Journal of Mathematical Chemistry, Vol. 43, No. 1, January 2008 (© 2006)
DOI: 10.1007/s10910-006-9176-8

Similarity analysis of DNA sequences based on the generalized LZ complexity of (0,1)-sequences

Chun Li*

Department of Mathematics, Bohai University, Jinzhou 121000 and Department of Applied Mathematics, Dalian University of Technology, Dalian 116024, P. R. China
E-mail: lchlmb@yahoo.com.cn

Jun Wang

Department of Applied Mathematics and College of Advanced Science and Technology, Dalian University of Technology, Dalian 116024, P. R. China

Received 12 June 2006; revised 27 June 2006

Based on the permutation of a binary alphabet, four generalized LZ complexities of a (0,1)-sequence are introduced. Since the logical representation of a DNA primary sequence includes four logical sequences, a DNA primary sequence can be characterized by a 16-D vector whose entries are the complexities corresponding to the logical sequences. The utility of the new quantitative characterization of DNA sequences is illustrated by an examination of the similarity among the full β-globin genes of 11 different species.

KEY WORDS: DNA, complexity, (0,1)-sequence, permutation

1. Introduction

The analysis of similarity/dissimilarity of DNA sequences can be divided into two groups: the sequence alignment and the invariant-based comparison. In the former, a distance function or a score function is used to represent insertion, deletion, and substitution of letters in the compared structures. Such approaches have been hitherto widely used. However, the computational complexity and the inherent ambiguity of the alignment cost criteria are still the bottleneck problems. The latter is based on the quantitative characterization of DNA sequences by ordered sets of invariants derived from the sequences, such as the leading eigenvalues of all kinds of matrices [1–18]. However, a trouble we must face is that the calculation of the matrix or eigenvalues will become more and more difficult with the length of the sequence longer.

*Corresponding author.

Table 1
The full β-globin genes of 11 species.

Species	Database	ID	Location	Length (bp)
Human	EMBL	HSHBB	62187–63610	1424
Chimpanzee	EMBL	PTGLB1	4189–5532	1344
Gorilla	EMBL	GBGLOBIN	4538–5881	1344
Lemur	EMBL	LMHBB	154–1595	1442
Rat	EMBL	RNGLB	310–1505	1196
Mouse	EMBL	MMBGL1	275–1462	1188
Goat	EMBL	CHHBBAA	279–1749	1471
Bovine	EMBL	BTGL02	278–1741	1464
Rabbit	EMBL	OCBGLO	277–1419	1143
Opossum	EMBL	DVHBBB	467–2488	2022
Gallus	EMBL	GGGL02	465–1810	1346

Here we introduce a new similarity measure that is based on the generalized LZ complexity of (0,1)-sequences. The examination of the similarities/dissimilarities among the full β-globin genes of 11 different species (see table 1) shows the utility of our approach.

2. Generalized LZ complexity of (0,1)-sequences

As pointed out in [19], the complexity of a symbolic sequence reflects an ability to represent a sequence in a compact form based on some structural features of this sequence. The general approach to estimating the complexity of symbolic sequences (texts) was suggested by Kolmogorov. He proved that there exists an optimal algorithm or program for the text generation. Kolmogorov complexity is the length of the shortest code generating a given sequence. 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 [19–21]. The complexity measure proposed by Lempel and Ziv was an explicitly computable implementation of this approach for finite sequences, and many text compression algorithms were based on their measure [19, 21–27]. Lempel and Ziv suggested measuring the complexity of a sequence by the minimal number of steps required for its synthesis in a certain process. In their approach, at each step of that process two operations were allowed: generation of a new symbol, and copying a fragment from the part of the sequence that has already been synthesized. Based on this, the "DSIC" complexity was introduced (see [19,22,23]), which is defined as the least number of events required to generate a given sequence algorithmically. Events are as follows: (i) generation of a new symbol; and (ii) copying of a longest fragment from the already generated sequence portion in the following orientations: direct (D), symmetrical (S),

inverted (I) or direct complementary (C). Moreover, Gusev et al. [23] generalized the LZ complexity measure by taking into account isomorphic repeats in the text. It is showed that the isomorphic measure can be used for recognition of the local structural regularities in DNA sequences.

