Journal of Mathematical Chemistry, Vol. 43, No. 2, February 2008 (© 2007)
DOI: 10.1007/s10910-007-9235-9

Volatilities of codons and its application in similarity analysis of biological sequences

Yi Zhang

Department of Applied Mathematics, Dalian University of Technology, Dalian 116024, People's Republic of China

Jun Wang*

Department of Applied Mathematics and College of Advanced Science and Technology, Dalian University of Technology, Dalian 116024, People's Republic of China
E-mail: junwang@dlut.edu.cn

Received 22 November 2006; Revised 19 January 2007

Volatilities of codons provide us a new way to characterize codons. In this article, we propose a new method measuring volatilities of codons base on the physics–chemical distances between amino acids and mutation frequencies between codons, then by which, we give a new graphical representation scheme for codon sequences. Finally, in order to show the effectiveness of our scheme, we analyze similarity among the coding codon sequences of exon 1 of beta-globin gene of human and those of other 10 species, find the result is consistent with those shown in the literature.

KEY WORDS: Volatilities of codons, distance, mutation frequency, similarity

1. Introduction

With more and more DNA databases becoming available in public databases, DNA sequences analysis is increasingly showing its value. In many biological study fields such as sequences compare or genes identification, as an important visible means, graphical representation methods are widely used to directly obtain information from the DNA sequences [1–14]. On the other hand, matrix methods [15–23] are often used to characterize DNA sequences. Narrowing attention to the graphical representation of proteins, Randic et al. proposed highly condensed graphical representation in refs. [24, 25] and a novel graphical representation of proteins that produces an 8×8 tabular representation of 64 codons, and the corresponding table of amino acids in ref. [26]. Enlightened by their work, we try to introduce a new index to graphical representation of codon sequences.

* Corresponding author

In ref. [27], Plotkin et al. use the "volatility" of a codon to qualify the chance that the most recent nucleotide mutation to that codon caused an amino-acid substitution, certainly, volatility provides us a new way to characterize codons. But as shown below, being applied to the graphical representation and similarity analysis, the volatility is not effective enough, partially due to the low variance between volatilities of different codons. Base on the physics–chemical distances between amino acids and mutation frequencies between different codons, we give an new scheme to measure volatilities of 61 codons in the universal genetic code, and produce a new graphical scheme to characterize codon sequences. Finally, we can see that the new scheme used here are quite

Table 1
The coding sequences of the exon 1 of beta-globin gene of 11 different species.

Species	Coding sequence	length
Bovine	ATGCTGACTGCTGAGGAGAAGGCTGCCGTCACCGCCTTTTGGGG CAAGGTGAAAGTGGATGAAGTTGGTGGTGAGGCCCTGGGCAG	86
Chimpanzee	ATGGTGCACCTGACTCCTGAGGAGAAGTCTGCCGTTACTGCCC TGTGGGGCAAGGTGAACGTGGATGAAGTTGGTGGTGAGGCCC TGGGCAGGTTGGTATCAAGG	105
Gallus	ATGGTGCACTGGACTGCTGAGGAGAAGCAGCTCATCACCGG CCTCTGGGGCAAGGTCAATGTGGCCGAATGTGGGGCCGAAG CCCTGGCCAG	92
Goat	ATGCTGACTGCTGAGGAGAAGGCTGCCGTCACCGGCTTCT GGGGCAAGGTGAAAGTGGATGAAGTTGGTGCTGAGGCCC TGGGCAG	86
Gorilla	ATGGTGCACCTGACTCCTGAGGAGAAGTCTGCCGTTACTG CCCTGTGGGGCAAGGTGAACGTGGATGAAGTTGGTGGTG AGGCCCTGGGCAGG	93
Human	ATGGTGCACCTGACTCCTGAGGAGAAGTCTGCCGTTACTG CCCTGTGGGGCAAGGTGAACGTGGATGAAGTTGGTGGTGA GGCCCTGGGCAG	92
Lemur	ATGACTTTGCTGAGTGCTGAGGAGAATGCTCATGTCACCTC TCTGTGGGGCAAGGTGGATGTAGAGAAAGTTGGTGGCGA GGCCTTGGGCAG	92
Mouse	ATGGTGCACCTGACTGATGCTGAGAAGTCTGCTGTCTCTTG CCTGTGGGGAAAGGTGAACTCCGATGAAGTTGGTGGTGA GGCCCTGGGCAG	92
Opossum	ATGGTGCACTTGACTTCTGAGGAGAAGAACTGCATCACT ACCATCTGGTCTAAGGTGCAGGTTGACCAGACTGGTGGTGA GGCCCTTGGCAG	92
Rabbit	ATGGTGCATCTGTCCAGTGAGGAGAAGTCTGCGGTCACTG CCCTGTGGGGCAAGGTGAATGTGGAAGAAGTTGGTGGTGAG GCCCTGGGC	90
Rat	ATGGTGCACCTAACTGATGCTGAGAAGGCTACTGTTAGTGG CCTGTGGGGAAAGGTGAACCCTGATAATGTTGGCGCTGAGG CCCTGGGCAG	92

