CODEN IEJMAT *Internet Electronic Journal of Molecular Design* **2007**, *6*, 1–12 ISSN 1538–6414
BioChem Press http://www.biochempress.com

Inter*net* Electronic Journal of Molecular Design

January 2007, Volume 6, Number 1, Pages 1–12

Editor: Ovidiu Ivanciuc

Special issue dedicated to Professor Lemont B. Kier on the occasion of the 75th birthday

Similarity Analysis of DNA Sequences based on the LZ Complexity

Jia Wen [1] and Chun Li [1]

[1] Department of Mathematics, Bohai University, Jinzhou 121000, P. R. China

Received: May 18, 2006; Revised: October 19, 2006; Accepted: December 7, 2006; Published: January 31, 2007

Citation of the article:
J. Wen and C. Li, Similarity Analysis of DNA Sequences based on the LZ Complexity, *Internet Electron. J. Mol. Des.* **2007**, *6*, 1–12, http://www.biochempress.com.

Inter*net* **Electronic** Journal of Molecular Design

BioChem Press
http://www.biochempress.com

Similarity Analysis of DNA Sequences based on the LZ Complexity

Jia Wen [1,*] and Chun Li [1]

[1] Department of Mathematics, Bohai University, Jinzhou 121000, P. R. China

Received: May 18, 2006; Revised: October 19, 2006; Accepted: December 7, 2006; Published: January 31, 2007

Internet Electron. J. Mol. Des. **2007**, *6* (1), 1–12

Abstract

Motivation. Almost all methods for similarity analysis and phylogenetic inference are usually based on the multiple alignment of sequences or the invariants of sequences. But the former is not useful to all types of data, e.g. the whole genome comparisons, while the latter is accompanied by the complex calculation. The motivation of this paper is to introduce a new approach for similarity analysis of DNA sequences.

Method. We propose a relative distance measure of (0,1)–sequence based on the LZ complexity to quantify the similarity degree of two different binary sequences. By transforming a DNA sequence into three binary sequences in term of classifications of nucleic acid bases, we can obtain the relative distance of corresponding characteristic sequences of any two DNA sequences. The distance matrices are thus obtained to reflect the similarities of DNA sequences. A similarity comparison is made for the 24 complete coronavirus genomes to show the utility of our method.

Results. As a result, we find that the 24 complete coronavirus genomes can be classified into four groups on the whole. In particular, SARS–CoVs are not closely related to any of the previously characterized coronaviruses and form a distinct group within the genus coronaviruses. The result is consistent with those of previous analyses.

Conclusions. On the basis of the findings, we conclude that the present method has apparently captured important features of DNA sequences considered and is useful for similarity analysis of DNA sequences.

Keywords. DNA; coronavirus; LZ complexity; relative distance; characteristic sequences.

1 INTRODUCTION

Compilation of DNA primary sequence data continues unabated and tends to overwhelm us with voluminous outputs that increase daily. Comparison of different DNA primary sequences remains one of the most important aspects for the analysis of DNA data banks. The traditional algorithms for similarity analysis and phylogenetic inference are guaranteed to find the 'optimal' alignment and are based mostly on dynamic programming [1–5]. Such approaches have been hitherto widely used. However, the computational complexity and the inherent ambiguity of the alignment cost

Dedicated to Professor Lemont B. Kier on the occasion of the 75th birthday.

* Correspondence author; E–mail: wenjia198021@163.com.

criteria are still the bottleneck problems. Recently, Randić *et al*. [6–13] proposed a sequence comparison approach that is grounded on characterization of biological sequences by ordered sets of invariants, rather than by a direct comparison of the sequences themselves. An important advantage of the characterization of structures by invariants, as opposed to use of codes, is the simplicity of the comparison based on invariants. However, as pointed out in [14,15], this approach involves a number of as yet unresolved questions. In particular, questions that need our attention are as follows: (1) some loss of information in the transfer of data from a biological sequence to its mathematical representation; (2) how to obtain suitable invariants to characterize biological sequences and how to select invariants suitable for sequence comparisons; (3) the calculations of some effective invariants become more and more difficult with the length of the sequence longer.

In this paper, we introduce a new distance measure for the similarity analysis of DNA sequences that is based on the symbolic sequence complexity. Unlike most existing methods, the proposed method does not require sequence alignment and avoids the complex calculation as in the calculation of invariants of the long DNA sequences. The utility of our method is illustrated by an examination of similarities among the 24 complete coronavirus genomes.

