J Math Chem (2009) 46:1203–1212
DOI 10.1007/s10910-008-9512-2

ORIGINAL PAPER

Normalized Lempel-Ziv complexity and its application in bio-sequence analysis

Yi Zhang · Junkang Hao · Changjie Zhou · Kai Chang

Received: 9 June 2008 / Accepted: 28 November 2008 / Published online: 13 December 2008

Abstract In this article, we propose a new method to measure DNA similarity based on a normalized Lempel-Ziv complexity scheme. The new method can weaken the effect of sequence length on complexity measurement and save computation time. Firstly, a DNA sequence is transformed into three (0,1)-sequences based on a scheme, which considers “A” and “non-A” , “G” and “non-G”, “C” and “non-C” bases respectively. Then, the normalized Lempel-Ziv complexity of the three (0,1)-sequences constitute a 3D vector. Finally, by the 3D vector, one may characterize DNA sequences and compute similarity matrix for them. The examination of similarities of two sets of DNA sequences illustrates the utility of the method in local and global similarity analysis.

Keywords Lempel-Ziv complexity · Normalized · Similarity · Relationship tree

1 Introduction

For a given bio-sequence with known structure and functions, similarity analysis may find similar regions in other sequences, from which we can infer that they may have the

Y. Zhang (✉) · C. Zhou
Department of Mathematics, Hebei University of Science and Technology, Shijiazhuang,
HeBei 050018, People’s Republic of China
e-mail: zhaqi1972@163.com

J. Hao
Physical Education Department, Hebei University of Science and Technology, Shijiazhuang,
HeBei 050018, People’s Republic of China

K. Chang
Department of Automatic Control, School of Information Science and Technology,
Beijing Institute of Technology, Beijing 100081, People’s Republic of China

same biological functions as the given sequence. The methods of similarity analysis of bio-sequences include sequence alignment, sequence descriptor comparison and complexity method. Sequence alignment is the procedure of comparing two (pairwise alignment) or more (multiple sequence alignment) bio-sequences by searching for a series of individual characters or character patterns that are in the same order in the sequences. In an optimal alignment, nonidentical characters and gaps are placed to bring as many identical or similar characters as possible into vertical register. Sequences that can be readily aligned in this manner are said to be similar [1,2]. The second class is derived from the quantitative characterization of DNA sequences by invariants (or descriptors), such as the maximum eigenvalues of some kinds of matrices [3–9].

Third, Otu and Sayood [10] proposed a new sequence distance measure based on the "relative information" between the sequences using Lempel-Ziv complexity to successfully construct phylogenetic trees. Liu and Wang [11] also applied the idea to the analysis of similarity of DNA sequences. In such similarity analysis, the Lempel-Ziv complexity can not be used as an invariant to characterize a bio-sequence because it is firmly correlated with sequence length (as shown by the two upper sub-figures in Fig. 1). Moreover, when we compute similarity matrix for a set of bio-sequences by relative information scheme, a process concatenating bio-sequences and much computation have to be involved [12,13], which makes it hard to analyze long (longer than 10,000 bases) sequences by relative information scheme. In this article, we used a normalized Lempel-Ziv complexity (denoted as NLZ for short) to characterize DNA sequences, and the new scheme has following characteristics.

(1) NLZ is independent of sequence length, and hence is a better invariant than Lempel-Ziv complexity in characterizing a DNA sequence.
(2) The new scheme can save runtime greatly in computing similarity matrix.
(3) The algorithm owns strong stability. For local and global similarity analysis, it shows perfect behavior.

In order to examine the validity of the new method in analyzing local and global similarity, we analyze the similarity for two sets of DNA sequences. Set I consists of the first exon sequences of β-globin genes of 10 species (shown in Table 1), and the set II includes 24 mitochondrial genomes.