useful in similarity analysis among the coding sequences of the exon 1 of human and ten other species beta-globins in table 1.

2. A new scheme measuring volatilities of codons

In ref. [27], Plotkin et al. use the "volatility" of a codon to qualify the chance that the most recent nucleotide mutation to that codon caused an amino-acid substitution, in other words, it is the proportion of point mutations in a gene, which do not yield a stop codon, which change an amino acid. Based on the universal genetic code shown in table 2, they define the volatility of codon c by the equation:

$$\text{volatility}(c) = \frac{1}{\text{no. of neighbors}} \times \sum_{\text{neighbors } c_i} D[\text{acid}(c), \text{acid}(c_i)], \qquad (1)$$

where they sum over those non-stop codons c_i that can mutate into c by a single point mutation. They use the simplest possible measure D: the Hamming metric, which equals zeros if two amino acids are identical, and one otherwise. Obviously, there are two points need to be improved in equation (1):

Table 2
The universal genetic code.

(1) GGG GGA GGU GGC / Glyline	(2) GAG GAA / Glutamic acid	GAU GAC / Aspartic acid
(3) GUG GUA GUU GUC / valine	(4) GCG GCA GCU GCC / Alanine	
(5) AGG AGA / Arginine	AGU AGC / Serine	(6) AAG AAA / Lysine
AAU AAC / Asparagine		
(7) AUG / Methionme	AUA AUU AUC / Isoleucine	(8) ACG ACA ACU ACC / Threonine
(9) UGG / Tryptophan	UGA / Stop	UGU UGC / Cysteine
(10) UAG UAA / Stop	UAU UAC / Tyrosine	
(11) UUG UUA / Leucine	UUU UUC / Phenylalanine	(12) UCG UCA UCU UCC / Serine
(13) CGG CGA CGU CGC / Arginine	(14) CAG CAA / Glutamine	CAU CAC / Histidine
(15) CUG CUA CUU CUC / Leucine	(16) CCG CCA CCU CCC / Proline	

(i) They implicitly suppose all the possible sense mutations (of a certain sense codon) occur in approximately equal frequency: $\frac{1}{\text{no. of neighburs}}$. In fact, just like discussed by Luo and Li [28], the mutation frequencies between nucleotides appeared in different codon positions are never uniform. They proposed the mutation frequency (MD) between a pair of triplets is expressed by four parameters : the transitional MD (MD caused by transitional mutation, $T \leftrightarrow C, G \leftrightarrow A$) is denoted by u, the transversional MD (MD caused by transversional mutation, $T \leftrightarrow A, T \leftrightarrow G, C \leftrightarrow G, C \leftrightarrow A$) is denoted by v, and wobble MD w_u, w_v that describe the additional effect of the third-letter mutation in a sense triplet (i.e., a triplet different from terminators, the latter are called non-sense triplets). Finally, they obtained $u = 2.2v$, $w_u = 8.1v$, $w_v = 3.7v$. Obviously, they also provide us with a more precise description of different mutation frequencies between different codons.

(ii) The Hamming distance between amino acids can only point out whether the mutation is a synonymous or a non-synonymous one. Generally speaking, from the physics–chemical view, the distances between different pairs of amino acids are never uniform, just like shown by Grantham [29]. Here, for completeness, we present the distance table in table 3.

Table 3
The distance between amino acids proposed by Grantham [29].

