Condensed Matrix: A Tool to Characterize DNA

Nadia L. Helal

Atomic Energy Authority, National Center for Nuclear Safety and Radiation Control. 3, Ahmed El-Zomer st., Nasr City 11762, P.O.Box 7551, Cairo-Egypt

Abstract

Motivation. This paper reports the development of new method for mathematical characterization of the primary DNA sequences.

Method. A condensed characterization of the primary sequence is based on 4*4 matrices the rows and columns of which are associated with the four nucleic bases A, G, C and T.

Results. The condensed matrices for the primary sequences of DNA is serving as a source of invariants that allow quantitative comparisons of DNA from different sources.

Conclusion. The sensitivity of the descriptor renders it suitable for using it as a parameter to index toxicity levels of various agents that induce changes in DNAs. The study was outlined on normal DNA.

Keywords. DNA sequence; DNA descriptors; condensed matrix; topological indices.

1 INTRODUCTION

DNA is usually presumed to be the critical macromolecular target for carcinogenesis and mutagenesis [1]. To predict sequence changes induced by different agents, it imperative to have quantitative measures to compare and contrast the different DNA sequences [2]. Earlier studies have shown different schemes to characterize DNA sequences so that members of different gene families and sequences can be described by unique numbers that have the potential to codify the DNA sequence into a set of numbers for quantitative comparisons between different species. This method is very useful in numerical characterization of DNA sequence, and is capable of handling large sequences with reasonable degree of accuracy, providing quantitative estimates to base alterations, deletions and additions for ready comparison and tabulation [3-6].

The method is based firstly on a DNA representative with a suitable mathematical object. Secondly, the selected mathematical object is described by various matrices that record distances among DNA sequences. Thirdly, one constracts various matrix invariants to serve as DNA descriptors. Comparison between sequences becomes thus comparison between matrices.

2 REDUCED DNA MATRICES

A direct base-by-base transformation of a primary DNA sequence to a matrix will result in a matrix having many rows and columns. For example the first exon of the primary DNA (shown in table 1), which has 50 nucleic bases leads to a symmetric 50*50 matrix with 1585 matrix entries. In case of human genes, the first exon is of longer length and would generate a symmetric matrix with over one million entries. Consider the beginning of the first exon of table 1:

T	G	G	A	A	T	T	G	T	G
A	G	C	G	G	A	T	A	A	C
A	A	T	T	T	C	A	C	A	C
A	G	G	A	A	A	C	A	G	C
T	A	T	G	A	C	C	A	T	G

The first exon of table 1 has 50 bases totaling 11T+ 12G+ 18A+ 9C. Thus the 50*50 DNA matrix will lead to diagonal submatrices of the following size: 11*11, 12*12, 18*18 and 9*9 corresponding to TT, GG, AA and CC respectively. All the off diagonal submatrices will be in this case rectangular, TG of size 11*12, TA of 11*18 and TC of 11*9. In our study, the four nucleic bases

A, G, C and T are considered as labels for a 4*4 symmetric matrix (with xy=yx). One can summarize pertinent information in a very condensed 4*4 matrix of the following form

	A	G	C	T				
A	AA	AG	AC	AT	18*18	18*12	18*9	18*11
G	GA	GG	GC	GT	12*18	12*12	12*9	12*11
C	CA	CG	CC	CT	9*18	9*12	9*9	9*11
T	TA	TG	TC	TT	11*18	11*12	11*9	11*11

The elements of the matrix in this case are ten elements. AA, AG, AC, AT, GG, GC, GT, CC, CT and TT. Each element of this matrix relates to a pair of submatrices of the original matrix associated with the DNA sequence (a 50*50 matrix in our case of Table 1 is considered). Constructed matrices are varied in this way depending on what property of original (n*n) matrix of the considered DNA sequence. In this work we consider the average distance between pairs of bases in construction of condensed AGCT (4*4) matrices. As shown in the condensed matrix , it is only along the diagonal that we have quadratic submatrices with zero diagonal entry, the size of which is given similarly by the total number of the corresponding bases. In this contribution we will consider the primary sequence of DNA as an input and will seek quantitative characterization for them.

