ORIGINAL PAPER

A generalization of Lempel-Ziv complexity and its application to the comparison of protein sequences

Chun Li · Zhengxing Li · Xiaoqi Zheng · Hong Ma · Xiaoqing Yu

Received: 29 October 2008 / Accepted: 19 March 2010 / Published online: 30 March 2010

Abstract In this paper, a complexity measure of symbolic sequences is proposed that generalizes the Lempel-Ziv complexity by taking into account a specific kind of the inexact copy in the text, and based on which, a new sequence distance measure for the similarity analysis is introduced. The utility of our approach is illustrated by an examination of the relationships among β-globin proteins of 13 species.

Keywords Amino acid · Comparison · Complexity · Protein sequence · Substring

1 Introduction

The sequence similarity between different species provides an important reference for the phylogenetic analysis although it doesn't decide the final result of the phylogenetic analysis completely. Approaches for comparative studies of biological 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, which have been hitherto widely used, are computer intensive. The latter is based on the quantitative characterization of biological sequences by ordered sets of invariants derived from the sequences. To obtain the invariant, one often follows the strategy below:

C. Li (✉) · Z. Li · H. Ma · X. Yu
Department of Mathematics, Bohai University, 121013 Jinzhou, People's Republic of China
e-mail: lchlmb@yahoo.com.cn

X. Zheng
Department of Mathematics, Shanghai Normal University, 200034 Shanghai,
People's Republic of China

Step 1: Represent a biological sequence by some mathematical object of fixed geometry, such as graph, or a set of lines;
Step 2: For the selected mathematical object, construct its numerical representation in the form of a matrix or set of matrices;
Step 3: From obtained matrices extract a set of invariants.

Graphical representations of DNA sequences were initiated about 25 years ago by Hamori et al. [1,2] and Gates [3], whose pioneering work was soon followed by introduction of alternative such representations (see [4–31]). While the graphical representations of proteins emerged only very recently [32–40]. It should be mentioned that most of graphical representations of DNA involve some degree of arbitrariness, such as the selection of directions to be assigned to individual bases. Therefore, extension of DNA graphical representations to those of proteins would increase enormously the number of possible alternative assignments for the 20 amino acids making such generalizations unacceptable, which is probably the most important reason why graphical representations of proteins have not been advanced. The matrices associated with a graph include the ED, D/D, L/L, and their 'higher order' matrices [13,15–19,23–38]. Once a real symmetric matrix M is given, one often uses some of matrix invariants, such as the leading eigenvalue and the ALE-index, as descriptors of the sequence [13,15–18,23–37]. However, a trouble we must face is that the calculation of some effective invariants will become more and more difficult with the length of the sequence longer.

In this paper, we propose a generalized Lempel-Ziv complexity of a protein sequence, and based on which, we introduce a new sequence distance measure for the similarity analysis. The examination of the relationships among β-globin proteins of 13 species (see Table 1) shows the utility of our approach.

Table 1 The β-globin proteins of 13 species

Species	Database	Accession number	Length (aa)
Human	GenBank	AAA16334	147
Gorilla	GenBank	CAA43421	121
Chimpanzee	GenBank	CAA26204	125
Lemur	GenBank	AAA36822	147
Rabbit	GenBank	CAA24251	147
European hare	GenBank	CAA68429	147
Goat	GenBank	AAA30913	145
Sheep	GenBank	NP_001091117	145
Bovine	GenBank	CAA25111	145
Mouse	GenBank	CAA24101	147
Rat	GenBank	CAA29887	147
Opossum	GenBank	AAA30976	147
Gallus	GenBank	CAA23700	147

2 Methods

2.1 Preliminaries

Let Ω be a finite alphabet. A sequence x with length n over the alphabet Ω is an ordered n-tuple $x = x_1x_2 \cdots x_n$ of symbols from Ω. The empty sequence, that is the string with zero symbols, is denoted by φ. The set of all sequences over an alphabet Ω is denoted as Ω^*. The concatenation of two sequences x and y forms a new sequence xy. We call w a substring of sequence x if x is of the form uwv for $u, v \in \Omega^*$. We also say that substring w occurs at position $|u| + 1$ of sequence x, where $|u|$ represents the length of sequence u. The starting position of w in x is the position $|u| + 1$ while position $|u| + |w|$ is said to be the end position of w in x. In general the substring of x starting at position i and ending at position j, inclusively, is denoted by $x[i : j]$.