	Trp	Met	Glu	Asp	Lys	Asn	Gln	His	Cys	Tyr	Phe	Ile	Gly	Val	Ala	Thr	Pro	Leu	Arg
Ser	177	135	80	65	121	46	68	89	112	144	155	142	56	124	99	58	74	145	110
Arg	101	91	54	96	26	86	43	29	180	77	97	97	125	96	112	71	103	102	
Leu	61	15	138	172	107	153	113	99	198	36	22	5	138	32	96	32	98		
Pro	147	87	93	108	103	91	76	77	169	110	114	95	42	68	27	38			
Thr	128	81	65	85	78	65	42	47	149	92	103	89	59	69	58				
Ala	148	84	107	126	106	111	91	86	195	112	113	94	60	64					
Val	88	21	121	152	97	133	96	84	192	55	50	29	109						
Gly	184	127	98	94	127	80	87	98	159	147	153	135							
Ile	61	10	134	168	102	149	109	94	198	33	21								
Phe	40	28	140	177	102	158	116	100	205	22									
Tyr	37	36	122	160	85	143	99	83	194										
Cys	215	196	170	154	202	139	154	174											
His	115	87	40	81	32	68	24												
Gln	130	101	29	61	53	46													
Asn	174	142	42	23	94														
Lys	110	95	56	101															
Asp	181	160	45																
Glu	152	126																	
Met	67																		

Considering the (i) and (ii), we define the volatility of codon c by the new equation:

$$\text{volatility}_{\text{new}}(c) = \sum_{\text{neighbors } c_i} p[c, c_i] \times D'[\text{acid}(c), \text{acid}(c_i)], \tag{2}$$

where,

(1) $p[c, c_i]$ is v when sense codons c and c_i are connected by a transversional mutation in the first or second codon position.
(2) $p[c, c_i]$ is $2.2v$ when sense codons c and c_i are connected by a transitional mutation in the first or second codon position.
(3) $p[c, c_i]$ is $4.7v$ ($10.3v$) when sense codons c and c_i are connected by a transversional (transitional) mutation in the third codon position.
(4) $D'[\text{acid}(c), \text{acid}(c_i)]$ is the distance between amino acids encoded by codons c and c_i, as shown in table 3.

Consequently, we calculate the volatility of 61 sense codons by the two schemes, respectively, and list them in table 4. Note, the $p[c, c_i]$s of a certain sense codon c is normalized, so that, as to all sense codons c_is, the sum of $p[c, c_i]$s equals 1. So, in our scheme, each of the 61 volatilities is unique.

Observing table 4, we can find that, our scheme makes more amino acids have synonymous codons with different volatilities, and make the variance between different volatilities are larger than the original scheme; moreover, in our scheme, there are no codons of different amino acids have identical volatility any more. So, from some aspect, the new scheme should be a reasonable improvement of the original one.

3. 2D graphical representation and characterization of codons sequences

Now, let's apply the newly defined volatility to 2D graphical representation of codon sequences. Generally speaking, as to a 2D graphical model, the leading eigenvalue of its **L/L** matrix can be interpreted as a "degree of folding" of the geometrical structure. Similarly, when we transform a codons sequence into a 2D graph, where codons are represented by their "$\text{volatility}_{\text{new}}$", the leading eigenvalue of its **L/L** matrix can be interpreted as a "degree of folding" of such geometrical structure. Based on the ground: similar species should have similar codons sequences, hence similar such "degree of folding", so, we expect that it is rational to use the leading eigenvalues in similarity analysis of codons sequences here.

By labeling each codon c by corresponding $\text{volatility}_{\text{new}}(c)$ respectively, one can obtain a numerical sequence for a codon sequence. In table 5, we present the numerical sequences for the first 10 codons of the human beta-globin gene (Sequence 6 in table 1).

Table 4
The volatilities of 61 sense codons, calculated by equation (1) and (2), respectively.