In table 2 we show matrix elements needed for symmetric square matrices construction . It shows the "distance" of each label from the neighboring labels of the same and different kind for the above sequence. We will refer to such labels by serial numbers. Consider the beginning of the first exon of table 1:

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	
T	G	G	A	A	T	T	G	T	G	A	G	C	G	G	……

The 50*50 S/S matrix diagonal elements i=j are zero and could be made symmetrical by assuming (S/S)ij=(S/S)ji. Here we consider quotient of serial "distance" between selected labels of one kind only and the sequence distance when all labels are counted in the primary sequence of DNA. The first entry in the sequence, T, will contribute to TT, TG, TC and TA submatrices of the 4*4 AGCT matrix. Elements of the TT submatrix are: 1/5, 2/6, ….Elements of TA are: 1/3, 2/4, 3/10,….and soon because the distance between T and the first neighbor T (in position 6) is 5, the distance between T and the second neighbor T (in position 7) is 6. Similarly, the TG submatrix elements are also obtained by subtracting the corresponding sequential numbers of the first G, second G, the third G, and soon from the sequential number of the first T.

TABLE 2. Part of the 50*50 matrix having as elements the serial distance for the first 15 nucleic bases of the first exon of table 1

	T1	G1	G2	A1	A2	T2	T3	G3	T4	G4	A3	G5	C1	G6	G7
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
T1	0	1	2	1	2	2	3	3	4	4	3	5	1	6	7
G1		0	1	1	2	1	2	2	3	3	3	4	1	5	6
G2			0	1	2	1	2	1	3	2	3	3	1	4	5
A1				0	1	1	2	1	3	2	2	3	1	4	5
A2					0	1	2	1	3	2	1	3	1	4	5
T2						0	1	1	2	2	1	3	1	4	5
T3							0	1	1	2	1	3	1	4	5
G3								0	1	1	1	2	1	3	4
T4									0	1	1	2	1	3	4
G4										0	1	1	1	2	3
A3											0	1	1	2	3
G5												0	1	1	2
C1													0	1	2
G6														0	1
G7															0

In table 3 the distances between different T nucleic bases are grouped in TT submatrix, the distances between T and G nucleic bases are grouped in TG submatrix and soon.

TABLE 3. Part of the rearranged 50*50 matrix having as elements the serial distance for the first 15 nucleic bases so that nucleic bases of the same kind are grouped together

	T1	T2	T3	T4	G1	G2	G3	G4	G5	G6	G7	A1	A2	A3	C1
	1	6	7	9	2	3	8	10	12	14	15	4	5	11	13
T1	0	1	2	3	1	2	3	4	5	6	7	1	2	3	1
T2		0	1	2	2	1	1	2	3	4	5	2	1	1	1
T3			0	1	2	1	1	2	3	4	5	2	1	1	1
T4				0	3	2	1	1	2	3	4	2	1	1	1
G1					0	1	2	3	4	5	6	1	2	3	1
G2						0	1	2	3	4	5	1	2	3	1
G3							0	1	2	3	4	2	1	1	1
G4								0	1	2	3	1	2	1	1
G5									0	1	2	3	2	1	1
G6										0	1	3	2	1	1
G7											0	3	2	1	1
A1												0	1	2	1
A2													0	1	1
A3														0	1
C1															0

In tables 4 and 5 we represent parts of TT and GT submatrices.

TABLE 4. S/S matrix for the submatrix TT of exon of table 1

	1	6	7	9	17	23	24	25	41	43	49
1	0	1/5	2/6	3/8	4/16	5/22	6/23	7/24	8/40	9/42	10/48
6		0	1/1	2/3	3/11	4/17	5/18	6/19	7/35	8/37	9/43
7			0	1/2	2/16	3/16	4/17	5/18	6/34	7/36	8/42
9				0	1/8	2/14	3/15	4/16	5/32	6/34	7/40
17					0	1/6	2/7	3/8	4/24	5/26	6/32
23						0	1/1	2/2	3/18	4/20	5/26
24							0	1/1	2/17	3/19	4/25
25								0	1/16	2/18	3/24
41									0	1/2	2/8
43										0	1/8
49											0

In Table 5 we illustrate the AG rectangular submatrix. It has 18 columns and 12 rows corresponding to the number of A and G respectively.