2 MATERIALS AND METHODS

Table 1. The accession number, abbreviation, name and length for each of the 24 coronavirus genomes

No	Accession	Group	Abbreviation	Genome	Length(nt)
1	NC_002645	I	HCoV–229E	Human coronavirus 229E	27,317
2	NC_ 002306	I	TGEV	Transmissible gastroenteritis virus	28,586
3	NC_003436	I	PEDV	Porcine epidemic diarrhea virus	28,033
4	U00735	II	BCoVM	Bovine coronavirus strain Mebus	31,032
5	AF391542	II	BCoVL	Bovine coronavirus isolate BCoV–LUN	31,028
6	AF220295	II	BCoVQ	Bovine coronavirus strain Quebec	31,100
7	NC_003045	II	BCoV	Bovine coronavirus	31,028
8	AF208067	II	MHVM	Murine hepatitis virus strain ML–10	31,100
9	AF201929	II	MHV2	Murine hepatitis virus stain 2	31,028
10	AF208066	II	MHVP	Murine hepatitis virus strain Penn 97–1	31,233
11	NC_001846	II	MHV	Murine hepatitis virus	31,276
12	NC_001451	III	IBV	Avian infectious bronchitis virus	27,608
13	AY278488	IV	BJ01	SARS coronavirus BJ01	29,725
14	AY278741	IV	Urbani	SARS coronavirus Urbani	29,727
15	AY278491	IV	HKU–39849	SARS coronavirus HKU–39849	29,742
16	AY278554	IV	CUHK–W1	SARS coronavirus CUHK–W1	29,736
17	AY282752	IV	CUHK–Su10	SARS coronavirus CUHK–Su10	29,736
18	AY283794	IV	SIN2500	SARS coronavirus Sin2500	29,711
19	AY283795	IV	SIN2677	SARS coronavirus Sin2677	29,705
20	AY283796	IV	SIN2679	SARS coronavirus Sin2679	29,711
21	AY283797	IV	SIN2748	SARS coronavirus Sin2748	29.706
22	AY283798	IV	SIN2774	SARS coronavirus Sin2774	29,711
23	AY291451	IV	TW1	SARS coronavirus TW1	29,729
24	NC_004718	IV	TOR2	SARS coronavirus	29,751

2.1 Materials

The 24 complete coronavirus genomes used in this paper are downloaded from NCBI, of which 12 are SARS–CoVs and 12 are from other groups of coronaviruses. The name, accession number, abbreviation, and genome length for the 24 genomes are listed in Table 1. According to the existing taxonomic groups, sequences 1–3 belong to group I, and sequences 4–11 are members of group II, while sequence 12 is the only representative of group III. Refer to Table 1 for details.

2.2 LZ Complexity

Let ϕ be a finite alphabet. A sequence ***S*** with length *n* over the alphabet ϕ is an ordered *n*–tuple $S = s_1 s_2 \ldots s_n$ of symbols from ϕ. To indicate a substring of ***S*** that starts at position *i* and ends at position *j*, we write S[*i*:*j*]. That is, $S[i:j] = s_i s_{i+1} \ldots s_j$ for $i \le j$.

A general approach to the analysis of symbolic sequence complexity was proposed by Kolmogorov [16,17]. However, Kolmogorov complexity is not a recursive function, that is, it is not incorporated in a computational scheme, and thus generally can only be approximated [17,18]. The complexity measure proposed by Lempel and Ziv was an explicitly computable implementation of this approach for finite sequences, and many text compression algorithms are based on their measure [19,20]. The LZ complexity of a non–empty sequence ***S***, denoted by c(*S*), is defined as the minimal number of steps in some (optimal) procedure of its synthesis

$$\boldsymbol{S} = S[1:i_1]S[i_1+1 : i_2] \ldots S[i_{k-1} + 1 : i_k] \ldots S[i_{m-1} + 1 : n]$$

where $S[i_{k-1}+1:i_k]$ is a fragment (component) generated at the *k*–th step, and at each step, two operations are allowed: copy the longest fragment from the part of ***S*** that has already been synthesized and generate an additional symbol which ensures the uniqueness of each component $S[i_{k-1}+1:i_k]$ (the source code is given in Appendix 1).

Lempel and Ziv [19] called the complexity decomposition of a sequence **S** based on the rule above the exhaustive history of ***S***, and proved that every sequence ***S*** has a unique exhaustive history.

For example, the LZ complexity of the sequence ***S***=011111001011010100010111 amounts to 7, and this sequence can be generated through the following steps, where * is used to separate the decomposition component:

(i) generate a novel symbol **0**: Φ + 0 → 0

(ii) generate a novel symbol **1**: 0 + 1 → 0*1

(iii) copy the longest fragment + generate a additional symbol 1111**0**:

0*1 + 11110 → 0*1*11110

(iv) copy the longest fragment + generate a additional symbol 01**0**:

0*1*11110 +010 → 0*1*11110*010

(v) copy the longest fragment + generate a additional symbol 110**1**:

0*1*11110*010 + 1101 → 0*1*11110*010*1101

(vi) copy the longest fragment + generate a additional symbol 010**0**:

0*1*11110*010*1101 + 0100 → 0*1*11110*010*1101*0100

(vii) copy the longest fragment 1011**1**:

0*1*11110*010*1101*0100 + 10111→0*1*11110*010*1101*0100*10111,

and this is just the exhaustive history of ***S***.

