Characterization of DNA Primary Sequences Based on the Average Distance between Bases

Paper By: Milan Randic and Subash C. Basak

July 16, 2000

Discussion and Report by: Divish P. Ranjan

April 15, 2005

Introduction

One of the most important tasks in DNA based computing involves comparing strings of nucleic bases. DNA comparison has numerous applications from crime investigations to medical/surgical uses. In most cases, efficient results in terms of time and computation cost is extremely desirable, if not necessary. Finding organ donors for emergency operations is one such situation. In this paper, Randic and Basak, describe a alternative approach to DNA Characterization for comparison of DNA sequences.

Background

Traditional methods for DNA sequence matching consider difference between two strings due to deletion-insertion, compression-expansion, and substitution of the string elements. String matching are widely studied algorithms(1). String matching method fall under class of methods that require direct DNA sequence comparison. In this paper, we will be introduced to a different method for sequence matching – based on characterization of DNA by ordered sets of invariants derived for DNA sequence. The advantage of this method is in the simplicity of sequence matching through numerical based on invariants. The flip-side is in loss of information due to simplification.

The difficulty is in finding the invariants to characterize the DNA sequence. Matrices are a way to construct such structural invariants. “Once a matrix has been constructed we can use a selection of matrix invariants as descriptors, which, upon ordering, offer a numerical characterization of the sequence.” (pg 561) There have been different methods outlined in the past for construction of such DNA based matrices based on graphical representations of DNA. Some worth mentioning are:

Graphically:

1. 2-D representation of DNA is obtained by assigning four nucleic bases the direction along the positive and negative x and y axes. (8)
2. Assign to the four nucleic acids the four tetrahedral direction in 3D space one obtains a three-dimensional representation of DNA. (9)
3. One can construct a matrix representing DNA by calculating the Euclidean (through space) and the graph theoretical (through bonds) distances between all pairs of nucleic acid bases.

Non Graphically:

4. Consider the primary sequence directly and map two numbers to each nucleic acid base: position of the base in the DNA sequence and the other giving the position of the base in the subsequence of the nucleic acid base of the same kind. For example all adenine (A) of a DNA sequence may be mapped as follows: 1/1, 2/8, 3/13, 4/20 .... so on. 2/8 means that it is the second A in the subsequence of only A's and the eighth element in the DNA subsequence. (11)
5. Count the frequency of occurrences of pairs of bases X-Y at various separations. The frequency of X-Y bases, when summarized, leads to a reduced 4x4 matrices of DNA sequence, each of such giving information on nucleic acid bases separated by different distances. X,Y = A,C,G and T. (10)

In this paper, we will see a modification of the fifth method by using average distances between pairs of bases. This will provide a set of manageable and scaled down, simplified numerical invariants. This paper describes how to derive such a structural invariant to characterize DNA sequences.

Characterization of DNA through Matrices

The goal as stated above is to create simplified matrices to numerically characterize DNA sequences. We consider different pairs of nucleic bases and can easily form a 4x4 matrix like below:

AA	AC	AG	AT
CA	CC	CG	CT
GA	GC	GG	GT
TA	TC	TG	TT

The XY pair relates to the information about the base pair XY. If order of XY bases doesn't matter the above matrix will be symmetric, and otherwise non-symmetric.

This idea can theoretically be extended to XYZ triplets making a 4x4x4 matrix. Below is the DNA sequence example that is presented in the paper:

Table 1. Exon-1 of Human Beta Globin Gene[a]

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
A	T	G	G	T	G	C	A	C	C	T	G	A	C	T
16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
C	C	T	G	A	G	G	A	G	A	A	G	T	C	T
31	32	33	34	35	36	37	38	39	40	41	42	43	44	45
G	C	C	G	T	T	A	C	T	G	C	C	C	T	G
46	47	48	49	50	51	52	53	54	55	56	57	58	59	60
T	G	G	G	G	C	A	A	G	G	T	G	A	A	C
61	62	63	64	65	66	67	68	69	70	71	72	73	74	75
G	T	G	G	A	T	G	A	A	G	T	T	G	G	T
76	77	78	79	80	81	82	83	84	85	86	87	88	89	90
G	G	T	G	A	G	G	C	C	C	T	G	G	G	C
91	92													
A	G													

[a] The nucleic bases are grouped in groups of five for better visibility.