2 Lempel-Ziv complexity [14] and normalized Lempel-Ziv complexity

Given a finite alphabet Ω, let U, V and W be sequences over it, $L(U)$ be the length of U, $U(i)$ be the i-th element of U and $U(i, j)$ be the subsequence of U starting at position i and ending at position j. Here $U(i, j) = \emptyset$, for $i > j$. Concatenating V and W can construct a new sequence $U = VW$, in this equation, V is named "a prefix" of U, and U is called "an extension" of V if there exists an integer i such that $V = U(1, i)$. An extension $U = VW$ of V is reproducible from V denoted by $V \rightarrow U$, if there exists an integer $P \leq L(V)$ such that $W(k) = U(p + k - 1)$, for $k = 1, 2, \ldots, L(W)$. A non-null sequence U is producible from its prefix $U(1, j)$, denoted by $U(1, j) \Rightarrow U$, if $U(1, j) \rightarrow U(1, L(U) - 1)$. For example: $01 \Rightarrow 0100$ with $p = 1$. Note that, the producibility allows for an extra different symbol at the end

Table 1 The coding sequences of the exon 1 of β-globin gene of 10 different species

Species	Coding sequence	Length
Bovine	ATGCTGACTGCTGAGGAGAAGGCTGCCGTCACCGCCTTTTGGGGCAAGGTGAAAGTGGATGAAGTTGGTGGTGAGGCCCTGGGCAG	86
Chimpanzee	ATGGTGCACCTGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGTGGGGCAAGGTGAACGTGGATGAAGTTGGTGGTGAGGCCCTGGGCAGGTTGGTATCAAGG	105
Gallus	ATGGTGCACTGGACTGCTGAGGAGAAGCAGCTCATCACCGGCCTCTGGGGCAAGGTCAATGTGGCCGAATGTGGGGCCGAAGCCCTGGCCAG	92
Gorilla	ATGGTGCACCTGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGTGGGGCAAGGTGAACGTGGATGAAGTTGGTGGTGAGGCCCTGGGCAGG	93
Human	ATGGTGCACCTGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGTGGGGCAAGGTGAACGTGGATGAAGTTGGTGGTGAGGCCCTGGGCAG	92
Lemur	ATGACTTTGCTGAGTGCTGAGGAGAATGCTCATGTCACCTCTCTGTGGGGCAAGGTGGATGTAGAGAAAGTTGGTGGCGAGGCCTTGGGCAG	92
Mouse	ATGGTTGCACCTGACTGATGCTGAGAAGTCTGCTGTCTCTTGCCTGTGGGCAAAGGTGAACCCCGATGAAGTTGGTGGTGAGGCCCTGGGCAGG	94
Opossum	ATGGTGCACTTGACTTCTGAGGAGAAGAACTGCATCACTACCATCTGGTCTAAGGTGCAGGTTGACCAGACTGGTGGTGAGGCCCTTGGCAG	92
Rabbit	ATGGTGCATCTGTCCAGTGAGGAGAAGTCTGCGGTCACTGCCCTGTGGGGCAAGGTGAATGTGGAAGAAGTTGGTGGTGAGGCCCTGGGC	90
Rat	ATGGTGCACCTAACTGATGCTGAGAAGGCTACTGTTAGTGGCCTGTGGGGAAAGGTGAACCCTGATAATGTTGGCGCTGAGGCCCTGGGCAG	92