From the point of view of mathematics, a DNA primary sequence is a string over alphabet $\Omega = \{A, C, G, T\}$, and copying is an identical permutation of Ω, while complementary is such a permutation: $p = (AT)(GC)$. They are only two out of 24 permutations of the alphabet Ω. But for a binary alphabet $\boldsymbol{B} = \{0, 1\}$, the number of permutations of its is only equal to 2, they are: $p_1 = (0)(1)$, $p_2 = (01)$. On the other hand, neither the "DSIC" complexity nor Gusev's isomorphic measure requires generating a new symbol at every step. This leads to that the generated sequence may have many identical components. Taking into account of these, for each permutation $p_k (k = 1, 2)$, we define two generalized LZ complexities of a (0,1)-sequence S as follows:

(a) $C_{dpk}(S)$, defined as the minimal number of steps required for its synthesis in a certain process, at each step of which only a "continuous operation" is allowed: copy a longest fragment from the part of the sequence that has already been synthesized by (*direct*) permutation p_k and then generate an additional symbol which ensures the uniqueness of each component.

(b) $C_{ipk}(S)$, the difference from $C_{dpk}(S)$ is that the copying is based on the *inverted* permutation p_k.

Thus, for any (0,1)-sequence S, we obtain four complexities: $C_{dp1}(S)$, $C_{ip1}(S)$, $C_{dp2}(S)$, $C_{ip2}(S)$. For example, the corresponding complexities of the sequence $S = 0001101001000101$ are 6, 5, 6, and 7, respectively.

3. Characterization of DNA sequences with 16-D vectors

It is easy to see that the four bases A, G, C and T can be classified into two classes based on the knowledge of logic: A and not-A (G or C or T). Denoting A (not-A) by 1 (0), we reduce a DNA primary sequence into a binary sequence. The obtained (0,1)-sequence is called the A-sequence of the DNA sequence. In this representation, some information of the DNA sequence structure may be lost; however, this will give prominence to local information from adenine. Besides this we can similarly define other three (0,1)-sequences: G-sequence, C-sequence and T-sequence, respectively. Like we did in [28], here we call the four (0,1)-sequences the logical representation (or simply **LR**) of the DNA sequence considered. Clearly, the **LR** gives all information of the corresponding DNA primary sequence.

As mentioned in section 2, one can obtain four generalized LZ complexities from any (0,1)-sequence, therefore, a DNA primary sequence can be

Table 2
The 16-D complexity vectors associated with 11 full β-globin genes in table 1.

Species	human	chim.	gorilla	lemur	goat	bovine	mouse	rat	rabbit	oposs.	gallus
$C_{dp1}(\mathbf{LR}^{A})$	113	105	106	111	114	118	89	96	89	158	100
$C_{ip1}(\mathbf{LR}^{A})$	109	104	105	105	117	116	91	92	83	152	98
$C_{dp2}(\mathbf{LR}^{A})$	224	216	215	230	234	230	206	209	203	278	241
$C_{ip2}(\mathbf{LR}^{A})$	224	217	214	228	232	229	203	206	201	281	241
$C_{dp1}(\mathbf{LR}^{C})$	100	94	93	99	104	104	93	92	87	140	108
$C_{ip1}(\mathbf{LR}^{C})$	103	97	97	95	103	104	91	88	84	135	101
$C_{dp2}(\mathbf{LR}^{C})$	246	235	236	245	260	229	202	201	207	328	205
$C_{ip2}(\mathbf{LR}^{C})$	245	234	235	246	264	232	204	200	209	325	205
$C_{dp1}(\mathbf{LR}^{G})$	108	100	102	113	120	116	97	94	92	136	123
$C_{ip1}(\mathbf{LR}^{G})$	99	93	95	110	113	112	91	90	92	136	116
$C_{dp2}(\mathbf{LR}^{G})$	228	218	219	223	229	229	197	194	171	328	170
$C_{ip2}(\mathbf{LR}^{G})$	227	216	221	224	228	231	195	197	167	329	169
$C_{dp1}(\mathbf{LR}^{T})$	127	123	123	128	127	123	109	113	108	172	100
$C_{ip1}(\mathbf{LR}^{T})$	127	122	121	130	124	123	108	107	107	172	96
$C_{dp2}(\mathbf{LR}^{T})$	186	185	185	182	208	191	165	162	154	255	268
$C_{ip2}(\mathbf{LR}^{T})$	185	182	182	182	210	192	165	160	152	251	266

