Available online at www.sciencedirect.com

ScienceDirect

Molecular Phylogenetics and Evolution 41 (2006) 64–69

MOLECULAR PHYLOGENETICS AND EVOLUTION

www.elsevier.com/locate/ympev

Clustering DNA sequences by feature vectors

Libin Liu [a], Yee-kin Ho [b], Stephen Yau [a,*]

[a] *Department of Mathematics, Statistics and Computer Science, University of Illinois at Chicago, M/C 249 Chicago, IL 60607-7045, USA*
[b] *Department of Biochemistry and Molecular Genetics, University of Illinois at Chicago, M/C 249 Chicago, IL 60607-7045, USA*

Received 22 July 2005; revised 14 May 2006; accepted 17 May 2006
Available online 26 May 2006

Abstract

We represent all DNA sequences as points in twelve-dimensional space in such a way that homologous DNA sequences are clustered together, from which a new genomic space is created for global DNA sequences comparison of millions of genes simultaneously. More specifically, basing on the contents of four nucleotides, their distances from the origin and their distribution along the sequences, a twelve-dimensional vector is given to any DNA sequence. The applicability of this analysis on global comparison of gene structures was tested on myoglobin, β-globin, histone-4, lysozyme, and rhodopsin families. Members from each family exhibit smaller vector distances relative to the distances of members from different families. The vector distance also distinguishes random sequences generated based on same bases composition. Sequence comparisons showed consistency with the BLAST method. Once the new gene is discovered, we can compute the location of this new gene in our genomic space. It is natural to predict that the properties of this new gene are similar to the properties of known genes that are locating near by. Biologists can do various experiments to test these properties.

Keywords: DNA sequences; Genomic space; Vector distance; Global comparison of gene structures

1. Background

With the development of technology, more and more biological data has been acquired. The number of sequences in GenBank has been growing exponentially in the past few years (http://www.ncbi.nlm.nih.gov). Many analysis methods have been proposed. One of them is graphical representation of DNA sequences. Early graphical representations suffer from the degeneracy (Gates, 1985; Liu et al., 2002). Recently, graphical representation without degeneracy has been proposed (Yau et al., 2003). That representation provides an efficient way to visualize, sort or compare short genes (Fig. 1). To analyze long biological sequences, sequences have to be numerically characterized. Traditionally, DNA sequences are transferred to the vector by using indicator vectors. For DNA/RNA sequence, a vector with length 4 represents every nucleotide. For example sequence AATGC will be represented as <1000 1000 0100 0010 0001>. For protein sequence, a vector with length 20 represents every amino acid. There are two disadvantages about the indicator vector. First of all, vector is much longer than original sequence. In the case of protein sequence, the vector will be 20 times longer than the original sequence. Secondly for the different biological sequences, we will get vectors with different length. This will bring difficulty in computation. In this report, DNA sequences will be analyzed at different levels of complexity. First level is to study A, G, C, T contents and their distributions along the primary sequences. The second level is to analysis on di-nucleotide (AA, TT, GG, CC, AT, ...) level and the third level is to study triplet codons. In a 4 bases system, basing on the contents of four nucleotides, their distances from the origin and their distribution along the sequence, a twelve-dimensional vector is given for any DNA sequence. The applicability of this analysis to global comparison of gene structure is tested on myolobin, β-globin, histone-4, lysozyme and rhodopsin families. Members of each family exhibit smaller vector distances relative to the distances of members from different families. The vector distance also distinguishes

[*] Corresponding author. Fax: +1 312 996 3065.
E-mail address: yau@uic.edu (S. Yau).