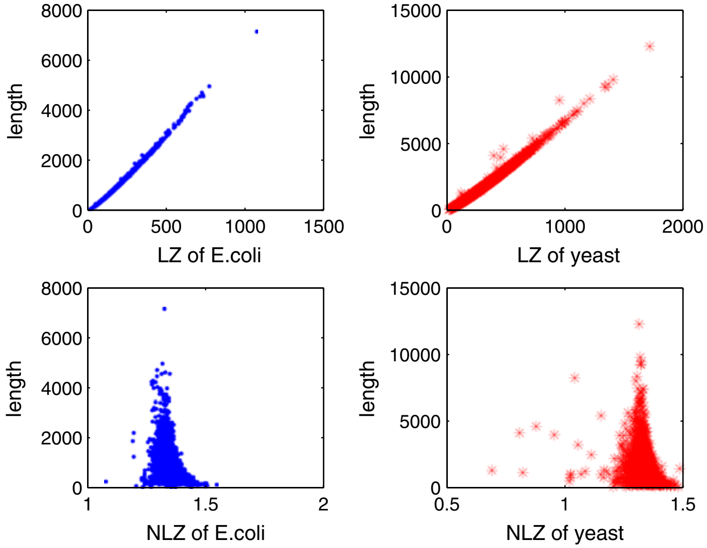

Fig. 1 A plot of LZ and NLZ versus sequence length for all *E. coli* and *yeast* genes

of the extension process, but the reproducibility does not. So, "reproducible" always means "producible", but the reverse may not always be true.

Given a non-null sequence U, it always can be made from a production process by iterative self-deleting-building process, where at the i-th step $U(1, h_{i-1}) \Rightarrow U(1, h_i)$, $\emptyset = U(1, 0) \Rightarrow U(1, 1)$. An m-step process producing U may lead to a parsing of U into $H(U) = U(1, h_1) \cdot U(h_1 + 1, h_2) \cdots U(h_{m-1} + 1, h_m)$, it is named the "history" of U, and $H_i(U) = U(h_{i-1} + 1, h_i)$ is called the i-th component of $H(U)$. A component $H_i(U)$ and the corresponding production step $U(1, h_{i-1}) \Rightarrow U(1, h_i)$ are called "exhaustive" if $U(1, h_{i-1}) \Rightarrow U(1, h_i)$ is false. If each of its components (with a possible exception of the last one) is exhaustive, then the history is also called exhaustive. Significantly, the exhaustive history of any non-null sequence is unique. For example, for the sequence $U = 00010110101$, its exhaustive history is $EXHI(U) = 0 \cdot 001 \cdot 011 \cdot 0101$. The number of components in the exhaustive history of U is denoted as $c(U)$, which is the least possible number of steps needed to generate U according to the rules of production process [14].

Although $c(U)$ is an important complexity indicator, it can not be directly used as an invariant in computing distance matrix, because it depends on sequence length strongly. Taking the *S. cerevisiae* (i.e. yeast) and *E. coli* K12 as examples, which are both taken from http://pedant.gsf.de/ of MIPS (the Munich Information Center for Protein Sequences). We first compute Lempel-Ziv complexity and measure sequence length for each gene of the two genomes respectively, then draw figures to show the correlation between length and Lempel-Ziv complexity (as represented by upper two sub-figures of Fig. 1). From the upper two sub-figures, one can see the Lempel-Ziv complexity is strongly correlated with sequence length. Minutely, the Spearman's rank correlation coefficients $r_s = 0.9993$ for *E. coli*, and $r_s = 0.9988$ for *yeast*, where $p < 10^{-12}$. Without normalization, comparing Lempel-Ziv complexities of two sequences is nearly equivalent to comparing their lengths. In fact, relative information schemes [10–13] merely use the difference between sequence LZ complexities to generate distance matrix, instead of using the Lempel-Ziv complexity as an indicator to characterize a sequence directly in similarity analysis. One may expect a new scheme measuring complexity, it can reduce the effect of length and give a more reasonable measurement for sequence complexity. In [14], Lempel and Ziv have stated that the asymptotic behavior of $c(U)$ in the case of uniformly distributed symbols is given by $\frac{n}{\log n}$, where n is the length of the sequence. Therefore we may take $\frac{c(U)}{\frac{n}{\log n}}$ as a normalized Lempel-Ziv complexity to characterize DNA sequences. From the Fig. 1, we see the correlation between Lempel-Ziv complexity and length is large, $r_s = 0.9993$ for *E. coli*, and $r_s = 0.9988$ for yeast, while the correlation between NLZ and length is small, $r_s = -0.4343$ for *E. coli*, and $r_s = -0.3087$ for yeast. Clearly, Fig. 1 illustrates that NLZ has much less dependence on sequence length than Lempel-Ziv complexity does. Following similarity analysis will show the advantage of NLZ in characterizing DNA sequences and similarity analysis.