2.3 Relative Distance Based on LZ Complexity

Given two sequences Q and S, by definition, the number of steps needed to build Q when appended to S is c(SQ)–c(S), and how much the degree of c(SQ)–c(S) is less than c(Q) will depend on the degree of similarity between S and Q [20].

For instance, S=110101001011, R=001111100010 and Q=110101101010.

The exhaustive histories of these sequences are as follows:

H(S)=1*10*10100*1011,

H(R)=0*01*11110*0010,

H(Q)=1*10*1011*01010,

which yield that c(S)=c(R)=c(Q)=4.

The exhaustive histories of sequences SQ, RQ and QR would be:

H(SQ)=1*10*10100*1011*1101011*01011,

H(RQ)=0*01*11110*0010*1101*0110*1010,

H(QR)=1*10*10110*1010*00*111*1100*010,

form which we obtain that: c(SQ)=6,c(RQ)=7 and c(QR)=8.

By above examples, we find that Q is similar to S than R for the reason why c(SQ)<c(RQ).

Therefore, one can use the following formula

$$D(P,Q) = f(P,Q) - \max\{f(P,P), f(Q,Q)\} \quad (1)$$

where $f(P,Q) = \dfrac{c(PQ) - c(P) + c(QP) + c(Q)}{c(PQ) + c(QP)}$, to describe the similarity degree of two sequences P and Q. For convenience, we call $D(P,Q)$ as the relative distance between the sequences P and Q.

3 RESULTS AND DISCUSSION

As we know, the four nucleic acid bases A, G, C and T can be divided into two classes according to their chemical structures, *i.e.*, purine R = {A, G} and pyrimidine Y = {C, T}. The bases can be also divided into another two classes, amino group M = {A, C} and keto group K = {G, T}. In addition, the division also can be made according to the strength of the hydrogen bond, i.e. weak H–bonds W ={A, T} and strong H–bonds S ={G, C}. By labeling the elements of R, M and W by 1, and that of Y, K and S by 0, respectively, three (0,1)–sequences corresponding to the same DNA sequence can be obtained. He and Wang [21,22] called them the characteristic sequences of the DNA sequence.

For example, the (M,K)–, (R,Y)–, and (W,S)–characteristic sequences of the segment ACTTTTAAAGTAAAGTGAGTGTAGCGTGGC, the first 30 bases of TEGV in Table 1, are 110000111001110001000010100001, 100000111101111011101011010110 and 101111111011110101010110001000, respectively.

Clearly, each characteristic sequence is a coarse–grained description for the DNA primary sequence. Although some information may be lost in characteristic sequence, we can get some information by comparing characteristic sequences that can not be obtained from the direct comparison of DNA sequences and observe some special nature in genome from different aspects.

For convenience, we denote by S_{MK} the (M,K)–characteristic sequence of a DNA sequence S. By definition, the LZ complexity of characteristic sequence S_{MK}^{TEGV}, the (M,K)–characteristic sequence of TEGV, is easily calculated as c(S_{MK}^{TEGV})=1947. Also, the LZ complexities of the characteristic sequences S_{MK}^{PEGV}, S_{MK}^{TEGV} S_{MK}^{PEGV}, S_{MK}^{PEGV} S_{MK}^{TEGV}, S_{MK}^{TEGV} S_{MK}^{TEGV} and S_{MK}^{PEGV} S_{MK}^{PEGV} are 1904, 3599, 3591, 1948 and 1905, respectively. By Eq.(1), we obtain $D(S_{MK}^{TEGV}, S_{MK}^{PEGV})$=0.9277. In the same way, we calculate the relative distances between any two among (M,K)–characteristic sequences of the 24 coronavirus genomic sequences, and list them in Table 2. Similarly, in Tables 3 and 4, we list similarities of the 24 coronavirus genomes based on the (R,Y)– and (W,S)–characteristic sequences, respectively.

From Tables 2, 3 and 4, we can obtain some information for each characteristic sequence. For example, the smallest entry (0.0001) in Table 2 is corresponding to Urbani–TW1, which indicates that the degree of the similarity between Urbani and TW1 is the highest among the 24 coronavirus genomes, while the same result cannot be obtained from Tables 3 and 4. On the other hand, Table 3 shows TW1 is more similar to SIN2774 than SIN2748, which is different from Tables 2 and 4. Meanwhile, differing from Tables 2 and 3, Table 4 shows BCoVL is more similar to HCoV–229E than TGEV. All these results illuminate that the three characteristic sequences reflect some essence of genome from different aspects, which is the function of the classifications.