2.2 The generalization of Lempel-Ziv complexity

A general approach to estimating the complexity of an object as a finite automation-generated model was suggested by A.N. Kolmogorov. However, Kolmogorov complexity is not a recursive function and thus can not be incorporated in a computational scheme [41]. The complexity measure proposed by Lempel and Ziv is an explicitly computable implementation of this approach for finite sequences, and many text compression algorithms are based on their measure [41–45]. The Lempel-Ziv complexity of a non-empty sequence x, denoted by c(x), is defined as the minimal number of steps in some (optimal) procedure of its synthesis

$$x = x[1 : i_1]x[i_1 + 1 : i_2] \cdots x[i_{k-1} + 1 : i_k] \cdots x[i_{m-1} + 1 : n]$$

where $x[i_{k-1}+1 : i_k]$ is a substring (component) generated at the k-th step, and at each step k, two operations are allowed: copying the longest fragment from the part of $\boldsymbol{x}$ that has already been synthesized plus generating an additional symbol which ensures the uniqueness of each component $x[i_{k-1} + 1 : i_k]$. The complexity decomposition of a sequence x based on this rule is called the exhaustive history of x [44–46].

The copy used in the Lempel-Ziv model is an *exact* copy of a substring that starts somewhere earlier in the sequence. When we consider the approximate version of this problem we do not require an exact matching but something that is “similar” in some way. We will now focus our attention on a generalization of the Lempel-Ziv complexity measure designed for the alphabet of the 20 natural amino acids.

As it is known, a protein primary sequence can be taken as a string of letters on an alphabet Ω = {A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y}. For better physical understanding and practical purposes, much effort has been made by considering minimalist models with a few types of amino acid residues to simplify the natural set of residues of 20 types. In these models the compositions are much simpler than the real ones. The simplest reduction is the well-known HP model. The studies of such a model enable people to understand some fundamental physics and mechanism of protein folding. However, as argued in a number of studies (see [47,48]), the HP model may be too simple and lacks enough consideration on the heterogeneity and the complexity of the natural set of residues, such as the interactions between the residues. Moreover, the minimal sets of residues for protein design suggested by biochemical experiments seem unfavorable to those with only two types of residues since a small number of types obviously introduces the homopolymeric degeneracy. In 1997, by using combinatorial chemistry along with a screening strategy, Riddle et al. searched and found out a subset of the natural amino acids that can be used to construct a well-ordered proteinlike molecule consisting of β sheets. This subset contains five amino acids: isoleucine, alanine, glycine, glutamic acid and lysine, which are simply represented as I, A, G, E, and K (see [47–49]). Three years after that, based on the statistical and the kinetic characteristics of the folding, and on the thermodynamic stability of the ground states of some reduced sequences, Wang and Wang [47,48] proved that the suggested five-letter code is valid in general and feasible for elucidating characteristics of real proteins with 20 kinds of amino acids. As a matter of fact, the 20 natural amino acids in this model are classified into five groups according to their interaction characteristics: group-I (with residues C, M, F, I, L, V, W, and Y), group-II (A, T, and H), group-III (G and P), group-IV (D and E), and group-V (S, N, Q, R, and K). Each group contains some residues which interact with others in a similar way. Moreover, for the five groups, letters I, A, G, E, and K are taken as the best representative letters, respectively. This selection is based on a physical reason, but not an arbitrary choice (see [47,48,50]).

Using the classification above, we define a homomorphism map f by $f(x) = f(x_1)f(x_2)\cdots f(x_n)$, where $x = x_1x_2\cdots x_n \in \Omega^*$ and

$$f(x_j) = \begin{cases} \text{I} & \text{if } x_j \in \{\text{C, M, F, I, L, V, W, Y}\} \\ \text{A} & \text{if } x_j \in \{\text{A, T, H}\} \\ \text{G} & \text{if } x_j \in \{\text{G, P}\} \\ \text{E} & \text{if } x_j \in \{\text{D, E}\} \\ \text{K} & \text{if } x_j \in \{\text{S, N, Q, R, K}\} \end{cases}, \quad (j = 1, 2, \ldots, n).$$

For two fragments $x = x_1x_2\cdots x_n$ and $y = y_1y_2\cdots y_n$, we call them “co-image” under the homomorphism map f if $f(x) = f(y)$. That is, $f(x_1) = f(y_1)$, $f(x_2) = f(y_2)$, …, $f(x_n) = f(y_n)$. We thus define the complexity measure $c_f(x)$, allowing for copying of fragments which are “co-image” under f.