3 Characterizing DNA sequences with normalized Lempel-Ziv complexity and constructing relationship tree

Similar to the scheme used in [5], for a given DNA sequence, its four bases can be divided into three classes according to their chemical structure, i.e., non-A = G, C, T, non-G = A, C, T and non-C = A, G, T. By labeling the elements of non-A, non-G and non-C by 1, and the elements of A, G, and C by 0, we can gain three (0,1)-sequences $S_{AA'}$, $S_{GG'}$, $S_{CC'}$. Based on the NLZ values $NLZ_{AA'}$, $NLZ_{GG'}$ and $NLZ_{CC'}$ of the three (0,1)-sequences, we construct a 3D vector $\{NLZ_{AA'}, NLZ_{GG'}, NLZ_{CC'}\}$ to characterize the DNA sequence uniquely. In Table 2, we listed the components of 3D vectors characterizing first exon sequences of β-globin genes of 10 species respectively.

Table 2 The $NLZ_{AA'}$, $NLZ_{GG'}$, $NLZ_{CC'}$ corresponding to the 10 DNA sequences shown in Table 1

Species	$NLZ_{AA'}$	$NLZ_{GG'}$	$NLZ_{CC'}$
Bovine	0.6733	0.7769	0.6215
Chimpanzee	0.6205	0.8421	0.5319
Gallus	0.6389	0.7864	0.7372
Gorilla	0.6336	0.8773	0.5361
Human	0.6389	0.8847	0.5406
Lemur	0.6881	0.8355	0.5406
Mouse	0.5317	0.7733	0.6767
Opossum	0.6881	0.7864	0.6881
Rabbit	0.6	0.8	0.55
Rat	0.6389	0.7864	0.6389

Table 3 The similarity/dissimilarity matrix for the 10 DNA sequences of Table 1 based on the quotient Q of the 3D vectors $\{NLZ_{AA'}, NLZ_{GG'}, NLZ_{CC'}\}$

	Bovine	Chimpanzee	Gallus	Gorilla	Human	Lemur	Mouse	Opossum	Rabbit	Rat
Bovine	0	0.1234	0.1216	0.1386	0.14	0.1013	0.1532	0.0689	0.1053	0.0397
Chimpanzee		0	0.2165	0.0377	0.0472	0.0685	0.1856	0.1807	0.0503	0.1227
Gallus			0	0.2242	0.2232	0.2114	0.1242	0.0696	0.1934	0.0985
Gorilla				0	0.0102	0.0689	0.2052	0.1873	0.0855	0.1382
Human					0	0.0696	0.2088	0.1859	0.0938	0.1399
Lemur						0	0.2199	0.1566	0.0956	0.121
Mouse							0	0.1583	0.1476	0.1149
Opossum								0	0.1653	0.0696
Rabbit									0	0.0983
Rat										0

Once bio-sequences are represented by vectors, the similarity between them can be described via their vectors. Here, we use these 3D vectors mentioned above to investigate similarities and dissimilarities for 10 coding sequences of Table 1. The underlying assumption is that if two vectors point to a similar direction in the 3D space and have similar magnitudes, then the two DNA sequences represented by the two 3D vectors are similar. Taking into account of the two aspects, similar to Randic et al. done in [3], we examine the similarities and dissimilarities by the quotients of the Euclidean distances between the end-points of the vectors and the cosines of the correlation angle of two vectors. For convenience, we denote such quotient by Q. Clearly, the smaller is the Q, the more similar are the DNA sequences. In Table 3, we list similarity matrix for 10 DNA sequences of set I.

