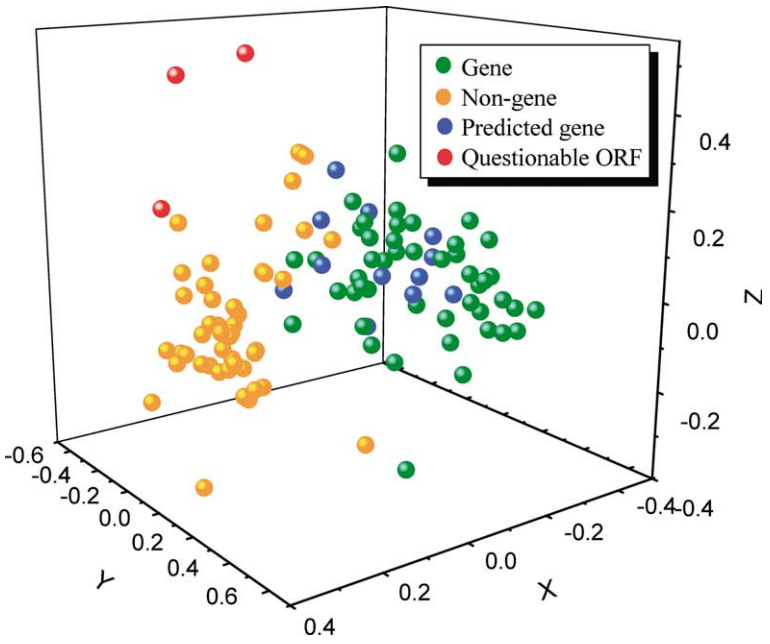

Fig. 1. Distribution of the mapping points corresponding to genes, non-genes, predicted genes, and questionable ORFs for the SARS-CoV, TOR2 strain in a 3-dimensional (3-D) space. Each gene or ORF is mapped onto a point in a 9-D space. To visualize the distribution, the mapping points are projected onto the 3-D space spanned by the first three principal axes based on the principal component analysis. The first, second, and third principal vectors are denoted by the *X*-, *Y*-, and *Z*-axes, respectively. The fraction of the first three principal components accounts for 69.59% of the total inertia of the 9-D space. Green and orange balls represent the positive samples (genes) and negative samples (non-coding sequences), respectively. Blue balls correspond to the genes predicted by ZCURVE_CoV for the TOR2 strain, while red balls correspond to ORF 4, ORF 13, and ORF 14 annotated by Marra et al. [8]. It is clear that the three red balls are situated at the side of non-coding sequences, indicating that ORF 4, ORF 13, and ORF 14 are very unlikely to code for proteins.

protein for SARS-CoV. For Sars63, two nucleotide mutations are observed at the base positions 38 and 170, leading to amino acid mutations of Glu → Gly and Pro → Leu in the strains SIN2677 and BJ01, respectively. See Fig. 1 in the supplementary materials for the detail.

The result of ClustalW alignment for Sars274 is shown in Fig. 2. Four nucleotide mutations, located at 31, 302, 406, and 783, respectively, at three different strains have been detected. The first three variations cause amino acid mutations (Fig. 2). The last substitution is a synonymous codon mutation which does not lead to amino acid change. The point mutations occurring at nucleotide positions 31, 302, and 406, respectively, cause amino acid changes. At the 31st position, G → A (TOR2) ⇒ Gly → Arg. Similarly, at the 302nd position, T (U) → A (HKU-39849) ⇒ Met → Lys; and at

```
            1                  31              302             406             783             825
TOR2        ATGGATTTGT.....CTCTTAGATCA.....AGGTATGGAGG.....AATCCAAGAACC.....GATCCAATTTA.....GCCTTTGTAA
HKU-39849   ATGGATTTGT.....CTCTTGGATCA.....AGGTAAGGAGG.....AATCCAAGAACC.....GATCCAATTTA.....GCCTTTGTAA
Urbani      ATGGATTTGT.....CTCTTGGATCA.....AGGTATGGAGG.....AATCCAAGAACC.....GATCCAATTTA.....GCCTTTGTAA
SIN2748     ATGGATTTGT.....CTCTTGGATCA.....AGGTATGGAGG.....AATCCAAGAACC.....GATCCAATTTA.....GCCTTTGTAA
SIN2774     ATGGATTTGT.....CTCTTGGATCA.....AGGTATGGAGG.....AATCCAAGAACC.....GATCCAATTTA.....GCCTTTGTAA
SIN2679     ATGGATTTGT.....CTCTTGGATCA.....AGGTATGGAGG.....AATCCAAGAACC.....GATCCAATTTA.....GCCTTTGTAA
SIN2677     ATGGATTTGT.....CTCTTGGATCA.....AGGTATGGAGG.....AATCCAAGAACC.....GATCCAATTTA.....GCCTTTGTAA
SIN2500     ATGGATTTGT.....CTCTTGGATCA.....AGGTATGGAGG.....AATCCAAGAACC.....GATCCAATTTA.....GCCTTTGTAA
CUHK-Su10   ATGGATTTGT.....CTCTTGGATCA.....AGGTATGGAGG.....AATCCAAGAACC.....GATCCAATTTA.....GCCTTTGTAA
CUHK-W1     ATGGATTTGT.....CTCTTGGATCA.....AGGTATGGAGG.....AATCCAAGAACC.....GATCCAATTTA.....GCCTTTGTAA
BJ01        ATGGATTTGT.....CTCTTGGATCA.....AGGTATGGAGG.....AATCCCAGAACC.....GATCCCATTTA.....GCCTTTGTAA
```