For example, the generalized Lempel-Ziv complexity of the sequence
$x =$ MVHLTPEEKPDEVDSG amounts to seven, and this sequence can be generated through the following steps, where * is used to separate the decomposition component:

(i) generating a novel symbol M: $\varphi + \text{M} \rightarrow \text{M}$

(ii) “copying” the longest fragment + generating a novel symbol H: M + VH → M*VH

(iii) “copying” the longest fragment + generating a additional symbol P: M*VH + LTP → M*VH*LTP

(iv) generating a novel symbol E: M*VH*LTP + E → M*VH*LTP*E

(v) “copying” the longest fragment + generating a additional symbol K:

$$\text{M*VH*LTP*E} + \text{EK} \rightarrow \text{M*VH*LTP*E*EK}$$

(vi) “copying” the longest fragment + generating a additional symbol V:

$$\text{M*VH*LTP*E*EK} + \text{PDEV} \rightarrow \text{M*VH*LTP*E*EK*PDEV}$$

(vii) “copying” the longest fragment:
M*VH*LTP*E*EK*PDEV + DSG → M*VH*LTP*E*EK*PDEV*DSG, and this is just the exhaustive history of x.

3 Results and discussion

For any given sequences w and v, by definition, the number of steps needed to build w when appended to v is $c_f(vw) - c_f(v)$. It is not difficult to see that $c_f(vw) - c_f(v) \leq c_f(w)$ always holds. This shows that the steps required to extend v to vw are always less than the steps required to build w from φ. Therefore, the relative similarity degree of two sequences w and v can be described by the following formula:

$$d_r(w, v) = d(w, v) - \frac{1}{2}(d(w, w) + d(v, v)), \tag{1}$$

where $d(w, v) = \frac{c_f(wv) - c_f(w) + c_f(vw) - c_f(v)}{c_f(wv) + c_f(vw)}$. For convenience, we call $d_r(w, v)$ as the relative distance between sequences w and v.

In general, given m protein sequences $S_1, S_2, \ldots, S_m$, we first make pair-concatenation operation on the m sequences, and then we get $m^2 + m$ sequences. Without loss of generality, we let

$$\begin{aligned} &S_{11} = S_1 S_1, \quad S_{12} = S_1 S_2, \ldots, \quad S_{1m} = S_1 S_m, \ldots, \\ &S_{m1} = S_m S_1, \quad S_{m2} = S_m S_2, \ldots, \quad S_{mm} = S_m S_m. \end{aligned}$$

By this means, for any $i, j = 1, 2, \ldots, m$

$$\begin{aligned} d_r(S_i, S_j) &= d(S_i, S_j) - \frac{1}{2}\left(d(S_i, S_i) + d(S_j, S_j)\right) \\ &= \frac{c_f(S_{ij}) - c_f(S_i) + c_f(S_{ji}) - c_f(S_j)}{c_f(S_{ij}) + c_f(S_{ji})} \\ &\quad - \frac{1}{2}\left[\frac{c_f(S_{ii}) - c_f(S_i)}{c_f(S_{ii})} + \frac{c_f(S_{jj}) - c_f(S_j)}{c_f(S_{jj})}\right] \end{aligned} \tag{2}$$

For the 13 protein sequences in Table 1, we calculate the corresponding generalized Lempel-Ziv complexities c_f's and list them in Table 2. Then by Eq. (2) we calculate the relative distances between any two of the 13 protein sequences. Consequently, a 13×13 real symmetric matrix $D = \left(d_r(S_i, S_j)\right)_{13\times 13}$ is obtained (see Table 3). The relationship tree (see Fig. 1) is constructed using the UPGMA program included in MEGA 4.0. The branch lengths are not scaled according to the distances and only the topology of the tree is concerned.

Table 2 The c_f's of the corresponding sequences

c_f	Human	Gorilla	Chimp	Lemur	Rabbit	E_hare	Goat	Sheep	Bovine	Mouse	Rat	Opossum	Gallus
	42	36	38	43	43	43	41	41	42	44	42	43	40
Human	43	43	44	57	52	51	60	60	55	59	56	62	64
Gorilla	43	37	38	56	51	50	57	57	53	56	55	58	60
Chimp	46	40	39	58	54	53	60	60	56	59	57	61	61
Lemur	58	56	57	44	59	59	62	62	61	62	59	64	68
Rabbit	53	52	53	58	44	47	58	58	56	56	55	62	65
E_hare	52	51	52	58	47	44	59	59	56	56	55	62	65
Goat	60	57	58	62	57	58	42	42	50	62	59	62	65
Sheep	60	57	58	62	57	58	42	42	50	62	59	62	65
Bovine	55	52	53	61	55	55	51	51	43	62	59	63	66
Mouse	61	58	59	61	56	57	65	65	64	45	57	64	68
Rat	57	55	56	57	55	55	60	60	60	55	43	63	66
Opossum	63	60	61	65	62	63	65	65	64	63	65	44	65
Gallus	63	59	60	66	65	64	63	63	64	65	65	63	41