The 92 bases Exon will be condensed to a 4x4 matrix which provide structural invariants for DNA characterization and make DNA sequence matching possible without direct sequence comparison.

Average X-Y Base Distance

To derive invariants, the paper presents the idea of constructing distance matrix for each base pair X-Y. Distance matrices are widely applied in shortest path finder, networking and other practical algorithmic problems.(2) For the Exon-1 illustrated in Table 1, we look at the distance matrix for AA or adenine-adenine nucleic bases. There are 17 instances of adenine in Exon-1 and this presents us with the a 17x17 symmetric distance matrix. Table 3 presents the matrix. The numbers on the top and to the left of the matrix represent the position of adenine in Exon-1.

Table 3. Submatrix That Is Collecting Information on All A−A Separation Distances in the Primary DNA Sequence of Table 1

	1	8	13	20	23	25	26	37	52	53	58	59	65	68	69	80	91
1	0																
8	7	0															
13	12	5	0														
20	19	12	7	0													
23	22	15	10	3	0												
25	24	17	12	5	2	0											
26	25	18	13	6	3	1	0										
37	36	29	24	17	14	12	11	0									
52	51	44	39	32	29	27	26	15	0								
53	52	45	40	33	30	28	27	16	1	0							
58	57	50	45	38	35	33	32	21	6	5	0						
59	58	51	46	39	36	34	33	22	7	6	1	0					
65	64	57	52	45	42	40	39	28	13	12	7	6	0				
68	67	60	55	48	45	43	42	31	16	15	10	9	3	0			
69	68	61	56	49	46	44	43	32	17	16	11	10	4	1	0		
80	79	72	67	60	57	55	54	43	28	27	22	21	15	12	11	0	
91	90	83	78	71	68	66	65	54	39	38	33	32	26	23	22	11	0

The above 17x17 matrix forms part of the 92x92 distance matrix for Exon-1. Subdividing into base pairs, we have the following matrix structure:

AA	AC	AG	AT	17×17	17×19	17×35	17×21
CA	CC	CG	CT	19×17	19×19	19×35	19×21
GA	GC	GG	GT	35×17	35×19	35×35	35×21
TA	TC	TG	TT	21×17	21×19	21×35	21×21

Another example is the distance matrix for AC base pair. There are 17 instance of adenine and 19 of cytosine forming a 17x19 matrix. Some interesting properties of this matrix that may help find numerical errors are as follows:

1. The difference between the successive rows in the AC sub-matrix is constant for all the rows and the columns till the position in the column where the row label becomes larger than the column label. Then the the sense of the difference is reversed and the relative magnitudes of successive rows or columns are reversed.
2. A similar regularity can be found also for the difference between the successive columns, except for the rows which have a label that is larger than the first column and smaller than the next column when instead of difference we have a constant sum.

The average distance method adds all the matrix values and divides it by the number of matrix elements to give the average matrix element value of the average distance between XY base pair. For AA matrix, we have 8564/(17 x 17) = 29.633218. For the AC matrix, we have 9994/(17 x 19) = 30.941176. Using these values, we can construct a symmetric 4x4 matrix of average distances between base pairs as show below:

Table 5. Condensed 4 × 4 Matrix: the Elements of Which Show the Average Separation between X−Y Nucleic Acid Bases (X, Y = A, C, G, T)[a]

AA	AC	AG	AT	29.633218	30.941176	30.998319	29.708683
	CC	CG	CT		30.116343	32.348872	30.441103
		GG	GT			15.193469	30.394558
			TT				28.961451
AA	AC	AG	AT	30.281500	30.806250	31.510714	30.071429
	CC	CG	CT		29.890000	31.871429	30.114286
		GG	GT			15.193469	30.394558
			TT				28.961451

[a] The top part corresponds to the DNA sequence of Table 1, and the bottom part corresponds to the hypothetical DNA sequence in which adenine at the position 58 is replaced by cytosine.

In the above matrix, the 2nd matrix shows the average distance values for the case when the adenine in the 58th position is substituted with cytosine. This matrix will be represented as A/C matrix. This will allow for comparison of two DNA sequences which differ by one nucleic base. show casing the sensitivity of the invariants.