TABLE 5. Truncated part of the S/S matrix for the submatrix AG of exon of table 1

	4	5	11	16	18	19	21	22	27	29	31	34	35
2	2/2	2/3	4/9	7/14	7/16	7/17	7/19	7/20	7/25	7/27	7/29	9/32	9/33
3	1/1	1/2	3/8	6/13	6/15	6/16	6/18	6/19	6/24	6/26	6/28	8/31	8/32
8	1/4	1/3	2/3	7/8	7/10	7/11	7/13	7/14	7/19	7/21	7/23	9/26	9/27
10	2/6	2/5	1/1	4/6	4/8	4/9	4/11	4/12	4/17	4/19	4/21	6/24	6/25
12	3/8	3/7	1/1	3/4	2/6	2/7	2/9	2/10	2/15	2/17	2/19	4/22	4/23
14	4/10	4/9	2/3	2/2	2/4	2/5	2/7	2/8	2/13	2/15	2/17	4/20	4/21
15	5/11	5/10	3/4	1/1	1/3	1/4	1/6	1/7	1/12	1/14	1/16	3/19	3/20
32	6/28	6/27	4/21	1/16	1/14	1/13	1/11	1/10	1/7	1/3	1/1	2/2	2/3
33	7/29	7/28	5/22	2/17	2/15	2/14	2/12	2/11	2/8	2/4	2/2	1/1	1/2
39	8/35	8/34	6/28	3/23	3/21	3/20	3/18	3/17	3/12	3/10	3/8	1/5	1/4
44	9/40	9/39	7/33	4/28	4/26	4/25	4/23	4/22	4/17	4/15	4/13	2/10	2/9
50	10/46	10/45	8/39	5/34	5/32	5/31	5/29	5/28	5/23	5/21	5/19	3/16	3/15

3 INVARIANTS OF REDUCED MATRICES

Mathematical characterization of molecules has led to hundreds of molecular descriptors, and their number continues to grow [7-10]. These descriptors, often referred to as topological indices (TI), play an important role in structure-property and structure-activity studies [11-14]. Their advantage is that they are easily available and can be quickly computed from existing or virtual structures. As already mentioned. Our task is to replace the 50*50 matrix by a 4*4 matrix, the elements of which will be extracted from the larger matrix. A pair of labels is associated with each entry in this table. As shown in the condensed matrix. We have AA, AG, AC and AT in the first row. We will assign to each element of the 4*4 matrix numerical value derived from the corresponding submatrix of the 50*50 matrix. Various submatrix invariants can be selected. As an invariant of choice we consider the Wiener number, W, which is given as the sum of the matrix elements of the distance matrix above the main diagonal [15-17].

To illustrate the approach. In Table 4 we represent the AG rectangular submatrix that records the distance between (A) and Guanine (G). It has 18 columns and 12 rows corresponding to the number of A and G, respectively. The sum of all 18*12 entries is 68.3307, which gives the average value of the matrix element of AG submatrix 68.3307/216=0.3163. Similarly, in Table 5. Because base T occurs 11 times in the first DNA exon, we obtained a symmetric 11*11 the distance matrix TT. From this 11*11 the average matrix element of all the entries in the matrix is divided by 11*11 to give 0.2579. In this way the initial 50*50 matrix with over 1500 entries is reduced to symmetrical (4*4) matrix with, at most, ten different entries. Table 6 gives the elements of the 4*4 matrix for the first exon of table 1. One expects different matrices for different sequences that facilitate comparison that is hidden in lengthy sequence of the primary DNA.

TABLE 6.The reduced 4*4 matrix for the first exon of table 1

	A	G	C	T
A	0.4166	0.3163	0.5152	0.4592
G		0.3564	0.2871	0.3319
C			0.3461	0.2870
T				0.2579

Because the condensed matrix of DNA is associated with some loss of information. One cannot recreat a structure from a list of invariants. These “undesirable” features of mathematical characterization of a structure by invariants are, in part, compensated for by the fact that one can always supplement a list of invariants by adding additional invariants. Randic [18, 19] has proposed the construction of additional (4*4) AGCT condensed matrix. In this matrix, alternative matrix invariants is used to build contraction of submatrices of the large matrix to a single entry for the corresponding reduced matrix. In table 7 the matrix elements for the limiting matrix of the first exon of table 1 is shown. It is obtained by counting entries 1 in each submatrix separately and dividing it by *pq*, where *p* and *q* are the number of rows and columns for each submatrix.

TABLE 7. The limiting binary matrix for the reduced (4*4) matrix for the exon 1 of table 1

	A	G	C	T
A	7/324	8/216	13/162	10/198
G		12/144	4/108	7/132
C			5/81	2/99
T				6/121

4 CONCLUSION

In this work we succeeded in replacing the primary sequence of DNA by condensed matrices. Such matrices allow one to make qualitative and quantitative comparisons between different sequences of

DNA, whether between within the same or between different species. The method used in this paper can be used as a marker for the toxicity of DNA-damaging agents. Comparison of two DNA sequences is now transformed into a comparison of the corresponding sequences of mathematical descriptors of DNA which is a straightforward mathematical exercise. The loss of information that accompanies such condensation can be recovered by considering additional (4*4) condensed matrices either derived by using different matrix invariants or by algebraic manipulation of existing matrix elements. Given a gene for comparison, we could create tables and graphs and index each gene compared to a standard through in the appropriate tables. This will have benefit in tabulating and classifying gene libraries.