1055-7903/$ - see front matter © 2006 Elsevier Inc. All rights reserved.
doi:10.1016/j.ympev.2006.05.019

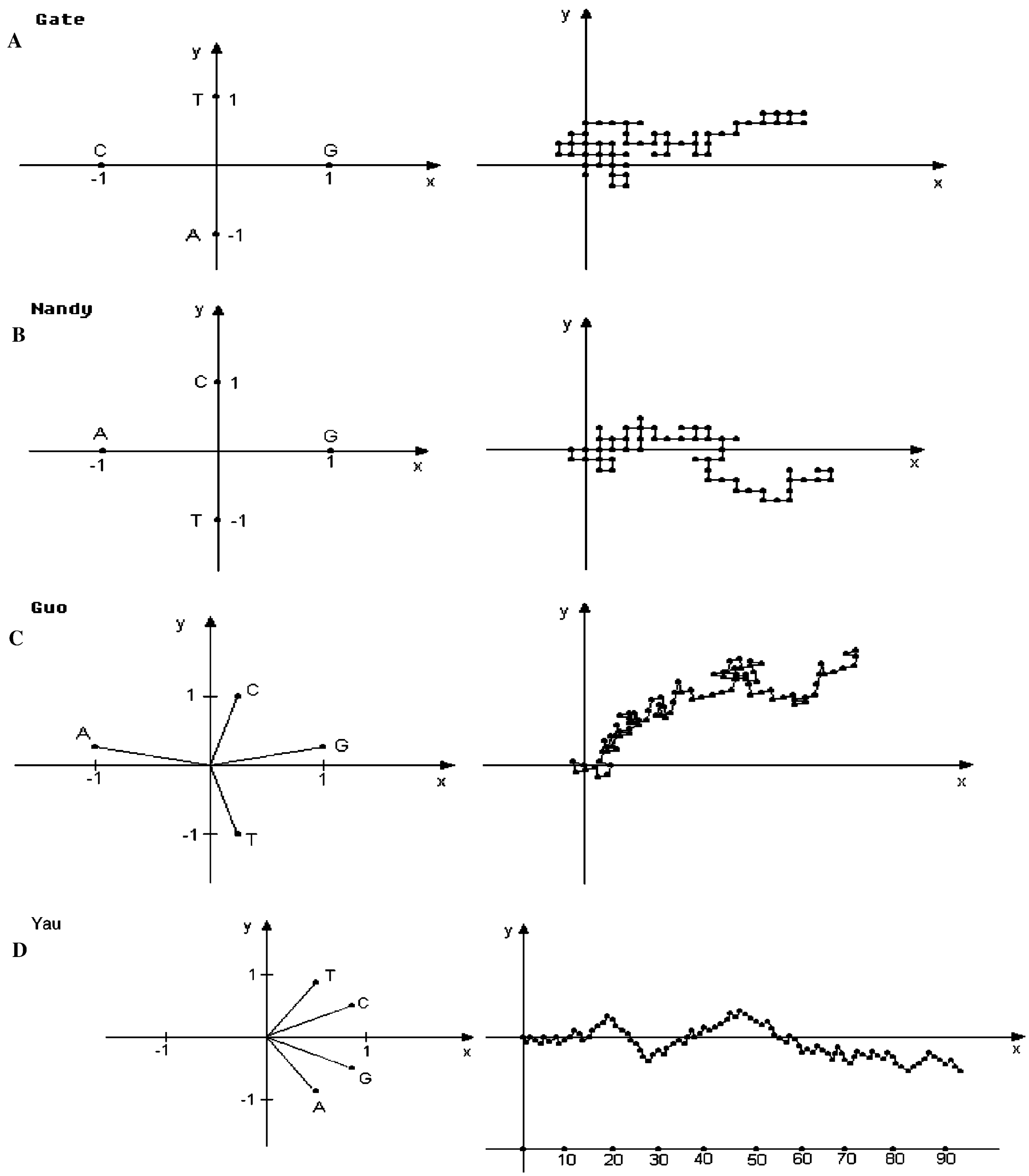

Fig. 1. Graphic representation of DNA sequence.

random sequences generated based on the same bases composition. Furthermore, the analysis is sensitive to detect single base change. The clustering results of homologous sequences are consistent with BLAST computation (Altschul et al., 1990). Unlike BLAST, our novel system creates a new genomic space for global DNA sequence comparison of millions genes simultaneously. Once the new gene is discovered, we can compare the location of this new gene in our genomic space. It is natural to predict that the properties of this new gene are similar to the properties of known genes that are locating near by. Biologists can do various experiments to test these properties.