Invariants of 4x4 Matrix

From the 4x4 matrices in Table 5, there are various invariants that can be derived. In the table below, we see a few of these invariants that seem like a good choice for characterization of the DNA sequence:

Table 6. Selection of Matrix Invariants Derived from Condensed 4 × 4 Matrix[a]

matrix invariant		A/C
the maximal row sum	123.281396	122.681965
the minimal row sum	108.935218	108.970170
the average row sum	118.392476	118.465938
the leading eigenvalue	118.638256	118.707621
other eigenvalues:	−0.399856	−0.449911
	−1.150524	−0.772788
	−13.183404	−13.158502
trace (the sum of eigenvalues)	103.904481	104.32642
average matrix element	29.598119	29.616485

[a] The last column corresponds to the case of A/C substitution.

We observe that the average row sum value and the leading eigenvalue is very close indeed. This is

because when the individual matrix elements of the 4x4 matrix do not differ much, the average row sum estimates the leading eigenvalue. Leading eigenvalues have been found to be very useful in characterization of molecular branching. Extensive material is available for referencing where leading eigenvalue have been used similarly. Please refer to the references to get a list of some selected works.

Leading Eigenvalue for DNA "profile"ing

In this paper too, leading eigenvalue is decided upon as the invariant to characterize the DNA sequence. The MATLAB function A.*B multiplies two matrix element wise, that is element a(i,j) of matrix A is multiplied with element b(i,j) of matrix B. If, A = B, then A.*B gives A^2 The leading eigenvalue of A^2 can be used as a structural invariant as well. This process can be repeated to give $A^3, A^4, A^5 \ldots$ and corresponding leading eigenvalues $\lambda_1^3, \lambda_1^4, \lambda_1^5 \ldots$ providing additional structural invariance. With increasing order of matrix A, the leading eigenvalues grow exponentially in magnitude and form a divergent series. To obtain a convergent series, we normalize the leading eigenvalues using the following formula: $\frac{\lambda_1^k}{k!^2}$. Using this, a convergent leading eigenvalue series is obtained as shown below:

Table 7. Eigenvalues, the Normalization Factors, and the Normalized Leading Eigenvalues of the "Higher Order" Condensed Matrices[a]

	eigenvalue	normalization	profile original	after A/C substitution
1	118.64	1	118.64	118.71
2	3577.67	$1/2^2$	894.42	895.22
3	10,875.22	$1/6^2$	3020.89	3023.57
4	3,320,808.00	$1/24^2$	5765.29	5768.33
5	101,705,087.43	$1/120^2$	7062.85	7061.40
6	3,121,851,701.12	$1/720^2$	6022.09	6014.04
7	96,005,201,290.02	$1/5040^2$	3779.49	3768.59
8	2,957,422,941,020	$1/40320^2$	1819.17	1810.29
9	91,249,864,640,008	$1/362880^2$	692.96	687.88
10	2,819,911,366,087,310	$1/3628800^2$	214.15	211.95
11	$8.728054\ 10^{16}$	$1/(11!)^2$	54.78	54.03
12	$2.705684\ 10^{18}$	$1/(12!)^2$	11.79	11.58
13	$8.400690\ 10^{19}$	$1/(13!)^2$	2.17	2.12
14	$2.612354\ 10^{21}$	$1/(14!)^2$	0.34	0.33
15	$8.136323\ 10^{22}$	$1/(15!)^2$	0.05	0.05
16	$2.538064\ 10^{24}$	$1/(16!)^2$	0.01	0.01

[a] The last column corresponds to the case of A/C substitution.

The above sequence of invariants presents a means for characterization of the DNA sequence. This is referred to as DNA "profile".

Comparison of original DNA sequence with A/C substitution

As stated before, the reason for the substitution of the 58th element of Exon with cytosine was to study the sensitivity of DNA profile. Since A and C are the only nucleic bases being modified, the other nucleic base pairs GT, TG, TT, GG remain unperturbed. This is can be confirmed by comparing the top and bottom GG, TT, GT average distance values of Table 5. All other nucleic base pairs average distances are changed. We also observed in Table 6 that the leading eigenvalue are quite similar for original and substituted DNA sequence. The use of higher order leading eigenvalues address this issue. In table 7, notice how the difference in eigenvalue is magnified with increasing order. If these two profiles being compared are viewed as vectors in 16 dimensional space, then the Euclidean distance between them is 17.69.

