J Math Chem (2010) 48:521–529
DOI 10.1007/s10910-010-9688-0

ORIGINAL PAPER

Numerical characterization of DNA sequences: connectivity type indices derived from DNA line graphs

R. Natarajan · R. Jayalakshmi · M. Vivekanandan

Received: 20 April 2009 / Accepted: 22 April 2010 / Published online: 7 May 2010

Abstract The four-letter code sequence of a single strand of a DNA sequence was converted into a line graph, and the vertices of the line graph were assigned weights according to the dissociation constant (pK_a) of the corresponding nitrogenous base represented by each of the vertices. Connectivity type indices were computed for the weighted line graphs and the numerical descriptors thus calculated were used for alignment-free sequence comparison. The numerical descriptors proposed in this study were calculated very fast even for whole genomes, and thus, the methodology enabled alignment-free comparison of long DNA sequences without much computational load. Sequence comparison using numerical descriptors derived from the weighted line graphs is illustrated using 23 mitochondrial genomic sequences. The cladogram obtained from the hierarchical clustering carried out using the numerical descriptors grouped evolutionarily similar sequences together.

Keywords Numerical characterization · DNA · Connectivity index ·
Sequence comparison

R. Natarajan (✉)
Center for Mathematical Sciences, Arunapuram, Pala, Kerala 686574, India
e-mail: rnataraj@lakeheadu.ca

R. Natarajan
Department of Chemical Engineering, Lakehead University, Thunder Bay, Ontario P7B 5E1, Canada

R. Jayalakshmi
Department of Biotechnology, Bharathidasan University, Tiruchirappalli, Tamil Nadu 620024, India

M. Vivekanandan
Vivekananda College of Arts and Science for Women, Thiruchengode, Namakal District,
Tamil Nadu 637205, India

1 Introduction

Traditionally evolutionary relationships among organisms had been inferred using morphological characteristics. In the last few years scientists have switched more to genomic data for studying evolutionary models (phylogeny) owing to the availability of enormous amount of sequence data. Evolutionary models obtained using DNA sequence data are often referred to as substitution models because mutation is considered as a substitution of one nucleotide for another at a particular site in a DNA sequence. Genes of similar type, homologous genes, are normally used for comparison of similarity among the sequences. Sequences of homologous genes from various organisms are often unequal in length, and therefore, correspondences among sequences positions are not evident. In order to make the alignment perfect, the local and global dynamic programming algorithms use insertion of gaps, whereas multiple sequence analysis uses insertion of gaps and deletion of residues. A matrix constructed based on the aligned columns is then used for phylogenetic analysis. Several sequence alignment algorithms were developed [1], and the computational load of such sequence alignments increases with increase in length of the sequences. In order to reduce intense computation, bioinformatics tools such as BLAST [2] and FASTA [3] were developed based on heuristic approach.

Graphical representation methods were introduced by Hamori for visual comparison of DNA sequences [4,5] which was followed by several modifications and additional graphical representation methods were introduced [6,7]. In graphical representation methods, a DNA sequence is plotted, usually, as a random walk on a Euclidean plane. The idea behind it was to read the DNA sequence base by base and plot succeeding points on the graph. These rectangular walks had the inherent limitation that sequence consisting of bases that alternated between two types along one axis caused overlapping path in one or other of these representations. Thus, a repetitive sequence showed up only one step along any one of the axes and led to loss of information. Moreover, there were chances of two sequences leading to identical plots. Though attempts were made [8] to remove degeneracy in graphical representations of DNA sequences, it was difficult to visualize the lengthier sequences and visual comparison of large sequences was not possible. As visual comparison introduced obscurity, Randić introduced [9–11] numerical characterization of the DNA graphs and several authors [12–18] followed this approach to suggest different ways of numerical characterization of DNA sequences. Most of these approaches on numerical characterization of DNA graphs were reviewed by Nandy et al. [19]

Numerical descriptors were also derived directly from the primary sequences using information theory [20,21]. In the information theoretic approach, a DNA sequence was considered as a linear sequence of n symbols from a finite set of four alphabets. Probability distributions of combinations of (segment of) L symbols (*L-tuple*) were computed. Information theory based measures were used on DNA sequences from large databases and the approach was also extended to protein sequences [21]. All these alignment-free sequence comparison methods treat biological sequences as a string of letters and the biological molecules are reduced to linguistic representations without any consideration for physicochemical properties, 3-D structure and long range interactions. In this paper, numerical characterization of DNA sequences using one of the physicochemical properties of nitrogenous bases namely, dissociation constants (pK_a) is proposed as a first step in deriving DNA descriptors that encode more information. Application of the new descriptors in alignment-free sequence comparison was tested using 23 mitochondrial genomic sequences.