Table 2. The similarity matrix of the (M,K)–characteristic sequences of the 24 coronavirus genomic sequences

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
1	0	0.9220	0.9157	0.9250	0.9259	0.9259	0.9261	0.9277	0.9266	0.9257	0.9273	0.9289
2		0	0.9224	0.9247	0.9247	0.9259	0.9251	0.9290	0.9261	0.9265	0.9289	0.9256
3			0	0.9277	0.9278	0.9280	0.9276	0.9277	0.9265	0.9269	0.9272	0.9271
4				0	0.0945	0.0574	0.0963	0.8834	0.8790	0.8801	0.8827	0.9271
5					0	0.1254	0.0206	0.8820	0.8792	0.8799	0.8817	0.9273
6						0	0.1272	0.8834	0.8799	0.8808	0.8827	0.9280
7							0	0.8820	0.8787	0.8796	0.8821	0.9277
8								0	0.5069	0.4015	0.0268	0.9276
9									0	0.1750	0.5038	0.9260
10										0	0.4004	0.9261
11											0	0.9271
12												0
13												
14												
15												
16												
17												
18												
19												
20												
21												
22												
23												
24												

	13	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**
1	0.9266	0.9266	0.9265	0.9264	0.9265	0.9263	0.9263	0.9263	0.9263	0.9262	0.9266	0.9267
2	0.9268	0.9270	0.9268	0.9270	0.9270	0.9267	0.9267	0.9267	0.9265	0.9267	0.9270	0.9270
3	0.9271	0.9271	0.9270	0.9271	0.9268	0.9268	0.9268	0.9266	0.9268	0.9267	0.9272	0.9272
4	0.9262	0.9257	0.9257	0.9258	0.9258	0.9257	0.9255	0.9257	0.9257	0.9255	0.9257	0.9259
5	0.9265	0.9261	0.9261	0.9260	0.9260	0.9260	0.9258	0.9260	0.9260	0.9258	0.9261	0.9262
6	0.9273	0.9268	0.9268	0.9271	0.9270	0.9268	0.9267	0.9268	0.9268	0.9267	0.9268	0.9271
7	0.9266	0.9262	0.9262	0.9262	0.9261	0.9260	0.9259	0.9260	0.9260	0.9259	0.9262	0.9263
8	0.9276	0.9274	0.9275	0.9275	0.9274	0.9274	0.9272	0.9274	0.9274	0.9272	0.9274	0.9275
9	0.9264	0.9263	0.9263	0.9265	0.9265	0.9262	0.9262	0.9262	0.9262	0.9261	0.9265	0.9267
10	0.9258	0.9254	0.9255	0.9256	0.9255	0.9254	0.9254	0.9254	0.9255	0.9253	0.9256	0.9259
11	0.9276	0.9273	0.9276	0.9273	0.9273	0.9273	0.9272	0.9273	0.9273	0.9272	0.9273	0.9276
12	0.9258	0.9258	0.9253	0.9258	0.9257	0.9258	0.9258	0.9258	0.9258	0.9257	0.9258	0.9258
13	0	0.0083	0.0122	0.0073	0.0093	0.0073	0.0088	0.0083	0.0098	0.0083	0.0083	0.0093
14		0	0.0044	0.0049	0.0049	0.0025	0.0039	0.0034	0.0049	0.0034	0.0001	0.0015
15			0	0.0088	0.0088	0.0069	0.0083	0.0078	0.0093	0.0078	0.0044	0.0054
16				0	0.0039	0.0044	0.0059	0.0054	0.0069	0.0054	0.0049	0.0054
17					0	0.0044	0.0059	0.0054	0.0069	0.0054	0.0049	0.0054
18						0	0.0015	0.0010	0.0025	0.0010	0.0025	0.0039
19							0	0.0025	0.0039	0.0025	0.0039	0.0054
20								0	0.0034	0.0020	0.0034	0.0049
21									0	0.0034	0.0049	0.0064
22										0	0.0034	0.0049
23											0	0.0015
24												0

Table 3. The similarity matrix of the (R,Y)–characteristic sequences of the 24 coronavirus genomic sequences

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
1	0	0.9264	0.9255	0.9296	0.9294	0.9300	0.9294	0.9274	0.9300	0.9304	0.9271	0.9285
2		0	0.9277	0.9302	0.9303	0.9296	0.9294	0.9269	0.9309	0.9309	0.9277	0.9266
3			0	0.9298	0.9308	0.9299	0.9297	0.9262	0.9271	0.9274	0.9259	0.9277
4				0	0.2691	0.0665	0.2681	0.9191	0.9203	0.9195	0.9194	0.9298
5					0	0.2834	0.0824	0.9179	0.9206	0.9205	0.9185	0.9288
6						0	0.2817	0.9197	0.9208	0.9198	0.9200	0.9301
7							0	0.9186	0.9204	0.9202	0.9193	0.9286
8								0	0.7643	0.6804	0.0386	0.9275
9									0	0.2101	0.7647	0.9297
10										0	0.6835	0.9296
11											0	0.9281
12												0
13												
14												
15												
16												
17												
18												
19												
20												
21												
22												
23												
24												