2. Methods

To employ vector to characterize the DNA sequences, vectors have to be the same length no matter how different the original sequences are. In this paper, we associate to each DNA sequences a twelve-dimensional vector. This creates a new genomic geometry in twelve-dimensional space. The method described below is for the first level of understanding the distributions of four nucleotides A, T, C, G, but it is applicable for the analysis of the sixteen di-nucleotides and the sixty-four triplet codons systems.

Multiple numerical information that is directly related to the DNA sequence is selected to formulate the strategy for mathematical analysis.

(1) The nucleotide A, T, G, and C contents from the DNA sequence are chosen as the first parameter in the vector analysis. For any DNA sequence, it will have a defined base content and the total numbers of A, T, G, and C also dictate the length of the sequence. So the characterization vector contains n_{A}, n_{G}, n_{T}, and n_{C}. These four integers denote numbers of nucleic bases A, G, T, and C of the DNA sequence respectively. It is obvious that just the nucleotide contents is not sufficient to denote a specific DNA sequence since two different DNA sequences of the same length can have exactly the same nucleotide contents. So more parameters are needed.

(2) The second numerical parameter is the total distances of each nucleotide base to the first nucleotide. Suppose that we have two DNA sequences. The first one has two adenine nucleotides at the positions 2 and 3, while the second one has two adenine nucleotides at positions 8 and 9. In this example, both DNA sequences will have two adenine bases, but the total distance generated from these two cases are different which is a special characteristic to the sequence. Four sets of total distance measurement for A, T, G, and C to the original can be generated respectively. Total distance T_i is defined as:

$$T_i = \sum_{j=1}^{n_i} t_j$$

$i = \mathrm{A, G, T, C}$; t_j is the distance from the first nucleotide to the jth nucleotide i in the DNA sequence.

The characteristics of the four sets of total distances T_{A}, T_{G}, T_{C}, and T_{T} are dictated by the DNA sequence that reflect the information how far is each nucleic base from the first nucleotide. If two DNA sequences are similar, the information about total distance should also be similar. However, it is important to point out that the measurement of total distances alone is also insufficient to denote the DNA sequence for comparison. For example we can have one DNA sequence which has two adenines in the positions 4 and 6, while the other DNA sequence also has two adenines, but in the positions 3 and 7. Those two DNA sequences have same number of adenine nucleotides (A) and are appeared at different positions in the DNA sequences but their total distances T_{A} from the origin are the same. It is obvious that another numerical parameter is needed from the DNA graphs to further define the DNA sequence.

(3) The third parameter selected for the vector analysis is the distribution of each nucleotide alone the DNA sequence. If the distribution of each nucleotide base is different, DNA sequences cannot be similar even though they may have the same nucleotide contents and the same total distance measurement. Therefore, the information about distribution has also been included in the vector analysis. The variance of distance for each nucleic base used to describe the distribution is defined as following:

$$D_i = \sum_{j=1}^{n_i} \frac{(t_j - \mu_i)^2}{n_i}$$

where $i = \mathrm{A, T, G, C}$; t_j is the distance from the first nucleotide to the jth nucleotide i in the DNA sequence

and

$$\mu_i = \frac{T_i}{n_i}$$

. As described above, each set of numerical parameter is not sufficient to denote specific DNA sequence. However, a combine characterization vector that contains all of three sets of parameters could be used to characterize similarity between DNA sequences. So the characterization vector, which contains twelve-dimensional information, is given as follow:

$$< n_{\mathrm{A}}, T_{\mathrm{A}}, D_{\mathrm{A}}, n_{\mathrm{G}}, T_{\mathrm{G}}, D_{\mathrm{G}}, n_{\mathrm{T}}, T_{\mathrm{T}}, D_{\mathrm{T}}, n_{\mathrm{C}}, T_{\mathrm{C}}, D_{\mathrm{C}} >$$

The characterization vector can be used as a numerical measure of similarity of different DNA sequences. In order to compare the similarity and difference among DNA sequences, we introduce the distance between vectors as an index for comparison. If two DNA sequences are similar, the distance between these two characterization vectors should be small. Otherwise, large distance between the characterization vectors is expected for non-homologous DNA sequences. The distance of two characterization vectors is defined as:

$$L = \sqrt{\sum_j \sum_i (j_i - j_i')^2} \quad i = \mathrm{A, G, T, C};\ j = n, T, D.$$

In this way, we have created a twelve-dimensional genomic space, of which points are the DNA sequences. The novelty of this genomic space is that it allows us to assign a distance between two DNA sequences which measure the similarity between them.