Mutations:

TOR2	Codon	(31)	GGA → AGA	Amino acid	(11)	Gly → Arg
HKU-39849	Codon	(302)	ATG → AAG	Amino acid	(101)	Met → Lys
BJ01	Codon	(406)	AAG → CAG	Amino acid	(135)	Lys → Gln

Fig. 2. Nucleotide mutations of the predicted gene Sars274 based on the alignment of corresponding coding sequences in 11 complete genome sequences. A total of four point mutations are detected, of which one is a silent mutation and the other three cause amino acid changes in the putative genes. The point mutations occur at nucleotide positions 31, 302, 406, and 783, respectively. At the 31st position, G → A (TOR2) ⇒ Gly → Arg. Similarly, at the 302nd position, T (U) → A (HKU-39849) ⇒ Met → Lys; at the 406th position, A → C (BJ01) ⇒ Lys→ Gln; and at the 783rd position, A → C (BJ01), but no amino acid change.

the 406th position, A → C (BJ01) ⇒ Lys → Gln. On the other hand, it was reported by Marra et al. [8] that there exist three *trans*-membrane regions spanning approximately at nucleotide positions 102 → 168 (residues 34 → 56), 231 → 297 (77 → 99), and 309 → 375 (103 → 115), respectively, in Sars274 sequence. Therefore, the mutations occur outside of the predicted *trans*-membrane regions. Note that the second mutation of amino acid (Met → Lys) is essential, as reflected by the fact that Met is a relatively strong hydrophilic amino acid, whereas Lys is a strong hydrophobic one. At present, we cannot know whether these mutations cause severe conformational changes in the tertiary structure of this putative protein. The high mutation rate of Sars274 implies that either it might be a relatively unimportant protein for SARS-CoV, or the mutations do not lead to biological function changes dramatically. Finally, for the time being we still cannot rule out the possibility that all or a part of these mutations are caused by sequencing errors.

Supplementary materials

The detailed supplementary materials related to this study are available from the website http://tubic.tju.edu.cn/sars/, which includes the following content:

(a) Table 1. The 48 seed ORFs and the six coronavirus genomes from which the seed ORFs are derived.

(b) Table 2. The Fisher coefficients and threshold obtained from the seed ORFs.

(c) Table 3. Results of gene-finding using ZCURVE_CoV for the 11 SARS-CoV complete genomes.

(d) Fig. 1. The results of multiple sequence alignment of the six predicted genes coding for non-structural proteins, Sars274, Sars63, Sars122, Sars44, Sars39, and Sars84, respectively.

Online service and availability of the program ZCURVE_CoV

A web interface of the ZCURVE_CoV system has been constructed. When a user pastes a SARS-CoV genome sequence to the input window of the website, the gene-finding result will be returned to the user immediately. A user may also download the executable version of the program ZCURVE_CoV and run it on the computers under the platforms of either Windows (95/98/NT/Me/2000 or higher), or Linux (Redhat 7.1 or higher), or SGI IRIX 6.5. For more detailed information, visit: http://tubic.tju.edu.cn/sars/.