Table 3 Lower triangles of the relative distance matrix

d_r	Human	Gorilla	Chimp.	Lemur	Rabbit	E_hare	Goat	Sheep	Bovine	Mouse	Rat	Opossum	Gallus
Human	0												
Gorilla	0.0679	0											
Chimp.	0.0867	0.0249	0										
Lemur	0.2379	0.2698	0.2715	0									
Rabbit	0.1675	0.2081	0.2188	0.2422	0								
E_hare	0.1518	0.1929	0.2044	0.2422	0.0624	0							
Goat	0.2848	0.2991	0.3058	0.2993	0.2463	0.2588	0						
Sheep	0.2848	0.2991	0.3058	0.2993	0.2463	0.2588	0.0000	0					
Bovine	0.2131	0.2320	0.2416	0.2803	0.2112	0.2112	0.1547	0.1547	0				
Mouse	0.2606	0.2736	0.2812	0.2702	0.2007	0.2076	0.3077	0.3077	0.2947	0			
Rat	0.2334	0.2658	0.2676	0.2442	0.2043	0.2043	0.2790	0.2790	0.2709	0.2094	0		
Opossum	0.2970	0.3056	0.3119	0.3106	0.2837	0.2893	0.3153	0.3153	0.3077	0.2925	0.3129	0	
Gallus	0.3305	0.3356	0.3304	0.3570	0.3380	0.3330	0.3431	0.3431	0.3454	0.3451	0.3502	0.3280	0

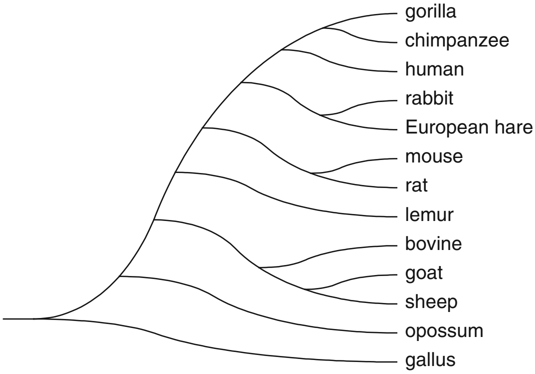

Fig. 1 The relationship tree of the 13 protein sequences

Observing Fig. 1, we find that gallus, the only non-mammalian representative, is situated at an independent branch, while the 12 mammals appear to cluster together and form a separate branch. A closer look at the subtree of mammals shows that human, gorilla, and chimpanzee tend to cluster together. Also, (European hare, rabbit) and (mouse, rat) tend to cluster together, respectively, while goat, sheep and bovine form a separate branch. On the other hand, opossum can be distinguished easily from the remaining mammals. This result is similar to that reported in other literature [13,17,23,28,29,46]. The conclusion one can draw from these findings is that the proposed Lempel-Ziv complexity measure may be a useful tool for protein comparative studies.

Acknowledgments This work was partially supported by the National Natural Science Foundation of China and the Natural Science Foundation of Liaoning Province.

References

1. E. Hamori, J. Ruskin, J. Biol. Chem. **258**, 1318–1327 (1983)
2. E. Hamori, Nature **314**, 585–586 (1985)
3. M.A. Gates, J. Theor. Biol. **119**, 319–328 (1986)
4. H.I. Jeffrey, Nucleic. Acid Res. **18**, 2163–2170 (1990)
5. R. Zhang, C.T. Zhang, J. Biomol. Struct. Dyn. **11**, 767–782 (1994)
6. P.M. Leong, S. Morgenthaler, Comput. Appl. Biosci. **12**, 503–511 (1995)
7. A. Nandy, Curr. Sci. **66**, 309–313 (1994)
8. A. Nandy, Curr. Sci. **66**, 821 (1994)
9. A. Roy, C. Raychaudhury, A. Nandy, J. Biosci. **23**, 55–71 (1998)
10. D. Bielińska-Wąż, T. Clark, P. Wąż, W. Nowak, A. Nandy, Chem. Phys. Lett. **442**, 140–144 (2007)
11. D. Bielińska-Wąż, W. Nowak, P. Wąż, A. Nandy, T. Clark, Chem. Phys. Lett. **443**, 408–413 (2007)
12. A. Nandy, S.C. Basak, B.D. Gute, J. Chem. Inf. Model. **47**, 945–951 (2007)
13. M. Randić, M. Vracko, A. Nandy, S.C. Basak, J. Chem. Inf. Comput. Sci. **40**, 1235–1244 (2000)
14. X.F. Guo, M. Randić, S.C. Basak, Chem. Phys. Lett. **350**, 106–112 (2001)
15. M. Randić, A.T. Balaban, J. Chem. Inf. Comput. Sci. **43**, 532–539 (2003)
16. M. Randić, M. Vracko, N. Lers, D. Plavsić, Chem. Phys. Lett. **368**, 1–6 (2003)
17. M. Randić, M. Vracko, N. Lers, D. Plavsić, Chem. Phys. Lett. **371**, 202–207 (2003)
18. M. Randić, M. Vracko, J. Zupan, M. Novic, Chem. Phys. Lett. **373**, 558–562 (2003)
19. M. Randić, Chem. Phys. Lett. **386**, 468–471 (2004)
20. M. Randić, Chem. Phys. Lett. **456**, 84–88 (2008)