characterized by a 16-D vector whose entries are the complexities corresponding to the four logical sequences of the DNA sequence. In table 2, we list the 16-D complexity vector representations of full β-globin genes of 11 species in table 1. If desired, one can introduce weighting procedure that will normalize magnitudes of the complexities to reduce variations caused by different lengths of sequences. For instance, one can consider instead of the complexity $C_{.p.}(S)$ a normalized complexity $C_{.p.}(S)/n$, where n is the length of the DNA sequence considered.

4. Similarities and dissimilarities

In this section, we illustrate the use of the new quantitative characterization of DNA sequences with an examination of the similarity among the full β-globin genes of 11 different species of table 1. A direct comparison of these sequences using computer codes is somewhat less straightforward due to the fact that the sequences have different lengths, from 1143 (rabbit) to 2022 bp (opossum). If we represent the sequences with the corresponding 16-D complexity vectors then different lengths of the sequences do not cause difficulties. Moreover, the complexity can be normalized as outlined above.

The analysis of similarity/dissimilarity between two DNA sequences represented by the 16-D vectors is based on the assumption that two DNA sequences are similar if the corresponding 16-D vectors point to a similar direction and have similar magnitudes. The similarity between these two vectors can

Table 3
The similarity/dissimilarity matrix for the 11 full β-globin genes.

Species	human	chim.	gorilla	lemur	goat	bovine	mouse	rat	rabbit	oposs	gallus
human	0	0.0117	0.0123	0.0146	0.0230	0.0244	0.0292	0.0276	0.0387	0.0344	0.1209
chim.		0	0.0053	0.0221	0.0195	0.0304	0.0229	0.0235	0.0352	0.0426	0.1152
gorilla			0	0.0231	0.0194	0.0309	0.0238	0.0248	0.0378	0.0418	0.1170
lemur				0	0.0274	0.0232	0.0336	0.0304	0.0364	0.0371	0.1214
goat					0	0.0338	0.0297	0.0333	0.0349	0.0489	0.1049
bovine						0	0.0365	0.0328	0.0501	0.0336	0.1137
mouse							0	0.0118	0.0316	0.0587	0.1089
rat								0	0.0319	0.0563	0.1124
rabbit									0	0.0708	0.1104
oposs.										0	0.1361
gallus											0

be measured by calculating the Euclidean distance between their end points. Clearly, the smaller is the Euclidean distance the more similar are the two DNA sequences. In table 3 we give the similarity/dissimilarity among the full β-globin genes of 11 different species of table 1 based on the normalized 16-D complexity vectors.

Observing table 3, we find that the three kind of Primates (human, chimpanzee, gorilla) are strongly similar to each other, and so are mouse and rat, which are expected because of their evolutionary relationship. On the other hand, the largest entries in the similarity/dissimilarity matrix appear in the row belonging to gallus (the only non-mammalian representative). Also, opossum (the most remote species from the remaining mammals) shows great dissimilarity with others. Similar results have been obtained by other authors (see [5,6,10,13,17,18]). It should be mentioned that, in the result of [29], the degree of similarity of human-chimpanzee is obviously lower than that of human and other some species including rat, rabbit and especially gallus. This seems to be a disappointing phenomenon in the evolutionary sense. In the present work, this phenomenon no longer occurs.