The practical application of using the characterization vector for DNA sequence comparison is straightforward. For example, each gene sequence from a genome can be given a characterization vector. A data bank of characterization vectors corresponding to individual gene sequence in the genome can be compiled. DNA sequence comparison can be accomplished on numerical level with comparison on a single number of the characterization vector. There is no need to conduct detailed nucleotide bases alignment between different DNA sequences. This novel system creates new genomic space for global DNA sequences comparison of millions of genes simultaneously. Once the new gene is discovered, we can compare the location of this new gene in our genomic space. It is natural to predict that the properties of this new gene are similar to the properties of known genes that are locating near by. Biologists can do various experiments to test these properties.

3. Results and discussion

The genomic space represents a novel approach in global DNA sequences comparison. In principle, each gene sequence will generate a characterization vector. Direct comparison of the characterization vectors among different genes will give a measure of the degree of homology or difference of these genes. The method bypasses the need of sequence alignment, gap generation, and selective matrix used in statistical analysis. We have selected DNA coding sequences of a few gene families to test the applicability of the method in sequence comparison and the sensitivity of the method to reflect minute change in gene sequence.

Four members of the myoglobin family (cow, rat, human, and mouse) were subjected to genomic space analysis. The results are shown in Table 1. All four myoglobin sequences have the length of 435 bases. The distances between two characterization vectors of the corresponding pair are shown in Columns 1–4. As a control, we generated four random sequences with the same base composition and length by computer. The distances between the myoglobin genes and the random sequences are listed in Columns 5–8. The vector distances among the myoglobin family are clustered together ranging from 1.9782×10^3–4.3593×10^3 which are easily distinguishable from the random sequences ranging from 4.1240×10^3–7.9292×10^3. This result demonstrated that the method can easily distinguish homologous sequences from random sequences even they have the same base composition and length.

To examine different protein families, we have chosen myoglobin, β-globin, histone-4, lysozyme, and rhodopsin for analysis. Each family contains several members. The results are shown in Table 2. The distances of the characterization vectors between members of the same family are clustered together. The distances become significantly large between different families. Moreover, the maximum distance inside each member of the same family is less than the minimum distance between the different families. The genomic space is applicable for global DNA sequence comparison. The distance reflects the degree of homology between to DNA sequences.

As a control, we have conducted the sequence comparison using the BLAST search method. The percentages of homology among the myoglobin, β-globin, and histone-4 families are summarized in Table 3. Results from genomic space analysis are consistent with the BLAST search computation.

Sensitivity of genomic space analysis towards minimum base change in a DNA sequence was examined. We define the sensitivity as:

$$S = d^2/|v|^*|v'|$$

where d is the distance between original DNA sequence and after base change DNA sequence; $|v|$ is the absolute value of original DNA sequence's characterization vector; $|v'|$ is the absolute value of DNA sequence's characterization vector after base change.

Table 1
Comparison of DNA sequences of Myoglobin family and random generated DNA sequence (Unit: 10^3)

	Myoglobin family				Random sequence with same length			
	I	II	III	IV	V	VI	IX	X
	Bos taurus myoglobin	Rattus norvegicus myoglobin	Homo sapiens myoglobin	Mus musculus myoglobin	Random 1	Random 2	Random 3	Random 4
I	0	2.9282	3.7670	3.3271	5.1687	7.1735	6.5532	5.3596
II		0	4.3539	4.0183	4.8003	5.1079	4.7706	5.7353
III			0	1.9782	6.3274	7.8575	7.8671	4.9823
IV				0	5.6902	7.9292	7.3538	4.1240