5 REFERENCES

[1] L. Rhomberg, V. L. Dellarco, W. H. Farland, and R. S. Cortesi, The Significance of DNA Damage and Repair Mechanisms in Health Risk Assessment, in: *DNA Damage and Repair in Human Tissues*, Eds. M. Betsy Sutherland and D. Avril Woodhead, Plenum Press, New York **1990**, pp 225-232.

[2] A. Nandy, P. Nandy, and S. C. Basak, Quantitative Descriptor for SNP Related Gene Sequences, *Internet Electronic Journal of Molecular Design*.**2002**, *1*, 367-373.

[3] M. Randic and S. C. Basak, A comparative Study of Proteomics Maps Using graph Theoretical Biodescriptors, *J. Chem. Inf. Comput. Sci*. **2002***, 42*, 983-992.

[4] A. Nandy and S. C. Basak, Simple Numerical Descriptors for Quantifying Effect of Toxic Substances on DNA Sequences, *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 915-919.

[5] M. Randic, X. Guo and S. C. Basak, On Characterization of DNA Primary Sequences by Triplet of Nucleic Acid Bases, *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 619-626.

[6] M. Randic, F. Witzmann, M. Vracko and S. C. Basak, On Characterization of Proteomics Maps and Chemically Induced Changes in Proteomes Using Matrix Invariants: Application to Peroxisome Proliferators, Med. Chem. Res. **2001**, *10*, 456-479.

[7] N. Trinajstic, D. Klein and M. Randic, On Some Solved and Unsolved Problems of Chemical Graph Theory. *Int. J. Quantum Chem.: Quantum Chem. Symposium*. **1986**, *20*, 699-742.

[8] S. C. Basak and G. J. Niemi, Predicting Properties of Molecules Using Graph Invariants. *J. Math. Chem*.**1991**, *7*, 243-272.

[9] S. C. Basak, S. Bertelsen, G. D. Grunwald, Use of Graph Theoretic Parameters in Risk Assessment of Chemicals. *Toxicology Letters*. **1995**, *79*, 239-250.

[10] M. Randic, Orthogonal Molecular Descriptors. *New J. Chem*. **1991**, *15*, 517-525.

[11] D. M. Hawkins, S. C. Basak, and X. Shi, QSAR with few Compounds and Many Features. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 663-670.

[12] S. C. Basak and D. Mills, Quantitative Structure-Property Relationships (QSARs) for the Estimation of Vapor Pressure: A Hierarchical Approach Using Mathematical Structural Descriptors, *J. Chem. Inf. Comput. Sci*. **2001**, *41*, 692-701.

[13] S. C. Basak, D. R. Mills, A. T. Balaban, and B. D. Gute, Prediction of Mutagenicity of Aromatic and Heteroaromatic Amines from Structure: A Hierarchical QSAR Approach, *J. Inf. Comput. Sci.* **2001**, *41*, 671-678.

[14] A. R. Katritzky, R. Petrukhin, D. Tatham, S. C. Basak, E. Benfenati, M. Karelson and U. Maran, Interpretation of Quantitative Structure-Property and –Activity Relationships, *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 679-685.

[15] M. Randic, X. Guo, T. Oxley, H. Krishnapriyan and L. Naylor, Wiener Matrix Invariants, *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 361-367.

[16] M. Randic, On Molecular Branching. *Acta Chimica Slovenia*, **1997**, *44(1)*, 57-77.

[17] M. V. Diudea and M. Randic. Matrix Operator, W(M1, M2, M3), and Schultz-Type Indices. *J. Chem. Inf. Comput. Sci*, **1997**, *37*, 1095-1100.

[18] M. Randic, On Characterization of DNA Primary Sequences by a Condensed Matrix, *Chemical Physics Letters*, **2000**, *317*, 29-34.

[19] M. Randic and S. C. Basak, Characterization of DNA Primary Sequences Based on the Average Distances Between Bases. *J. Chem. Inf. Comput. Sci*. **2001**, *41*, 561-568.

Biographies

The Author is a lecturer in biophysics at the Atomic Energy Authority of Egypt. I have obtained a Ph. D degree in computational biophysics from the University of Salzburg, Austria. I'm undertaking postdoctoral research in mathematical chemistry at the Atomic Energy Authority.