2 Conversion of a DNA sequence into a line graph

A line graph *L*(*G*) is obtained by converting each edge in a graph *G* into a vertex. In a single strand of DNA the phosphodiester groups and the deoxyribose units remain constant and the variation occurs only due to difference in nitrogenous base (see Fig. 1a), and therefore, DNA sequences are represented by the four-letter code. A DNA sequence was converted into a line graph in which each nitrogenous base represented a vertex while the phosphate and sugar units were suppressed. Each vertex was then assigned a vertex weight based on the dissociation constant (pK_a) of the base corresponding to the vertex (Fig. 1b). Dissociation constants (pK_a) of the DNA bases were taken from the internet source www.sciencecollege.co.uk/SC/biochemicals.html. The motivation for using a line graph was from the line distance matrix representation proposed by Randić [22].

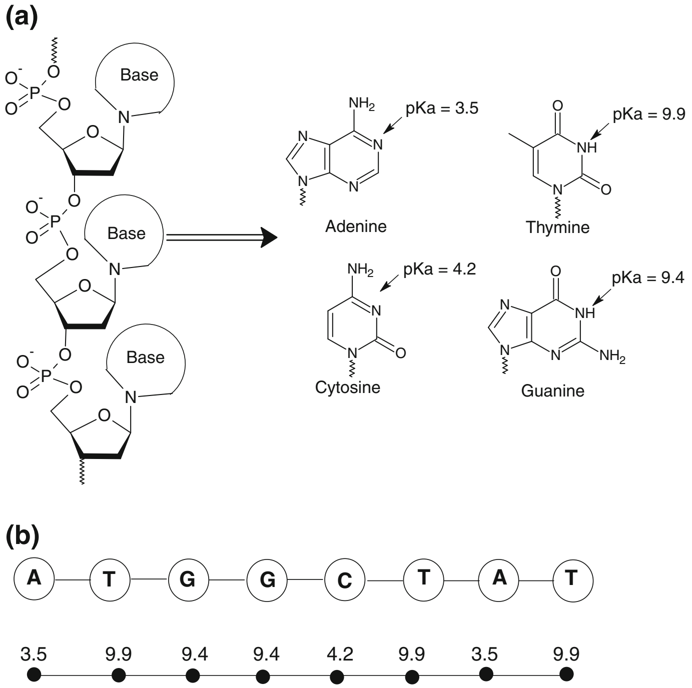

Fig. 1 **a** pK_a values of the four bases **b** conversion of a DNA sequence into a line graph using pK_a values of the four bases as the vertex weights

3 Connectivity indices for the DNA line graphs

In 1975 Randić [23] proposed molecular connectivity index to characterize branching in alkanes. The Randić connectivity index is calculated from the degrees δ using the relation given below:

$$^{1}\chi = \sum \frac{1}{\sqrt{\delta_i \delta_j}} \tag{1}$$

where *i* and *j* are the pairs of non-hydrogen atoms connected by a bond (edge) and the summation is over all the bonds in a molecule, and degree δ of a vertex is the number of edges incident on the vertex. Kier et al. developed [24] a generalized connectivity index $^{h}\chi$ considering paths of type $v_0, v_1, \ldots v_h$ of length *h* in the molecular graph. In the case of weighted graphs, vertices may be assigned weights based on several schemes such as bond-order, valency, etc. A generalized connectivity index $^{h}\chi$ of length *h* can be calculated from the equation.

$$^{h}\chi = \frac{1}{\sqrt{\delta_i \delta_j \ldots \delta_h}} \tag{2}$$

The connectivity indices are denoted as $^{h}\chi, ^{h}\chi^{v}$or $^{h}\chi^{b}$ to differentiate simple, valency, and bond-order based path connectivity, respectively. The new DNA descriptors proposed in this paper were calculated by extending the calculation of connectivity indices for molecular graphs to the DNA line graphs. The connectivity indices for a DNA sequence were calculated based on $\mathrm{pK_a}$ values of each of the four bases. Hence, the notation $^{h}\chi^{pKa}$ is suggested for the new set of descriptors proposed in this paper. Calculations of $^{h}\chi^{pKa}$ for the hypothetical DNA sequence shown in Fig. 1b are illustrated below:

$$\begin{aligned}
{}^{1}\chi^{\mathrm{pKa}} &= \frac{1}{\sqrt{3.5\times 9.9}} + \frac{1}{\sqrt{9.9\times 9.4}} + \frac{1}{\sqrt{9.4\times 9.4}} + \frac{1}{\sqrt{9.4\times 4.2}} + \frac{1}{\sqrt{4.2\times 9.9}} \\
&\quad + \frac{1}{\sqrt{9.9\times 3.5}} + \frac{1}{\sqrt{3.5\times 9.9}} \\
&= 1.0339 \\
{}^{2}\chi^{\mathrm{pKa}} &= \frac{1}{\sqrt{3.5\times 9.9.\times 9.4}} + \frac{1}{\sqrt{9.9\times 9.4\times 9.4}} + \frac{1}{\sqrt{9.4\times 9.4\times 4.2}} \\
&\quad + \frac{1}{\sqrt{9.4\times 4.2\times 9.9}} + \frac{1}{\sqrt{4.2\times 9.9\times 3.5}} + \frac{1}{\sqrt{4.2\times 3.5\times 9.9}} \\
&= 0.32860 \\
{}^{7}\chi^{\mathrm{pKa}} &= \frac{1}{\sqrt{3.5\times 9.9\times 9.4\times 9.4\times 4.2\times 9.9\times 3.5\times 9.9}} \\
&= 4.76\times 10^{-4}
\end{aligned}$$

Connectivity type indices for DNA sequences were previously proposed by Zhang et al. [25], but the basis of calculating the $\chi-$indices in the present paper is entirely different and not reported earlier.

4 Application of ${}^{h}\chi^{pKa}$ in sequence comparison

Twenty three genomic sequences of mitochondrial DNA were downloaded from the GenBank database using Entrez data-retrieval tool (http://www.ncbi.nlm.nih.gov/Entrez/). Table 1 gives the accession number, common names, and lengths of the sequences used in the present study. Connectivity-based indices of order zero to ten (${}^{h}\chi^{pKa}$ for $h = 0$ to 10) were calculated for each of the sequences using an in-house computer program developed in Visual Basic 6. The program took less than a minute per sequence for calculating the indices using a PC with Intel Core2DUO (E4500) 2.20 MHz processor and 1 GB RAM. The connectivity-based descriptors calculated for the 23 mitochondrial genomic sequences are given in Table 2. Ratio of (A+T) to

Table 1 Sequences used in the study

#	Accession no.	Species name	Common name	Seq. length
1	EU352212	*Aedes aegypti*	Mosquito_AA	16655
2	L20934	*Anopheles gambiae*	Mosquito_AG	15363
3	L06178	*Apis mellifera*	Honey Bee	16343
4	AF149768	*Bombyx mori*	Silk worm	15643
5	AF538716	*Brugia malayi*	Round worm	13657
6	AC186293	*Caenorhabditis briggsae*	*C.briggasae*	14420
7	X54252	*Caenorhabditis elegans*	*C.elegans*	13794
8	U37541	*Drosophila melanogaster*	Fruit fly	19517
9	AJ276844	*Plasmodium falciparum*	*P.falciparum*	5967
10	AJ312413	*Tribolium castaneum*	Beetle	15881
11	AY526085	*Bos taurus*	Cattle	16338
12	U96639	*Canis familiaris*	Dog	16727
13	AF010406	*Ovis aries*	Sheep	16616
14	D38113	*Pan troglodytes*	Chimpanzee	16554
15	NC_001807	*Homo sapiens*	Human	16571
16	AY612638	*Macaca mulatta*	Rhesus Monkey	16564
17	NC_005089	*Mus musculus*	Mouse	16299
18	X14848	*Rattus norvegicus*	Rat	16300
19	X52392	*Gallus gallus*	Chicken	16775
20	M10217	*Xenopus laevis*	Frog	17553
21	AJ508398	*Monodelphis domestica*	Opossum	17079
22	X83427	*Ornithorhynchus anatinus*	Platypus	17019
23	AJ001588	*Oryctolagus cuniculus*	Rabbit	17245

Table 2 Connectivity indices of order zero to ten for the DNA line graphs of the genomic sequences used in the study