From Table 3, we see that the three kinds of primates (human, chimpanzee, gorilla) DNA sequences are strongly similar to each other, for the smallest entries are associated with the pairs human–chimpanzee, human–gorilla and gorilla–chimpanzee. Gallus shows great dissimilarity with other among the 10 species, because almost all entries belonging to gallus are large. This is coincident with the results reported by Randic et al. [3], He and Wang [15] and Liu [16]. Next to these, the opossum shows great dissimilarity with others for it is the only pouched animal listed here. Lemur shows greater similarities to human, gorilla and chimpanzee than to the other species, perhaps because Lemur is a primate.

For comparison we list some results of the examinations of the degree of similarity of human and other several species in Table 4. As one can see there exists an overall agreement among similarities obtained by different approaches, despite some variation among them. Especially, lemur, as shown in [5], is similar to human than the other four species. It might also suggest lemur, one kind of primate animal, is related with human more closely. Besides, Table 4 shows the functions d_1^{**} and d^* of [10] fail to distinguish opossum from non-pouched animals. This may mean the newly defined NLZ is a more effective scheme in analyzing DNA similarity than d_1^{**} and d^*. Moreover, via calculating the correlation coefficients between our data and other scientists'

Table 4 The degree of similarity of the coding sequences of several species with the coding sequence of human, the data were normalized; the last column shows the correlation coefficients between this work and other methods

Species	Gallus	Opossum	Lemur	Rabbit	Rat	Correlation coefficients
This work	1.0000	0.8329	0.3118	0.4203	0.6268	
Based on [10, function d_1^{**}]	1.0000	0.9337	0.9648	0.6319	0.9131	0.4772
Based on [10, function d^*]	1.0000	0.8823	0.9412	0.5294	0.8823	0.4791
From [3, Table 12]	1.0000	0.8955	0.5922	0.6323	0.9685	0.8868
From [15, Table 9]	1.0000	0.9381	0.9092	0.5066	0.8307	0.5754
From [8, Table 6]	1.0000	0.7681	0.5835	0.0003	0.8063	0.7154
From [6, Table 5]	1.0000	0.6943	0.8781	0.5695	0.8450	0.3989

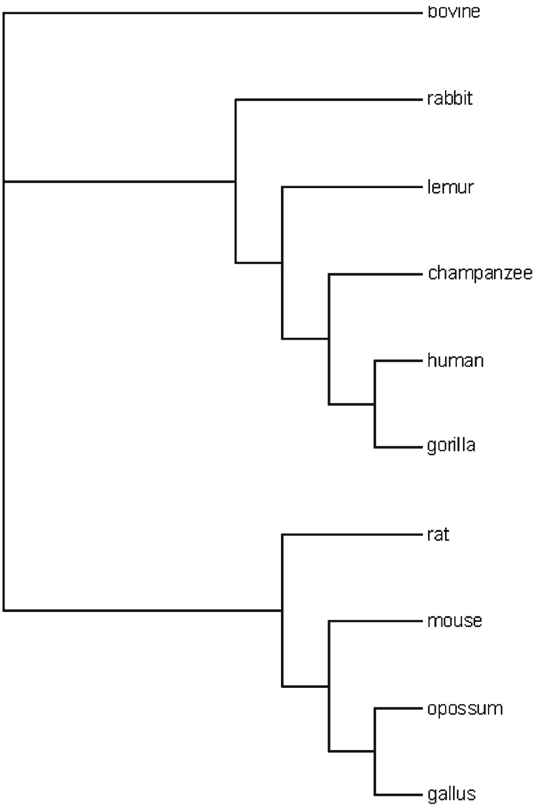

Fig. 2 The relationship tree for the 10 DNA sequences in Table 1 by KITSCH method based on our scheme

data, one may see our new method is strongly correlated with Randic's popularly accepted scheme (correlation coefficient is 0.8868). It also presents the significance and validity of NLZ.