Conclusion

Severe acute respiratory syndrome (SARS) is an extremely severe disease that has spread to many countries around the world. Accumulating evidence has shown that SARS is caused by a new coronavirus, i.e., SARS-CoV. A new system to recognize protein-coding genes in SARS-CoV genomes, called ZCURVE_CoV, has been reported in this paper. By applying the program to 11 complete SARS-CoV genomes, six genomes not annotated previously have been annotated, and some problems of previous annotations in the remaining five genomes have been pointed out and discussed. It is shown that the three protein-coding ORFs annotated by Marra et al. [8], i.e., ORF 4, ORF 13, and ORF 14, are very unlikely to code for proteins. In addition to ORF1a, ORF1b, and the four genes coding for the major structural proteins S, E, M, and N, the new system ZCURVE_CoV also predicts 5–6 putative genes coding for non-structural proteins. Aligning each of the non-structural gene sequences based on the 11 complete genomes, some mutations have been detected. The biological implications of the mutations have been discussed.

Acknowledgments

We are grateful to the scientists all over the world, who discovered and isolated the SARS coronavirus and sequenced the SARS-CoV genomes. We are indebted to Prof. Jingchu Luo in Peking University for the timely updated SARS-related information provided. The authors also thank Prof. Xi-Tai Huang and Prof. He-Mu Wang in Nankai University for their help in this work. The present study was supported in part by the 973 Project of China (Grant 1999075606).

References

[1] J.S. Peiris et al., Coronavirus as a possible cause of severe acute respiratory syndrome, Lancet 361 (2003) 1319–1325.
[2] T.G. Ksiazek et al., A novel coronavirus associated with severe acute respiratory syndrome, N. Engl. J. Med. 348 (2003) 1953–1966.
[3] C. Drosten et al., Identification of a novel coronavirus in patients with severe acute respiratory syndrome, N. Engl. J. Med. 348 (2003) 1967–1976.
[4] K.W. Tsang et al., A cluster of cases of severe acute respiratory syndrome in Hong Kong, N. Engl. J. Med. 348 (2003) 1977–1985.
[5] N. Lee et al., A major outbreak of severe acute respiratory syndrome in Hong Kong, N. Engl. J. Med. 348 (2003) 1986–1994.
[6] S.M. Poutanen et al., Identification of severe acute respiratory syndrome in Canada, N. Engl. J. Med. 348 (2003) 1995–2005.
[7] P.A. Rota et al., Characterization of a novel coronavirus associated with severe acute respiratory syndrome, Science 300 (2000) 1394–1398.
[8] M.A. Marra et al., The genome sequence of the SARS-associated coronavirus, Science 300 (2003) 1399–1404.
[9] E'd. Qin et al., A complete sequence and comparative analysis of strain (BJ01) of the SARS-associated virus, Chinese Sci. Bull. 48 (2003) 941–948.
[10] J. Besemer, M. Borodovsky, Heuristic approach to deriving models for gene finding, Nucleic Acids Res. 27 (1999) 3911–3920.
[11] S.L. Salzberg, A.L. Delcher, S. Kasif, O. White, Microbial gene identification using interpolated Markov models, Nucleic Acids Res. 26 (1998) 544–548.

[12] F.B. Guo, H.Y. Ou, C.-T. Zhang, ZCURVE: a new system for recognizing protein coding genes in bacterial and archaeal genomes, Nucleic Acids Res. 31 (2003) 1780–1789.

[13] C.-T. Zhang, R. Zhang, Analysis of distribution of bases in the coding sequences by a diagrammatic technique, Nucleic Acids Res. 19 (1991) 6313–6317.

[14] C.-T. Zhang, J. Wang, Recognition of protein coding genes in the yeast genome at better than 95% accuracy based on the *Z* curve, Nucleic Acids Res. 28 (2000) 2804–2814.

[15] G.M. Jenkins, E.C. Holmes, The extent of codon usage bias in human RNA viruses and its evolutionary origin, Virus Res. 92 (2003) 1–7.

[16] J. Ziebuhr, E.J. Snijder, A.E. Gorbalenya, Virus-encoded proteinases and proteolytic processing in the *Nidovirales*, J. Gen. Virol. 81 (2000) 853–879.

[17] J.D. Thompson, D.G. Higgins, T.J. Gibson, 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 (1994) 4673–4680.