Table 2
The distances among members and families (unit: 10^4)

	Myoglobin family				β-Globin			Histone-4		Lysozyme			Rhodopsin	
	I Bos taurus myoglobin	II Rattus norvegicus myoglobin	III Homo sapiens myoglobin	IV Mus musculus myoglobin	V Homo sapiens hemoglobin, β	VI Bos taurus hemoglobin, β	VII Danio rerio embryonic 1 β-globin (bE1)(be1)	VIII Homo sapiens histone-4	IX Mus musculus histone-4	X Danio rerio lysozyme	XI Mus musculus lysozyme	XII Rattus norvegicus lysozyme	XIII Bos taurus rhodopsin kinase	XIV Homo sapiens rhodopsin kinase
I	0	0.2928	0.3767	0.3327	0.7185	0.9948	0.8709	4.4171	4.4232	1.783	1.307	1.304	34.526	34.863
II		0	0.4354	0.4018	0.7169	0.9268	0.9254	4.4197	4.4270	1.724	1.291	1.306	34.506	34.884
III			0	0.1978	0.9505	1.1053	1.0253	4.4305	4.4397	1.737	1.242	1.210	34.582	34.907
IV				0	0.9929	1.1174	1.0433	4.4535	4.4630	1.767	1.286	1.251	34.542	34.866
V					0	0.4535	0.3786	4.2520	4.2504	1.574	1.311	1.420	34.686	35.031
VI						0	0.5031	4.2718	4.2701	1.459	1.332	1.476	34.730	35.070
VII							0	4.2544	4.2500	1.568	1.318	1.431	36.761	35.095
VIII								0	0.0623	3.016	3.250	3.29	38.874	39.233
IX									0	3.022	3.258	3.301	38.878	39.237
X										0	0.637	0.8080	36.061	36.381
XI											0	0.229	35.759	36.082
XII												0	35.754	36.072
XIII													0	2.006
XIV														0

S is the dimensionless variable that reflects percentage of change if bases in sequence are altered. The results of sensitivity among the myoglobin, β-globin, and histone-4 families are shown in Table 4. The degree of sensitivity varies based on the length of DNA sequence. The longer the sequence is, the lower the sensitivity becomes. It also depends on the position of the base changed along the sequence. Base alteration near the 5′ end has shorter distance from the origin. As a result, the sensitivity is low. We also analyze the sensitivity changes corresponding to one, two and three bases changes. The results are shown in Fig. 2. There are corresponding increases of sensitivity as the number of bases is changed. Overall, the genomic space analysis is sensitive to a single base change at any position of the sequence.

4. Authors' contribution

The results of DNA sequences comparison among homologous sequences gives close distances between their characterization vectors which is easily distinguishable from non-homologous or random sequences. Moreover, the numerical value is capable to detect one single base mutation on a given DNA sequence. The single codon system is easily expandable from the simple four bases of A, T, G and C to the sixteen base-pairs nearest neighbor analysis and further to the sixty four triplet codons related to the translated protein sequences. The different levels of the DNA sequence analyses is directly linked to their physical, chemical, and biological activities.

The mathematical analysis using the vector system is versatile. We use a twelve-dimensional vector analyses. For the sixteen coordinates system, one can expand the analyses to forty-eight dimensions accordingly. In the current analysis, we have chosen the base composition, the distances of the bases to the 5′ end, and the distribution of bases along the sequence as the mathematical parameters for the vector analysis. One could add more characteristic parameters to increase the sensitivity of the analysis or delete parameter from the analysis when it is no longer contributing to the sensitivity.

Results in comparing DNA sequences correlate closely with the BLAST method. Quantitative correlation of the vector distances of the characterization vectors to the percentage of homology of the BLAST method has not been established. It is likely that they will be variation between the two methods. The BLAST search utilizes statistical model with an established matrix to score sequence alignment. Base substitution and gap generation are allowed when the final result gives a better score. The main advantage of our new method is that it allows global comparison of multiple sequences (all the genomic sequences) simultaneously by direct comparison of the distances of their characteristic vectors.