	13	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**
1	0.9286	0.9282	0.9277	0.9283	0.9283	0.9284	0.9283	0.9284	0.9286	0.9283	0.9280	0.9281
2	0.9289	0.9286	0.9278	0.9288	0.9288	0.9289	0.9292	0.9291	0.9292	0.9289	0.9280	0.9283
3	0.9281	0.9283	0.9281	0.9280	0.9283	0.9282	0.9282	0.9284	0.9285	0.9284	0.9280	0.9280
4	0.9328	0.9316	0.9319	0.9323	0.9321	0.9321	0.9321	0.9322	0.9324	0.9321	0.9319	0.9320
5	0.9328	0.9316	0.9320	0.9325	0.9326	0.9327	0.9327	0.9327	0.9330	0.9326	0.9319	0.9321
6	0.9330	0.9318	0.9320	0.9326	0.9323	0.9323	0.9323	0.9323	0.9324	0.9322	0.9319	0.9322
7	0.9325	0.9317	0.9319	0.9321	0.9324	0.9325	0.9326	0.9325	0.9328	0.9324	0.9319	0.9320
8	0.9295	0.9291	0.9285	0.9291	0.9294	0.9298	0.9295	0.9294	0.9298	0.9297	0.9290	0.9293
9	0.9214	0.9311	0.9303	0.9312	0.9312	0.9313	0.9313	0.9314	0.9313	0.9316	0.9308	0.9311
10	0.9306	0.9302	0.9396	0.9305	0.9303	0.9306	0.9306	0.9306	0.9307	0.9307	0.9300	0.9304
11	0.9302	0.9298	0.9293	0.9296	0.9299	0.9304	0.9302	0.9302	0.9304	0.9303	0.9297	0.9298
12	0.9279	0.9278	0.9271	0.9277	0.9277	0.9280	0.9279	0.9280	0.9281	0.9279	0.9275	0.9275
13	0	0.0147	0.0171	0.0088	0.0098	0.0118	0.0137	0.0108	0.0132	0.0137	0.0103	0.0108
14		0	0.0132	0.0127	0.0079	0.0093	0.0113	0.0083	0.0108	0.0113	0.0059	0.0064
15			0	0.0176	0.0108	0.0127	0.0147	0.0117	0.0142	0.0147	0.0093	0.0093
16				0	0.0069	0.0118	0.0137	0.0108	0.0132	0.0137	0.0103	0.0103
17					0	0.0049	0.0069	0.0039	0.0064	0.0069	0.0034	0.0034
18						0	0.0039	0.0030	0.0034	0.0039	0.0049	0.0059
19							0	0.0049	0.0054	0.0059	0.0069	0.0079
20								0	0.0044	0.0049	0.0039	0.0049
21									0	0.0054	0.0064	0.0074
22										0	0.0069	0.0078
23											0	0.0025
24												0

Table 4. The similarity matrix of the (W,S)–characteristic sequences of the 24 coronavirus genomic sequences

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
1	0	0.9284	0.9278	0.9282	0.9258	0.9284	0.9255	0.9315	0.9323	0.9326	0.9313	0.9292
2		0	0.9293	0.9299	0.9296	0.9297	0.9287	0.9302	0.9319	0.9314	0.9308	0.9263
3			0	0.9324	0.9318	0.9327	0.9309	0.9283	0.9309	0.9292	0.9290	0.9273
4				0	0.2928	0.0774	0.2942	0.9153	0.9189	0.9183	0.9162	0.9290
5					0	0.3052	0.0936	0.9141	0.9184	0.9187	0.9146	0.9283
6						0	0.3025	0.9146	0.9181	0.9181	0.9157	0.9285
7							0	0.9138	0.9183	0.9181	0.9149	0.9267
8								0	0.7678	0.6865	0.0374	0.9290
9									0	0.2076	0.7709	0.9287
10										0	0.6900	0.9293
11											0	0.9289
12												0
13												
14												
15												
16												
17												
18												
19												
20												
21												
22												
23												
24												