To further check our result, we construct a relationship tree (shown in Fig. 2) for the 10 DNA sequences using the KITSCH method in Phylip 3.65 package.

Table 5 The accession number and length of 24 mitochondrial genomes

Species	Accession number	Length (bp)
Blue-whale	NC001601	16,402
Fin-whale	NC001321	16,398
Opossum	NC003039	17,191
Baboon	NC001992	16,521
Bonobo	NC001644	16,563
Bornean-orangutan	NC001646	16,389
Cat	NC001700	17,009
Chimpanzee	NC001643	16,554
Common-gibbon	NC002082	16,472
Cow	NC006853	16,338
Dog	NC002008	16,727
Donkey	NC001788	16,670
Gorilla	NC001645	16,364
Gray-seal	NC001602	16,797
Harbor-seal	NC001325	16,826
Hippopotamus	NC000889	16,407
Horse	NC001640	16,660
Human	NC001807	16,571
White-rhinoceros	NC001808	16,832
Indian-rhinoceros	NC001779	16,829
Pig	NC000845	16,613
Platypus	NC000891	17,091
Sheep	NC001941	16,616
Sumatran-orangutan	NC002083	16,499

It is not difficult to see that the similarity obtained by our method is coincident with that implicated in the tree. In fact, for the 10 sequences, our scheme gets the same relationship tree as Liu and Wang's (see formula (1) of [11]) does. Noticeably, our algorithm can save computation time greatly. For example, in calculating the similarity matrix for the 10 DNA sequences, by our algorithm, one only needs to calculate the complexity for 30 (0,1)-sequences, each is about 100 bp long; while by the formula (1) of [11], besides calculating complexity for 10 DNA (each about 100 bp), one needs to calculate complexity for other 90 DNA sequences, each approximately includes 200 bases. It shows that our scheme is efficient and concise in analyzing local similarity of DNA sequences.

In order to verify the validity of our method in analyzing global similarity of DNA sequences, we apply it to the set II, which includes 24 mitochondrial genomes (shown in Table 5).

By our method, a relationship tree for the 24 DNA sequences is got, we show it in Fig. 3. For concision, corresponding 3D vectors and similarity matrixes are not shown.

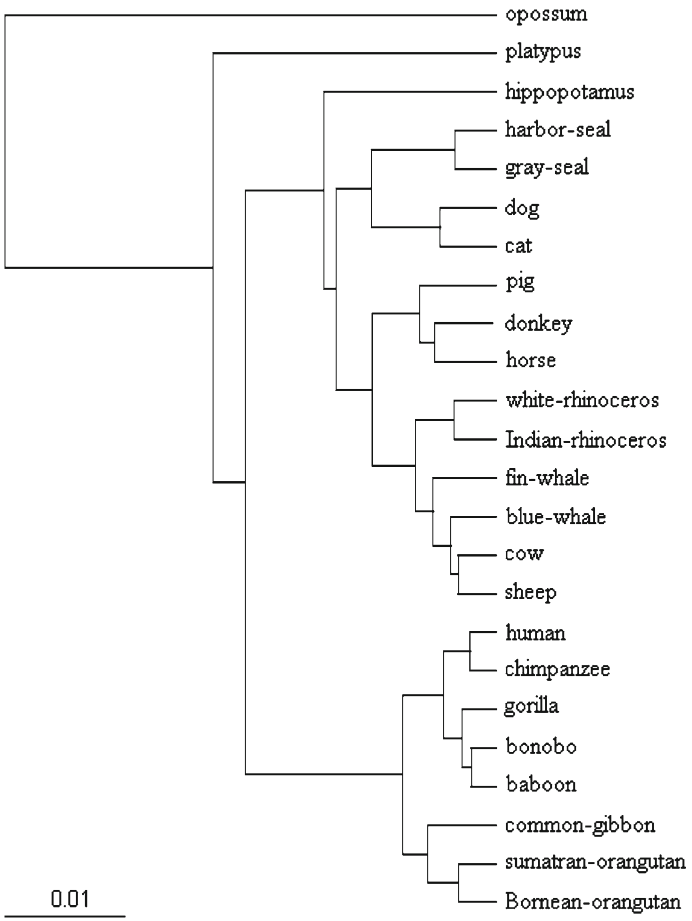

Fig. 3 The relationship tree for the 24 mitochondrial genomes by KITSCH method based on our scheme