5. Conclusion

In this paper, we introduce a new measure for similarity analysis that is based on the generalized LZ complexity of (0,1)-sequences. Unlike most existing comparison approaches, the proposed method does not require multiple alignment, and avoids the complex calculation as in the calculation of invariants of higher order matrices. The examination of similarities/dissimilarities among the full β-globin genes of 11 different species shows the utility of our approach.

Acknowledgment

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

References

[1] M. Randic and G. Krilov, Chem. Phys. Lett. 272 (1997) 115–119.
[2] M. Randic, M. Vracko, A. Nandy and S.C. Basak, J. Chem. Inf. Comput. Sci. 40 (2000) 1235.
[3] M. Randic and M. Vracko, J. Chem. Inf. Comput. Sci. 40 (2000) 599–606.
[4] M. Randic, Chem. Phys. Lett. 317 (2000) 29.
[5] M. Randic, X.F. Guo and S.C. Basak, J. Chem. Inf. Comput. Sci. 41 (2001) 619.
[6] P-an He and J. Wang, J. Chem. Inf. Comput. Sci. 42 (2002) 1080.
[7] P-an He and J. Wang, Internet Elect. J. Mol. Des. 1 (2002) 668.
[8] A.T. Balaban, D. Plavsic and M. Randic, Chem. Phys. Lett. 379 (2003) 147–154.
[9] M. Randic, M. Vracko, N. Lers and D. Plavsic, Chem. Phys. Lett. 368 (2003) 1.
[10] M. Randic, M. Vracko, N. Lers and D. Plavsic, Chem. Phys. Lett. 371 (2003) 202.
[11] M. Randic, J. Zupan and A.T. Balaban, Chem. Phys. Lett. 397 (2004) 247–252.
[12] M. Randic, N. Lers, D. Plavsic, S.C. Basak and A.T. Balaban, Chem. Phys. Lett. 407 (2005) 205–208.
[13] C. Li and J. Wang, Comb. Chem. High T. Scr. 6 (2003) 795–799.
[14] C. Li and J. Wang, J. Chem. Inf. Model. 45 (2005) 115–120.
[15] C. Li, N.N. Tang and J. Wang, Directed graphs of DNA sequences and their numerical characterization, J. Theor. Biol. 241 (2006) 173–177.
[16] M. Ji and C. Li, TB-curve, a New 2-D Graphical Representation of DNA Sequences, *J. Math. Chem.*, Online First (to appear: 40 (2006)).
[17] N. Liu and T.M. Wang, Chem. Phys. Lett. 408 (2005) 307–311.
[18] Y.H. Yao and T.M. Wang, Chem. Phys. Lett. 398 (2004) 318–323.
[19] Y.L. Orlov and V.N. Potapov, Nucleic Acids Res. 32 (2004) W628–W633.
[20] T.M. Cover and J.A. *Thomas, Elements of Information Theory* (Tsinghua University Press, Beijing 2003).
[21] T. Jiang, Y. Xu and M.Q. Zhang, *Current Topics in Computational Molecular Biology* (Tsinghua University Press and The MIT Press, Tsinghua and Cambridge 2002) pp. 157–171.
[22] V.N. Babenko, P.S. Kosarev, O.V. Vishnevsky, V.G. Levitsky, V.V. Basin and A.S. Frolov, Bioinformatics 15 (1999) 644–653.
[23] V.D. Gusev, L.A. Nemytikova and N.A. Chuzhanova, Bioinformatics 15 (1999) 994–999.
[24] A. Lempel and J. Ziv, IEEE T. Inform. Theory 22 (1976) 75–81.
[25] J. Ziv and A. Lempel, IEEE T. Inform. Theory 23 (1977) 337–343.
[26] J. Ziv and A. Lempel, IEEE T. Inform. Theory 24 (1978) 530–536.
[27] H.H. Otu and K. Sayood, Bioinformatics 19 (2003) 2122–2130.
[28] C. Li and J. Wang, A naturally logical representation for DNA primary sequences, *Comb. Chem. High T. Scr.* (2006) (in press).
[29] B. Liao and T.M. Wang, Chem. Phys. Lett. 388 (2004) 195–200.