	GGG	GGA	GGU	GGC	GAG	GAA	GAU	GAC
volatility	$\frac{6}{9}$	$\frac{5}{8}$	$\frac{6}{9}$	$\frac{6}{9}$	$\frac{7}{8}$	$\frac{7}{8}$	$\frac{8}{9}$	$\frac{8}{9}$
$\text{volatility}_{\text{new}}$	34.47	28.952	27.865	27.865	37.594	37.594	42.683	42.683
	GUG	GUA	GUU	GUC	GCG	GCA	GCU	GCC
volatility	$\frac{6}{9}$	$\frac{6}{9}$	$\frac{6}{9}$	$\frac{6}{9}$	$\frac{6}{9}$	$\frac{6}{9}$	$\frac{6}{9}$	$\frac{6}{9}$
$\text{volatility}_{\text{new}}$	17.117	17.744	19.488	19.488	19.979	19.979	20.655	20.655
	AGG	AGA	AGU	AGC	AAG	AAA	AAU	AAC
volatility	$\frac{7}{9}$	$\frac{6}{8}$	$\frac{8}{9}$	$\frac{8}{9}$	$\frac{7}{8}$	$\frac{7}{8}$	$\frac{8}{9}$	$\frac{8}{9}$
$\text{volatility}_{\text{new}}$	58.014	56.613	59.801	59.801	47.601	47.86	51.972	51.972
	AUG	AUA	AUU	AUC	ACG	ACA	ACU	ACC
volatility	$\frac{9}{9}$	$\frac{7}{9}$	$\frac{7}{9}$	$\frac{7}{9}$	$\frac{6}{9}$	$\frac{6}{9}$	$\frac{6}{9}$	$\frac{6}{9}$
$\text{volatility}_{\text{new}}$	22.683	20.342	22.192	22.192	19.601	20.228	19.302	19.302
	UGG		UGU	UGC			UAU	UAC
volatility	$\frac{7}{7}$		$\frac{7}{8}$	$\frac{7}{8}$			$\frac{6}{7}$	$\frac{6}{7}$
$\text{volatility}_{\text{new}}$	177.53		103.47	103.47			57.668	57.668
	UUG	UUA	UUU	UUC	UCG	UCA	UCU	UCC
volatility	$\frac{6}{8}$	$\frac{5}{7}$	$\frac{8}{9}$	$\frac{8}{9}$	$\frac{5}{8}$	$\frac{4}{7}$	$\frac{6}{9}$	$\frac{6}{9}$
$\text{volatility}_{\text{new}}$	23.387	21.563	31.822	31.822	30.103	24.475	32.626	32.626
	CGG	CGA	CGU	CGC	CAG	CAA	CAU	CAC
volatility	$\frac{5}{9}$	$\frac{4}{8}$	$\frac{6}{9}$	$\frac{6}{9}$	$\frac{7}{8}$	$\frac{7}{8}$	$\frac{8}{9}$	$\frac{8}{9}$
$\text{volatility}_{\text{new}}$	23.096	16.394	32.021	32.021	22.826	22.826	28.363	28.363
	CUG	CUA	CUU	CUC	CCG	CCA	CCU	CCC
volatility	$\frac{5}{9}$	$\frac{5}{9}$	$\frac{6}{9}$	$\frac{6}{9}$	$\frac{6}{9}$	$\frac{6}{9}$	$\frac{6}{9}$	$\frac{6}{9}$
$\text{volatility}_{\text{new}}$	16.996	16.641	17.865	17.865	22.149	22.149	22.185	22.185

For a given codon sequence with n codons, in a 2D space, we draw n points, which coordinates are $(i, \text{volatility}_{\text{new}}(c_i))$ respectively, where $i \in 1, 2, \ldots n$; $\text{volatility}_{\text{new}}(c_i)$ is the volatility of the ith codon appeared in the sequence. Then we connect dots in pairs, whose abscissae are consecutive integers, and obtain a zigzag-like curve.

Figure 1 shows the 2D graphical representation of the segment consisting of the first 10 codons of the human beta-globin gene based on table 5. Therefore, a

Table 5
The numerical sequences for the first 10 codons of the human beta-globin gene.

ATG	GTG	CAC	CTG	ACT	CCT	GAG	GAG	AAG	TCT	...
22.683	17.117	28.363	16.996	19.302	22.185	37.594	37.594	47.601	32.626	...

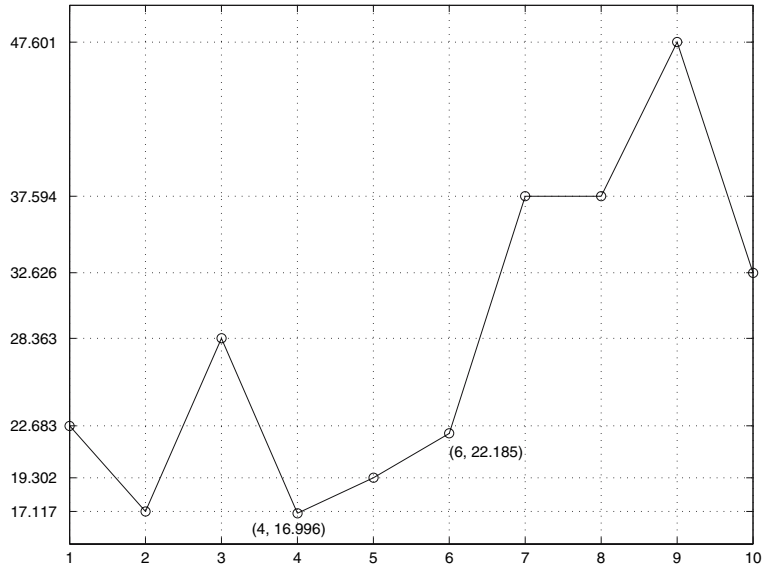

Figure 1. The 2D graphical representation of the first 10 codons of the human beta-globin gene based on table 5.