	13	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**
1	0.9286	0.9281	0.9281	0.9282	0.9280	0.9280	0.9283	0.9283	0.9280	0.9280	0.9286	0.9281
2	0.9280	0.9287	0.9286	0.9282	0.9288	0.9284	0.9287	0.9288	0.9284	0.9286	0.9290	0.9287
3	0.9274	0.9270	0.9268	0.9269	0.9270	0.9271	0.9272	0.9272	0.9271	0.9271	0.9268	0.9268
4	0.9290	0.9290	0.9289	0.9290	0.9289	0.9291	0.9294	0.9287	0.9290	0.9291	0.9291	0.9293
5	0.9291	0.9291	0.9292	0.9287	0.9289	0.9289	0.9290	0.9292	0.9289	0.9289	0.9292	0.9292
6	0.9291	0.9292	0.9291	0.9290	0.9289	0.9293	0.9295	0.9289	0.9292	0.9295	0.9293	0.9294
7	0.9288	0.9288	0.9289	0.9285	0.9285	0.9288	0.9292	0.9292	0.9288	0.9288	0.9289	0.9289
8	0.9296	0.9298	0.9295	0.9290	0.9291	0.9294	0.9296	0.9296	0.9294	0.9295	0.9295	0.9295
9	0.9319	0.9317	0.9318	0.9313	0.9314	0.9319	0.9323	0.9219	0.9318	0.9321	0.9318	0.9318
10	0.9301	0.9297	0.9298	0.9295	0.9294	0.9298	0.9301	0.9299	0.9298	0.9300	0.9298	0.9300
11	0.9303	0.9303	0.9300	0.9296	0.9297	0.9300	0.9302	0.9302	0.9300	0.9301	0.9300	0.9300
12	0.9287	0.9285	0.9285	0.9281	0.9286	0.9287	0.9288	0.9290	0.9285	0.9287	0.9288	0.9287
13	0	0.0201	0.0211	0.0136	0.0166	0.0166	0.0191	0.0166	0.0176	0.0176	0.0161	0.0166
14		0	0.0116	0.0151	0.0101	0.0096	0.0121	0.0096	0.0111	0.0106	0.0061	0.0076
15			0	0.0181	0.0111	0.0111	0.0136	0.0111	0.0126	0.0121	0.0076	0.0086
16				0	0.0111	0.0136	0.0161	0.0136	0.0146	0.0146	0.0131	0.0136
17					0	0.0066	0.0091	0.0066	0.0081	0.0076	0.0061	0.0066
18						0	0.0046	0.0041	0.0035	0.0030	0.0056	0.0071
19							0	0.0066	0.0061	0.0056	0.0081	0.0096
20								0	0.0056	0.0051	0.0056	0.0071
21									0	0.0046	0.0071	0.0086
22										0	0.0066	0.0081
23											0	0.0035
24												0

Table 5. The similarity matrix of the 24 coronavirus genomic sequences

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
1	0	0.9256	0.9230	0.9276	0.9270	0.9281	0.9270	0.9288	0.9296	0.9295	0.9285	0.9289
2		0	0.9265	0.9283	0.9282	0.9284	0.9277	0.9287	0.9296	0.9296	0.9291	0.9262
3			0	0.9300	0.9301	0.9302	0.9294	0.9274	0.9282	0.9278	0.9273	0.9274
4				0	0.2214	0.0670	0.2222	0.9060	0.9061	0.9060	0.9062	0.9286
5					0	0.2401	0.0655	0.9047	0.9062	0.9064	0.9050	0.9281
6						0	0.2392	0.9060	0.9064	0.9063	0.9062	0.9289
7							0	0.9049	0.9059	0.9060	0.9055	0.9277
8								0	0.6894	0.6003	0.0342	0.9280
9									0	0.1976	0.6899	0.9281
10										0	0.6024	0.9283
11											0	0.9280
12												0
13												
14												
15												
16												
17												
18												
19												
20												
21												
22												
23												
24												

	13	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**
1	0.9279	0.9276	0.9274	0.9276	0.9276	0.9276	0.9276	0.9277	0.9276	0.9275	0.9277	0.9276
2	0.9279	0.9281	0.9277	0.9280	0.9282	0.9280	0.9282	0.9282	0.9281	0.9281	0.9280	0.9280
3	0.9275	0.9275	0.9273	0.9273	0.9274	0.9274	0.9274	0.9274	0.9275	0.9274	0.9274	0.9274
4	0.9293	0.9287	0.9288	0.9290	0.9289	0.9289	0.9290	0.9288	0.9290	0.9289	0.9289	0.9291
5	0.9295	0.9289	0.9291	0.9290	0.9292	0.9292	0.9292	0.9293	0.9293	0.9291	0.9291	0.9292
6	0.9298	0.9293	0.9293	0.9295	0.9294	0.9295	0.9295	0.9293	0.9295	0.9294	0.9294	0.9296
7	0.9293	0.9289	0.9290	0.9289	0.9290	0.9291	0.9292	0.9292	0.9292	0.9290	0.9290	0.9291
8	0.9289	0.9288	0.9285	0.9285	0.9286	0.9288	0.9288	0.9288	0.9288	0.9288	0.9286	0.9288
9	0.9299	0.9297	0.9295	0.9296	0.9297	0.9298	0.9299	0.9298	0.9298	0.9299	0.9297	0.9299
10	0.9288	0.9284	0.9283	0.9285	0.9284	0.9286	0.9287	0.9286	0.9287	0.9287	0.9284	0.9287
11	0.9293	0.9291	0.9290	0.9288	0.9290	0.9292	0.9292	0.9292	0.9292	0.9292	0.9290	0.9291
12	0.9274	0.9274	0.9270	0.9272	0.9273	0.9275	0.9275	0.9276	0.9275	0.9274	0.9274	0.9273
13	0	0.0143	0.0168	0.0099	0.0119	0.0119	0.0138	0.0119	0.0135	0.0132	0.0115	0.0122
14		0	0.0097	0.0109	0.0076	0.0071	0.0091	0.0071	0.0089	0.0084	0.0040	0.0051
15			0	0.0148	0.0102	0.0102	0.0122	0.0102	0.0120	0.0115	0.0071	0.0078
16				0	0.0073	0.0099	0.0119	0.0099	0.0115	0.0112	0.0094	0.0097
17					0	0.0053	0.0073	0.0053	0.0071	0.0066	0.0048	0.0051
18						0	0.0033	0.0026	0.0031	0.0026	0.0043	0.0056
19							0	0.0046	0.0051	0.0046	0.0063	0.0076
20								0	0.0045	0.0040	0.0043	0.0056
21									0	0.0045	0.0061	0.0074
22										0	0.0056	0.0069
23											0	0.0025
24												0