21. S.S.T. Yau, J. Wang, A. Niknejad, C. Lu, N. Jin, Y.K. Ho, Nucleic. Acids Res. **31**, 3078–3080 (2003)
22. Y.H. Wu, A.W. Liew, H. Yan, M. Yang, Chem. Phys. Lett. **367**, 170 (2003)
23. Y.H. Yao, T.M. Wang, Chem. Phys. Lett. **398**, 318–323 (2004)
24. M. Ji, C. Li, J. Math. Chem. **40**, 185–193 (2006)
25. C. Li, J. Wang, Comb. Chem. High T. Scr. **7**, 23–27 (2004)
26. C. Li, J. Wang, J. Chem. Inf. Model. **45**, 115–120 (2005)
27. C. Li, N.N. Tang, J. Wang, J. Theor. Biol. **241**, 173–177 (2006)
28. C. Li, J. Hu, J. Biochem. Mol. Biol. **39**, 292–296 (2006)
29. M. Randić, X.F. Guo, S.C. Basak, J. Chem. Inf. Comput. Sci. **41**, 619–626 (2001)
30. G. Jaklic, T. Pisanski, M. Randić, J. Comput. Biol. **13**, 1558–1564 (2006)
31. A. Nandy, M. Harle, S.C. Basak, ARKIVOC **(ix)**, 211–238 (2006)
32. M. Randić, SAR QSAR Environ. Res. **15**, 147–157 (2004)
33. M. Randić, J. Zupan, A.T. Balaban, Chem. Phys. Lett. **397**, 247–252 (2004)
34. M. Randić, A.T. Balaban, M. Novic, A. Zaloznik, T. Pisanski, Period Boil. **107**, 403–414 (2005)
35. M. Randić, D. Butina, J. Zupan, Chem. Phys. Lett. **419**, 528–532 (2006)
36. M. Randić, J. Zupan, D. Vikić-Topić, J. Mol. Graph. Model **26**, 290–305 (2007)
37. M. Randić, Chem. Phys. Lett. **444**, 176–180 (2007)
38. M. Novic, M. Randić, SAR QSAR Environ. Res. **19**, 317–337 (2008)
39. Z.G. Yu, V. Anh, K.S. Lau, J. Theor. Biol. **226**, 341–348 (2004)
40. G. Aguero-Chapin, H. Gonzalez-Diaz, R. Molina, J. Varona-Santos, E. Uriarte, Y. Gonzalez-Diaz, FEBS Lett. **580**, 723–730 (2006)
41. Y.L. Orlov, V.N. Potapov, Nucleic. Acids Res. **32**, W628–W633 (2004)
42. V.N. Babenko, P.S. Kosarev, O.V. Vishnevsky, V.G. Levitsky, V.V. Basin, A.S. Frolov, Bioinformatics **15**, 644–653 (1999)
43. V.D. Gusev, L.A. Nemytikova, N.A. Chuzhanova, Bioinformatics **15**, 994–999 (1999)
44. A. Lempel, J. Ziv, IEEE T. Inform. Theory **22**, 75–81 (1976)
45. H.H. Otu, K. Sayood, Bioinformatics **19**, 2122–2130 (2003)
46. C. Li, J. Wang, J. Math. Chem. **43**, 26–31 (2008)
47. J. Wang, W. Wang, Nat. Struct. Biol. **6**, 1033–1038 (1999)
48. J. Wang, W. Wang, Phys. Rev. E. **61**, 6981–6986 (2000)
49. D.S. Riddle, J.V. Santiago, S.T. Brayhall, N. Doshi, V.P. Grantcharova, Q. Yi, D. Baker, Nat. Struct. Biol. **4**, 805–809 (1997)
50. H.S. Chan, Nat. Struct. Biol. **6**, 994–996 (1999)