codon sequence can be represented uniquely by such a 2D graph corresponding to its number sequence.

For each codon sequence, grounded on such a graphical representation, we construct its ${}^{inf}L/{}^{inf}L$ matrix (which is the limit of the matrices sequences ${}^{k}L/{}^{k}L$ as k trends to infinity, just like we did in ref. [23]), and use its leading eigenvalue to characterize the codon sequence. We list the leading eigenvalues

Table 6
The leading eigenvalues of the ${}^{inf}L/{}^{inf}L$ matrices associated with 2D graphs for the codons sequences of table 1.

Species	Leading eigenvalues from equation 2	Leading eigenvalues from eqution 1
Bovine	2.84988800043170	*10.35898678169115*
Chimpanzee	2.83451262631253	*11.06215782470777*
Gallus	3.17327758390017	*9.17966485302680*
Goat	3.13230899487733	*10.35898678169115*
Gorilla	2.75722177609269	*9.67296265292199*
Human	2.75722177608345	*9.47158965042611*
Lemur	2.37066429502053	*6.88191362501445*
mouse	2.91995402130494	*6.68049730428715*
Opossum	3.35995403696471	*12.27634546752876*
Rabbit	2.54741640253734	*9.71587038893165*
Rat	2.53519141267502	*6.48900527609151*

corresponding to the 11 codon sequences (shown in table 1) on the left side of table 6. For comparison, on the right side of table 6, we list the other 11 leading eigenvalues, which is derived from the equation 1. As can be seen, our scheme is better in characterizing codon sequences than the other one, for example, the other one even fails to distinguish gallus (the only non-mammal animal here) from mammals.

Moreover, the leading eigenvalues on the left side can distinguish the 11 species clearly. The gallus and Opossum have the largest values and far differ from other species, for they are non-mammal and pouched animals, respectively. The values of three primates: chimpanzee, gorilla and human are very close. This might mean the leading eigenvalues play a special role in the classification of non-pouched mammal, pouched mammal and non-mammal. Certainly, this group of numerical results may provide biologists with some useful information on the chemical structure of codon sequences of different species.

4. Similarities and dissimilarities analysis with a single variable

Once bio-sequences are represented by a single variable, naturally, the similarity between them can be described by the "differences" between the values of the variation. Obviously, the similarity/dissimilarity table can be got easily, and in fact, there are no intrinsic differences from other literatures, so, for briefness, we only list the data comparing human with other species in table 7.

Table 7
The degree of dissimilarity of the coding sequence of human with those of other 10 species.

Species	Bovine	Chimpanzee	Gallus	Goat	Gorilla
Dissimilarity	0.092666	0.077291	0.41606	0.37509	9.2326×10^{-12}
Species	Lemur	Mouse	Opossum	Rabbit	Rat
Dissimilarity	0.38656	0.16273	0.60273	0.20981	0.22203

Table 8
The degree of similarity of the coding sequences of several species with the coding sequence of human, the data were normalized.

Species	Opossum	Gallus	Lemur	Goat	Rat	Rabbit
This work	0.60273	0.41606	0.38656	0.37509	0.22203	0.20981
From ref. [14, table 3]	0.148	0.109	0.087	0.061	0.043	0.042
From ref. [16, table 11]	0.0509	0.0475	0.0463	0.0367	0.0327	0.025
From ref. [16, table 12]	4.491	5.015	2.970	4.996	4.857	3.171
From ref. [18, table 9]	16.2481	17.3205	15.748	17.4356	14.3875	8.77496
From ref. [19, table 10]	0.164054	0.172133	0.160247	0.17295	0.142292	0.0905605

Additionally, we list some results of the examinations of the degree of similarity of human and other several species in table 8. As one can see there exists an overall agreement among similarities obtained by different approaches, especially, our result is consistent with that in ref. [14, table 3] and from ref. [16, table 11].