As pointed out by He and Wang [21,22], the three characteristic sequences contain all information of the DNA sequence. Therefore, we calculate the average of the relative distances corresponding to the three characteristic sequences among the 24 coronavirus genomes, and list them in Table 5.

From Table 5, we find that the 24 coronavirus genomes can be classified into four groups on the whole:

1) The SARS–CoVs appear to cluster together, and can be distinguished easily from other three groups of the coronaviruses.
2) IBV, belonging to another group, is independent from other coronaviruses.
3) HCoV–229E, TGEV, and PEGV, tend to cluster together because their distances are less than that between each of them and other coronaviruses.
4) BCoV, BCoVL, BCoVM, BCoVQ, MHV, MHV2, MHVM, and MHVP form a group. Furthermore, one can see that among the 8 coronaviruses classified into two subgroups: one includes BCoV, BCoVL, BCoVM, BCoVQ, and the other includes MHV, MHV2, MHVM, and MHVP, respectively.

As mentioned in subsection 2.1, 12 are SARS–CoVs and 12 are from other groups of coronaviruses among the 24 complete coronavirus genomes. According to the existing taxonomic groups, group I includes HCoV–229E, TGEV, and PEGV, and group II contains BCoV, BCoVL, BCoVM, BCoVQ, MHV, MHV2, MHVM, MHVP. All the viruses in these groups are mammalian viruses. Group III contains only avian viruses, of which only the genome of IBV has been completely sequenced. In 2003, Rota *et al*. [23] have performed phylogenetic analysis based on sequence alignments using different genes, their result showed that SARS–CoVs are not closely related to any of the previously characterized coronaviruses and form a distinct group (group IV) within the genus coronavirus. The same results were obtained by Grigoriev [24], Gu *et al*. [25] and Zheng *et al*. [26]. Our result is consistent with these. This implies that the proposed method has apparently captured important features of similarity for DNA sequences considered, and is useful for similarity analysis of DNA sequences.

4 CONCLUSIONS

Based on the characteristic sequence of DNA sequence and LZ complexity of symbolic sequence, we propose a new distance measure for the similarity analysis of DNA sequences. It is well known that the alignment of DNA sequences is computer intensive that is a direct comparison of DNA sequences. The structure considered in sequences alignment is only string's structure. Here, we use an approach that considers not only sequences' structure but also chemical structure for

DNA sequences. On the other hand, our method avoids the complex calculation and thus can be directly used to handle long DNA sequences. To show the utility of the method, we use it to examine the similarities among the 24 complete coronavirus genomes. As a result, the 24 complete coronavirus genomes are classified into four groups on the whole, which is consistent with results reported in other literature.

Acknowledgment

This work was partially supported by the Science Research Project of Educational Department of Liaoning Province and the Natural Science Foundation of Liaoning Province of China.

Appendix 1

Source code of the LZ algorithm

```
%% GetLZ.m

Seq=S %% Input a sequence S with length>1
Length=size(Seq);
Len=Length(2);
History=Seq(1,1:2);
CeH=2;
i=2;
while i<Len
        i;
        Component=Seq(1,i);
        for k=i:1:Len-1
            lenofH=size(History,2);
            Hnolastone=Seq(1,1:lenofH-1);
            if size(findstr(Component,Hnolastone),1)>0
                Component=Seq(1,i:k+1);
                History=Seq(1,1:k+1);
                if k==Len-1
                end
            else
                CeH=CeH+1;
                History=Seq(1,1:k+1);
                break;
            end
        end
        k;
        i=k+1;
end
CeH    %% Output the LZ complexity of the sequence S.
```

5 REFERENCES

[1] S. Neeleman and C. Wunsch, A general method applicable to the search for similarities in the amino acid sequences of two protein, *J. Mol. Biol.* **1970**, *48*, 444–453.
[2] T. Smith and M. Sternberg, Identification of common molecular subsequences, *J. Mol. Biol.* **1981**, *147*, 195–197.
[3] W. Pearson and D. Lipman, Improved tools for biological sequence comparison, *Proc. Natl. Acad. Sci.* **1988**, *85*, 2444–2448.
[4] S. Altschul, W. Gish, W. Miller, E. Myers, and D. Lipman, Basic local alignment search tool, *J. Mol. Biol.* **1990**, *215*, 403–410.
[5] S. Altschul and W. Gish, Local alignment statistics, *Methods Enzymol.* **1996**, *266*, 460–480.