A non-leading eigenvalue method of characterizing the DNA is by canonically ordering the nucleic base pairs based on alphabetic order (AA, AC, AG, AT, CC, CG, CT, GG, GT, TT) . This gives us a 10-dimensional vectors for the original DNA and the substituted DNA. Below is the comparison based on this methodology:

Table 8. Elements of the Condensed Matrix as Components of 10-Dimensional Vector for DNA Sequence of Table 1 and the Hypothetical Sequence Obtained by Substitution of a Single Adenine by Cytosine[a]

	original DNA	A/C substitution	difference
AA	29.6332	30.2815	−0.6483
AC	30.9412	30.8063	+0.1349
AG	30.9983	31.5107	−0.5124
AT	29.7087	30.0714	−0.3627
CC	30.1163	29.8900	+0.2263
CG	32.3489	31.8714	+0.4774
CT	30.4411	30.1143	+0.3268
GG	15.1935	15.1935	0
GT	30.3946	30.3946	0
TT	28.9615	28.9615	0

[a] The last column shows the difference between the two cases.

The differences column showcases the difference in magnitude and even signs caused by the one small change in the original profile.

Discussion of the Methodologies

The method presented in this paper is very innovative in it's utilization of existing ideas. The method of DNA characterization is not constructed from scratch by proving theorems or derivation of new formulas from existing ones. In fact, this paper combines different methods under one roof to create something new. It uses the following aspects of existing methodologies:

1. Reduction of a larger, maybe unmanageable problem to a smaller more manageable problem. This

is the essence of engineering, and science – molding a problem that may not readily be solvable to something we can tackle and solve.

2. Creating distance matrix, which is a widely used and a common idea amongst computational groups. Various algorithms use this idea. Adjacency matrices used in graph theory is a central place where application of distance matrices can be found. (1)
3. Average distance. This is needs no introduction. We take average of about any data. From rainfall data, batting averages, class average etc. Analyzing data by taking average is probably the oldest data analysis method.
4. Finding eigenvalues, determinants, traces etc. All these invariants introduced in this paper have a historical background. Leading eigenvalue and normalization are also common ideas in data manipulation and analysis.

All of the above methods have existing foundations. There exists optimal methods to implement most of the above ideas and combining such methods to obtain an effective way of DNA profiling is very innovative and practical. Another advantage of using such a method is of achieving ease of implementation. I found this idea of generating DNA characterization very creative for sequence comparison.

References

Books:

1. Cormen, T. H; Leiserson, C. E.; Rivest, R. L.; Stein, C; Introduction to Algorithms, 2nd Edition.
2. Peterson, L. L; Davie, B. S; Computer Networks: A Systems Approach, 3rd Edition.

Leading Eigenvalue in Molecular Branching:

3. Lovasz, L. ; Pelikan, J. I. On the eigenvalues of tree. *Period. Math. Hung. 1973, 3, 175-182.*
4. Randic M.; Vracko, M.; Novic, M. Eigenvalues as molecular descriptors. In *QSAR/QSPR by Molecular Descriptors;* Diudea, M. V., Ed.; Nova Publ.: In press.
5. Randic, M.; Kleiner, A. F.; DeAlba, L. M. Distance/distance matrices. *J. Chem. Inf. Comput. Sci. 1994, 34, 277-286.*
6. Randic, M. On Structural ordering and branching of acyclic saturated hydrocarbons. *J. Math. Chem. 1998, 24, 345-358.*
7. Randic, M.; Guo, X.; Bobst, S. Use of path matrices for characterization of molecular structures. *DIMACS Ser. Discrete Math. Theor. Comput. Sci. 2000, 51, 305-322.*

Constructing DNA Based Matrices:

8. Randic, M.; Novic, M.,; Vracko, M. Molecular Descriptors, New and Old Lecture Notes in Chemistry.
9. Randic, M.; Vracko, M.; Nandy, A.; Basak, S. C. On 3-D graphical representation of DNA primary sequences an their numerical characterization. *J. Chem. Inf. Comput. Sci. 2000, 40, 1235 – 1244.*
10. Randic, M.: Condensed representation of DNA primary sequences. *J. Chem. Inf. Comput. Sci. 2000, 40, 50 – 56.*
11. Randic, M. On Characterization of DNA primary sequences by condensed matrix. *Chem. Phys. Lett. 2000, 317, 29-34.*