Common name	${}^{0}\chi^{pKa}$	${}^{1}\chi^{pKa}$	${}^{2}\chi^{pKa}$	${}^{3}\chi^{pKa}$	${}^{4}\chi^{pKa}$	${}^{5}\chi^{pKa}$	${}^{6}\chi^{pKa}$	${}^{7}\chi^{pKa}$	${}^{8}\chi^{pKa}$	${}^{9}\chi^{pKa}$	${}^{10}\chi^{pKa}$	AT/GC
Mosquito_AA	7114.887	3044.922	1300.943	556.269	238.351	102.213	43.920	18.884	8.130	3.504	1.510	3.761
Mosquito_AG	6565.948	2810.316	1200.639	513.646	220.073	94.397	40.559	17.442	7.515	3.243	1.400	3.457
Honey Bee	6998.930	3001.254	1283.924	548.483	234.787	100.558	43.160	18.524	7.955	3.418	1.468	5.603
Silk worm	6742.813	2912.416	1257.016	543.304	235.491	102.292	44.561	19.423	8.470	3.701	1.618	4.354
Round worm	5178.230	1969.637	750.238	286.530	109.580	41.977	16.124	6.204	2.389	0.921	0.356	3.075
C.briggasae	6407.595	2853.430	1272.520	567.526	252.771	112.524	50.153	22.341	9.958	4.442	1.982	3.090
C.elegans	5548.976	2235.966	902.597	365.042	147.420	59.479	24.016	9.699	3.922	1.588	0.642	3.205
Fruit fly	8322.570	3553.384	1517.963	648.787	278.097	119.129	51.042	21.840	9.346	3.999	1.712	4.605
P.falciparum	2482.480	1032.154	429.172	178.384	74.125	30.827	12.826	5.334	2.219	0.922	0.384	2.166
Beetle	6927.712	3027.890	1319.458	576.303	252.013	110.277	48.357	21.176	9.271	4.063	1.780	2.531
Cattle	7114.477	3094.260	1345.300	587.085	256.138	111.730	48.899	21.381	9.336	4.088	1.787	1.537
Dog	7208.222	3101.797	1333.109	574.939	247.786	106.846	46.182	19.946	8.610	3.723	1.608	1.522
Sheep	7241.021	3152.334	1370.979	598.381	261.193	114.069	49.963	21.863	9.556	4.191	1.836	1.568
Chimpanzee	7263.371	3184.531	1395.425	613.344	269.298	118.256	52.088	22.919	10.081	4.445	1.958	1.289
Human	7276.104	3193.369	1401.289	617.109	271.571	119.484	52.736	23.249	10.244	4.525	1.998	1.248
Rhesus Monkey	7277.588	3196.798	1404.670	619.507	273.182	120.440	53.261	23.546	10.400	4.607	2.039	1.313
Mouse	7093.204	3084.867	1340.621	584.712	255.055	111.316	48.752	21.328	9.325	4.088	1.792	1.721
Rat	7128.181	3115.433	1360.132	596.126	261.189	114.425	50.320	22.108	9.703	4.270	1.877	1.584
Chicken	7376.295	3242.674	1424.651	628.572	277.117	122.118	54.038	23.891	10.554	4.677	2.071	1.176
Frog	7557.220	3255.444	1401.817	605.519	261.758	113.118	48.967	21.189	9.169	3.969	1.717	1.704
Opossum	7348.737	3159.808	1356.801	584.963	252.232	108.737	46.954	20.278	8.748	3.781	1.635	1.260
Platypus	7267.995	3104.922	1326.661	568.587	243.695	104.384	44.785	19.201	8.227	3.526	1.510	1.857
Rabbit	7456.772	3221.222	1390.846	602.088	260.665	112.880	49.018	21.254	9.215	4.002	1.737	1.693

(G+C) i.e., (sum of ‘A’s and ‘T’s) ÷ (sum of ‘G’s and ‘C’s) was also included as one of the numerical descriptors.

Comparison of similarity among sequences using a diverse pool of numerical descriptors enables clustering in a multidimensional space and this is expected to yield a much better result than one-dimensional analyses using a single descriptor. However, numerical descriptors that are highly inter-correlated encode redundant information and appropriate statistical tools should be used to extract mutually orthogonal descriptors. Principal Component Analysis (PCA) is normally used for data reduction and extraction of orthogonal parameters. PCA yields a set of eigenvalues and the eigenvectors where the elements of eigenvectors can be interpreted as correlation indices and they reflect the degree of association between the ith variable and the jth principal component. The objective of the interpretation is to select one variable to represent each eigenvector; this subset of variables will have low inter-correlation because the eigenvectors are uncorrelated. Hopefully, the variables will be selected with the thought of maximizing the variation between the predictor variables selected as the subset and criterion variable. Although the correlation between the predictor variables in the subset and the criterion variables cannot be greater than the correlation between all of the predictor variables and the criterion variable, the difference between the two correlations should not be statistically significant. Thus, principal component analysis (PCA) should yield a subset of predictor variables that reduces both the data collection and the inter-correlation.