In Fig. 3, the relationship tree of 24 species are reasonable. Two pouched out-groups, i.e. platypus and opossum, are grouped together, and located far away from others. All primates, including Human, Gorilla, Bonobo, Chimpanzee, Baboon, Common-gibbon, Sumatran-orangutan and Bornean-orangutan, are close in the relationship tree. Additionally, the close relationships between other species, including gray-seal and harbor-seal, fin-whale and blue-whale, Indian-rhinoceros and white-rhinoceros, horse and donkey, cat and dog, cow and sheep, are also reasonable. Most relationships are consistent with those shown in [17], besides a more reasonable one: cat and dog are grouped together and located near seals. Undoubtedly, these results illustrate the stability and reliability of our algorithm in analyzing global similarity of DNA sequences. As seen from above discussions, the new algorithm based on the normalized Lempel-Ziv complexity (NLZ) is helpful to local and global similarity analysis of DNA sequences. Perhaps, being modified, the NLZ scheme can deal with similar questions appeared in other fields, such as analyzing similarity for amino acid or codon sequences.

4 Conclusion

In this article, based on the normalized Lempel-Ziv complexity, a new method is brought forth to analyze local and global similarity for DNA sequences. Outgoing most existing complexity approaches, the new method can save computer runtime greatly. Moreover, in order to verify the validity of our method, the relationship trees for two sets of DNA sequences are shown. The reasonability of results confirm the stability and reliability of our algorithm.

Acknowledgements The author thanks all the anonymous referees for their valuable suggestions and support. This work is partially supported by Hebei Province Education Foundation (Z2008111).

References

1. D.W. Mount, *Bioinformatics: Sequence and Genome Analysis* (Cold Spring Harbor Laboratory, 2004)
2. M.S. Waterman, *Introduction to Computational Molecular Biology* (Chapman & Hall, London, 1995)
3. M. Randic, X.F. Guo, S.C. Basak, J. Chem. Inf. Comput. Sci. **41**, 619 (2001)
4. M. Randic, J. Chem. Inf. Comput. Sci. **40**, 50 (2000)
5. J. Wang, Y. Zhang, Chem. Phys. Lett. **423**, 50 (2006)
6. J. Wang, Y. Zhang, Chem. Phys. Lett. **425**, 324 (2006)
7. Y. Zhang, J. Wang, J. Math. Chem. **43**, 864 (2008)
8. C. Li, J. Wang, Comb. Chem. High Throughput Screen. **7**, 23 (2004)
9. B. Liao, T.M. Wang, J. Comput. Chem. **25**(11), 1364 (2004)
10. H.H. Otu, K. Sayood, **19**(16), 2122 (2003)
11. N. Liu, T.M. Wang, Chem. Phys. Lett. **408**, 307 (2005)
12. C. Li, A.-H. Wang, L. Xing, J. Comput. Chem. **28**, 508 (2007)
13. N. Liu, T.M. Wang, BMC Bioinformatics **7**, 493 (2006)
14. A. Lempel, J. Ziv, IEEE Trans. Inform. Theory **22**, 75 (1976)
15. P. He, J. Wang, J. Chem. Inf. Comput. Sci. **42**, 1080 (2002)
16. Y. Liu, Internet Electron. J. Mol. Des. **1**, 675 (2002)
17. J. Barral. P, L. Cantinib, A. Hasmy, J. Jimenezc, A. Marcano, J. Theor. Biol. **236**, 422 (2005)