5. Conclusion and discussion

In this paper, we have outlined a new method measuring the volatilities of codons grounded on different mutation frequencies between codons shown in ref. [28] and the physics–chemical distances between amino acids shown in ref. [29]. As can be seen, the 2D graph presented here is easy constructed, and thoroughly avoids self intersection. The use of a single variable (i.e., leading eigenvalues) in characterizing biological sequences can simplify computation required in similarity analysis. Moreover, as observed from table 6, the new-defined volatility is superior to the former in characterizing codon sequences.

In fact, Stoletzki et al. [30] have summarized that, many scientists think volatility defined by Plotkin et al. [27] is unlikely to measure selection, the first two reasons are it only depends on four or five amino acids and it has low variance. In this sense, our scheme seems to improve the index to a certain extent. For example, in our new scheme, 10 amino acids have synonymous codons with different volatilities, codons coding for different amino acids always have different volatilities, and the variance between different volatilities is larger than the former. Of course, the effectiveness of our new index in other parts of biology, e.g. inferring the level of natural selection on DNA sequences, still needs further study.

Acknowledgments

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

References

[1] R. Staden, Nucleic Acids Res. 14 (1986) 217–231.
[2] E. Hamori and J. Ruskin, J. Biol. Chem. 258 (1983) 1318–1327.
[3] E. Hamori, BioTechniques 7 (1989) 710–720.
[4] M. Randic and A.T. Balaban, J. Chem. Inf. Comput. Sci. 43 (2003) 532–539.
[5] X.F. Guo, M. Randic and S.C. Basak, Chem. Phys. Lett. 350 (2001) 106–112.
[6] R. Zhang and C.T. Zhang, J. Biomo. Struct. Dyn. 11 (1994) 767–782.
[7] M.A. Gates, J. Theor. Biol. 119 (1986) 319–328.
[8] A. Nandy, Curr. Sci. 66 (1994) 309–314.
[9] A. Nandy, Curr. Sci. 66 (1994) 821.

[10] P.M. Leong and S. Mogenthaler, Comput. Appl. Biosci. 12 (1995) 503–511.
[11] X. Guo and A. Nandy, Chem. Phys. Lett. 369 (2003) 361–366.
[12] Y. Wu, A.W. Liew, H. Yan and M.S. Yang, Chem. Phys. Lett. 367 (2003) 170–176.
[13] A. Nandy and P. Nandy, Chem. Phys. Lett. 368 (2003) 102–107.
[14] M. Randic, M. Vracko, N. Lers and D. Plavsic, Chem. Phys. Lett. 368 (2003) 1–6.
[15] M. Randic, M. Vracko, N. Lers and D. Plavsic, Chem. Phys. Lett. 371 (2003) 202–207.
[16] M. Randic, Chem. Phys. Lett. 317 (2000) 29–34.
[17] M. Randic, X.F. Guo and S.C. Basak, J. Chem. Inf. Comput. Sci. 41 (2001) 619–626.
[18] P. He and J. Wang, Internet Electron. J. Mol. Des. 1 (2002) 668–674.
[19] P. He and J. Wang, J. Chem. Inf. Comput. Sci. 42 (2002) 1080–1085.
[20] Y. Liu, Internet Electron. J. Mol. Des. 1 (2002) 675–684.
[21] M. Randic and G. Krilov, Int. J. Quantum. Chem. 75 (1999) 1017–1026.
[22] C. Li and J. Wang, Comb. Chem. High Throughput Screen. 6 (2003) 795–799.
[23] J. Wang and Y. Zhang, Chem. Phys. Lett. 423 (2006) 50–53.
[24] M. Randic, SAR QSAR Environ. Res. 15 (3) (2004) 147–157.
[25] M. Randic and J. Zupan, SAR QSAR Environ. Res. 15 (3) (2004) 191–205.
[26] M. Randic, J. Zupan and A.T. Balaban, Chem. Phys. Lett. 397 (2004) 247–252.
[27] J.B. Plotkin, J. Dushoff and H.B. Fraser, Nature 428 (2004) 942–945.
[28] L. Luo and X. Li, Biosystem 65 (2002) 83–97.
[29] R.A. Grantham, Sceince 185 (1974) 862–864.
[30] N. stoletzki, J. Welch, J. Hermisson and A.E. Walker, Mol. Biol. Evol. 22(10), (2005) 2022–2026.