The ${}^{h}\chi$-values in Table 2 for different orders ($h = 0$, 1 etc) differ in their magnitudes and this was expected to effect the results of PCA because PCA is scale dependant. Hence, the ${}^{h}\chi$-values were normalized using the following procedure:

$$
{}^{h}\chi_{normalized} = \frac{{}^{h}\chi}{n-h} \tag{3}
$$

where n is the sequence length and h is the order of connectivity type descriptor. The normalized ${}^{h}\chi$-values values, and AT/GC ratio were then scaled using the transformation log_e (*variable +3*) and thus all the descriptors were brought to same orders of magnitudes. PCA of the data matrix containing 23 observations and 12 columns extracted two factors and they accounted for 99% of total data variance. The component matrix indicated all χ-descriptors were highly correlated to the first principal component (PC-1) while AT/GC ratio was highly correlated to the second principal component (PC-2). Amongst the χ-descriptors, ${}^{5}\chi^{pK}$a and ${}^{6}\chi^{pKa}$ were almost perfectly correlated with PC1. Hence, ${}^{6}\chi^{pKa}$ and AT/GC were selected from the initial descriptors set. The two selected descriptors had very low correlation between them ($r = -0.278$) and were used to study the similarity of sequences. Hierarchical cluster analysis was carried out using SPSS software and Euclidean distance was used as a measure of similarity. The dendrogram was drawn using linkage within group method and Euclidean distance as the measure of similarity in the 2-D space. The dendrogram (phylogenetic analysis) obtained for 23 genomic sequences is given in Fig. 2. The alignment-free approach followed in this paper was found to cluster similar sequences together. For example almost all invertebrates (arthropods and

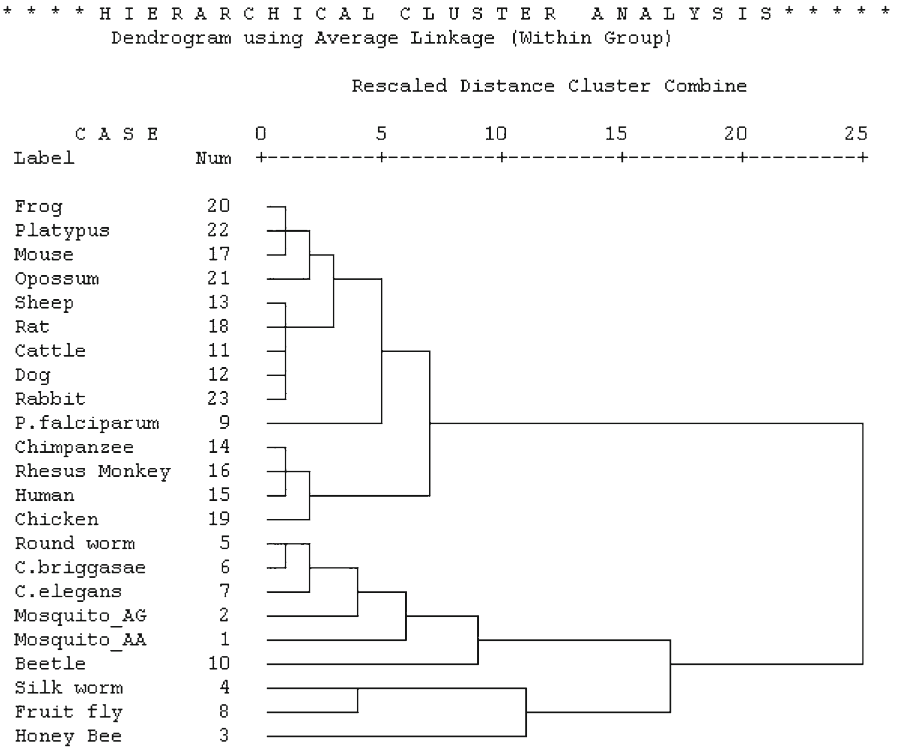

Fig. 2 Phylogenetic analysis of twenty three genomic sequences of mitochondrial DNA