[6] M. Randić and M. Vračko, On the similarity of DNA Primary Sequences, *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 599–606.
[7] M. Randić, M. Vračko, and J. Zupan, Compact 2–D graphical representation of DNA, *Chem. Phys. Lett.* **2003**, *373*, 558–562.
[8] M. Randić, N. Lerš, D. Plavšić, S. C. Basak, and A. T. Balaban, Four–color map representation of DNA or RNA sequences and their numerical characterization, *Chem. Phys. Lett.* **2005**, *407*, 205–208.
[9] X. F. Guo, M. Randić, and S. C. Basak, A novel 2–D graphical representation of DNA sequences of low degeneracy, *Chem. Phys. Lett.* **2001**, *305*,106–112.
[10] M. Randić and J. Zupan, Highly compact 2D graphical representation of DNA sequences, *SAR. QSAR. Environ. Res.* **2004**, *15*(*3*), 191–205.
[11] M. Randić, M. Vračko, N. Lers, and D. Plavšić, Analysis of similarity/dissimilarity of DNA sequences based on novel 2–D graphical representation, *Chem. Phys. Lett.* **2003**, *371*, 202–207.
[12] M. Randić, M. Vračko, N. Lers, and D. Plavšić, Novel 2–D graphical representation of DNA sequences and their numerical characterization, *Chem. Phys. Lett.* **2003**, *368*, 1.
[13] M. Randić, M. Vračko, A. Nandy and S. C. Basak, On 3-D graphical representation of DNA primary sequences and their numerical characterization, *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1235–1244.
[14] M. Randić, X. F. Guo, and S. C. Basak, On the characterization of DNA primary sequences by triplet of nucleic acid bases, *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 619–626.
[15] C. Li and J. Wang, New invariant of DNA sequences, *J. Chem. Inf. Model.* **2005**, *45*,115.
[16] V. D. Gusev, L. A. Nemytikova, and N. A. Chuzhanova, On the complexity measures of genetic sequences, *Bioinformatics*. **1999**, *15*, 994–999.
[17] M. C. Thomas and A. T. Joy, Elements of Information Theory, *Beijing: Tsinghua University Press*. **2003**, 239–265.
[18] T. Jiang, Y. Xu, and M. Q. Zhang, Current Topics in Computational Molecular Biology, *Tsinghua University Press & the MIT Press*. **2002**, 345–364.
[19] A. Lempel and J. Ziv, On the complexity of Finite Sequences, *IEEE*. **1976**, *22*, 75–81.
[20] H. Out and K. Sayood, A new sequence distance measure for phylogenetic tree construction, *Bioinformatics*. **2003**, *19*, 2122–2130.
[21] P. A. He and J. Wang, Characteristic Sequences for DNA Primary Sequence, *J. chem. Inf. Comput. Sci.* **2002**, *42*, 1081–1085.
[22] P. A. He and J. Wang, Numerical characterization of DNA primary sequence, *Internet Electron. J. Mol. Des.* **2002**, *1*, 668.
[23] P. A. Rota, M. S. Oberste, S. S. Monroe, W. A. Nix, R. Campagnoli, J. P. Icenogle, S. Penaranda, B. Bankamp, K. Maher, M. H. Chen, S. Tong, A. Tamin, L. Lowe, M. Frace, J. L. DeRisi, Q. Chen, D. Wang, D. D. Erdman, T. C. Peret, C. Burns, T. G. Ksiazek, P. E. Rollin, A. Sanchez, S. Liffick, B. Holloway, J. Limor, K. McCaustland, M. Olsen–Rasmussen, R. Fouchier, S. Gunther, A. D. Osterhaus, C. Drosten, M. A. Pallansch, L. J. Anderson, and W. J. Bellini, Characterization of a novel coronavirus associated with severe acute respiratory syndrome, *Science*. **2003**, *300*, 1394–1399.
[24] A. Grigoriev, Mutational patterns correlate with genome organization in SARS and other coronaviruses, *Trends Genet*. **2004**, *20*, 131–135.
[25] W. Gu, T. Zhou, J. Ma, X. Sun, and Z. Lu, Analysis of synonymous codon usage in SARS coronavirus and other viruses in the Nidovirales, *Virus Res*. **2004**, *101*, 155–161.
[26] W. X. Zheng, L. L. Chen, H. Y. Ou, F. Gao, and C. T. Zhang, Coronavirus phylogeny based on a geometric approach, *Mol. Phyl. Evol*. **2005**, *36*, 224–232.