Future work will emphasize in three directions: (1) To build a database of the characterization vectors for all coding sequences in genomes and to test the applicability of the

Table 3
Similarity results from BLAST

	Myoglobin family				β-Globin			Histone-4	
	I Bos taurus myoglobin	II Rattus norvegicus myoglobin	III Homo sapiens myoglobin	IV Mus musculus myoglobin	V Homo sapiens hemoglobin, β	VI Bos taurus hemoglobin, β	IX Danio rerio embryonic 1 β-globin (bE1)(be1)	X Homo sapiens histone-4	XI Mus musculus histone-4
I	100%	70%	78%	69%	NA	NA	NA	NA	NA
II		100%	72%	86%	NA	NA	NA	NA	NA
III			100%	72%	NA	NA	N	NA	NA
IV				100%	NA	NA	NA	NA	NA
V					100%	79%	54%	NA	NA
VI						100%	49%	NA	NA
IX							100%	NA	NA
X								100%	71%
XI									100%

Table 4
Sensitivity result for single base change

	Homo sapiens rhodopsin kinase (RHOK), mRNA (Length: 1152)	Bos taurus myoglobin (MB), mRNA (Length: 435)	Homo sapiens histone-4, H4, mRNA (Length: 207)
Change first A–G	0.6440×10^{-5}	0.3741×10^{-4}	0.1227×10^{-3}
Change first A–T	0.9696×10^{-5}	0.8573×10^{-4}	0.1597×10^{-3}
Change first A–C	0.6676×10^{-5}	0.6798×10^{-4}	0.1268×10^{-3}
Change random middle A–G	0.5086×10^{-5}	0.3453×10^{-4}	0.1608×10^{-3}
Change random middle A–T	0.6221×10^{-5}	0.4233×10^{-4}	0.1873×10^{-3}
Change random middle A–C	0.5268×10^{-5}	0.3401×10^{-4}	0.1615×10^{-3}
Change last A–G	0.2199×10^{-4}	0.1676×10^{-3}	0.6401×10^{-3}
Change last A–T	0.2752×10^{-4}	0.1736×10^{-3}	0.7847×10^{-3}
Change last A–C	0.2226×10^{-4}	0.1547×10^{-3}	0.6697×10^{-3}

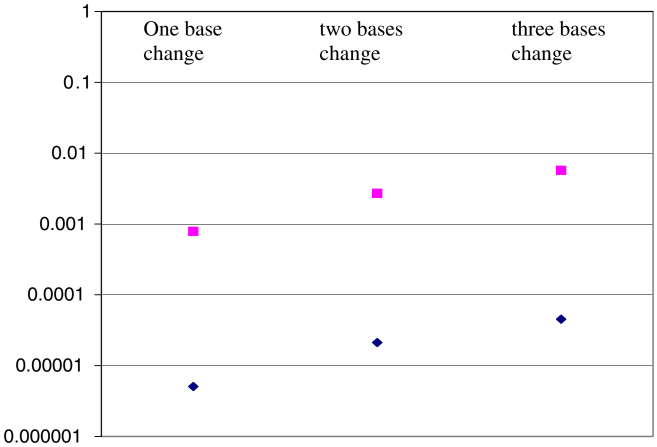

Fig. 2. Sensitivity analysis for one, two and three bases change.

system for global DNA sequences comparison, (2) To further characterize the sixteen coordinates system for double base-pairs and the sixty-four coordinates system for codon analysis and to extend the comparison to include amino acid sequences and protein structure.

References

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J., 1990. Basic alignment search tool. J. Mol. Biol. 215 (3), 403–410.

Gates, M.A., 1985. Simpler DNA sequence representations. Nature 31, 219.

Liu, Y., Guo, X., Xu, J., Pan, L., Wang, S., 2002. Some notes on 2-D graphic representation of DNA sequence. J. Chem. Inf. Comput. Sci. 42, 529–533.

Yau, S., Wang, J., Niknejad, A., Lu, C., Jin, N., Ho, Y., 2003. DNA sequence representation without degeneracy. Nucleic Acids Res. 31, 3078–3080.