nematodes) were diverged from vertebrates by grouping the insects in one close cluster and worms in another. Duck billed platypus is a monotreme and retains mixture of mammalian and reptilian features was paired with frog and mouse. This indicated that these mitochondrial genomic sequences might share some commonality in their gene families. Among the vertebrates, mammals and primates were separated (primates in same node), and distinguished from the rest of the species. It is quiet satisfying to note that the approach used in this study was able to identify similarity among the sequences. Moreover, the approach advocated in this paper was utilized to compare genomic sequences without much demand for large computation time as opposed to very large computation time required in sequence alignment methods. There is a wide scope to extend the calculation of the χ-indices for DNA graphs using other physicochemical properties and solvent perturbation parameters. It is also possible to assign weights to the edges based upon other interactions. Though, such secondary structural characteristics are not necessary for sequence comparison they will be useful while considering the numerical descriptors as biodescriptors. However, extension of the approach to protein sequences and incorporating secondary structural information will result in better numerical characterization.

5 Availability of computer program

The computer program for the calculation of the DNA descriptors explained in this paper may be obtained free of cost from the corresponding author (R.N.).

Acknowledgments One of the authors (R.N.) acknowledges the Department of Science and Technology, New Delhi, India for financial assistance under Project No. SR/S4/MS: 479/07. This contribution is paper No. 61 from the Centre for Mathematical Sciences Pala Campus, Kerala, India. Our thanks are due to Mr. T. M. Anbazhagan, Inflexion technology, Bangalore, India for his help in developing the computer program to calculate the descriptors.

References

1. S.F. Altschul, W. Gish, W. Miller, E.W. Myers, D.J. Lipman, J. Mol. Biol. **215**, 403 (1990)
2. S.F. Altschul, T.L. Madden, A.A. Schaffer, J. Zhang, Z. Zhang, W. Miller, D.J. Lipman, Nucleic Acids Res. **25**, 3389 (1997)
3. S. Henicoff, J.G. Henicoff, Proc. Natl. Acad. Sci. USA **89**, 10915 (1992)
4. E. Hamori, J. Ruskin, J. Biol. Sci. **258**, 1318 (1983)
5. E. Hamori, Nature **314**, 585 (1985)
6. M.A. Gates, J. Theor. Biol. **119**, 319 (1986)
7. A. Nandy, Curr. Sci. **66**, 309 (1994)
8. S.S.T. Yau, A. Niknejad, C. Lu, N. Jin, Y. Ho, Nucleic Acids Res. **31**, 3078 (2003)
9. M. Randić, M. Novic, D. Vikic-Topic, D. Plavsic, SAR QSAR Environ. Res. **17**, 583 (2006)
10. M. Randić, M. Vraćko, N. Lers, D. Plavsic, Chem. Phy. Lett. **368**, 1 (2002)
11. M. Randić, M. Vraćko, A. Nandy, S.C. Basak, J. Chem. Inf. Comput. Sci. **40**, 1235 (2000)
12. F.l. Bai, Y.z. Liu, T.M. Wang, Math. Biosci. **209**, 282 (2007)
13. Y. Chunzin, B. Liao, T.M. Wang, Chem. Phys. Lett. **379**, 412 (2003)
14. C. Li, J. Wang, Comb. Chem. High Throughput Screen **6**, 795 (2003)
15. B. Liao, T.M. Wang, J. Mol. Struct. THEOCHEM. **681**, 209 (2004)
16. Z.H. Qi, T.R. Fan, Chem. Phys. Lett. **442**, 434 (2007)
17. J. Song, H. Tang, J. Biochem. Bioph. Methods **63**, 228 (2005)
18. J. Song, J. Biol. Syst. **15**, 287 (2007)
19. A. Nandy, M. Harle, S.C. Basak, ARKIVOC **(ix)**, 211 (2006)
20. B.E. Blaisdell, Proc. Natl. Acad. Sci. **83**, 5155 (1986)
21. S. Vinga, J. Almeida, Bioinformatics **19**, 513 (2003)
22. M. Randić, J. Zupan, T. Pisanski, J. Math. Chem. **43**, 674 (2008). doi:10.1007/s10910-006-9219-1
23. M. Randić, J. Am. Chem. Soc. **97**, 6609 (1975)
24. L.B. Kier, W.J. Murray, M. Randić, L.H. Hall, J. Pharm. Sci. **65**, 1226 (1976)
25. B.H. Zhang, H.S. Wang, L. Xu, Chemom. Intell. Lab. Syst. **87**, 194 (2007)