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

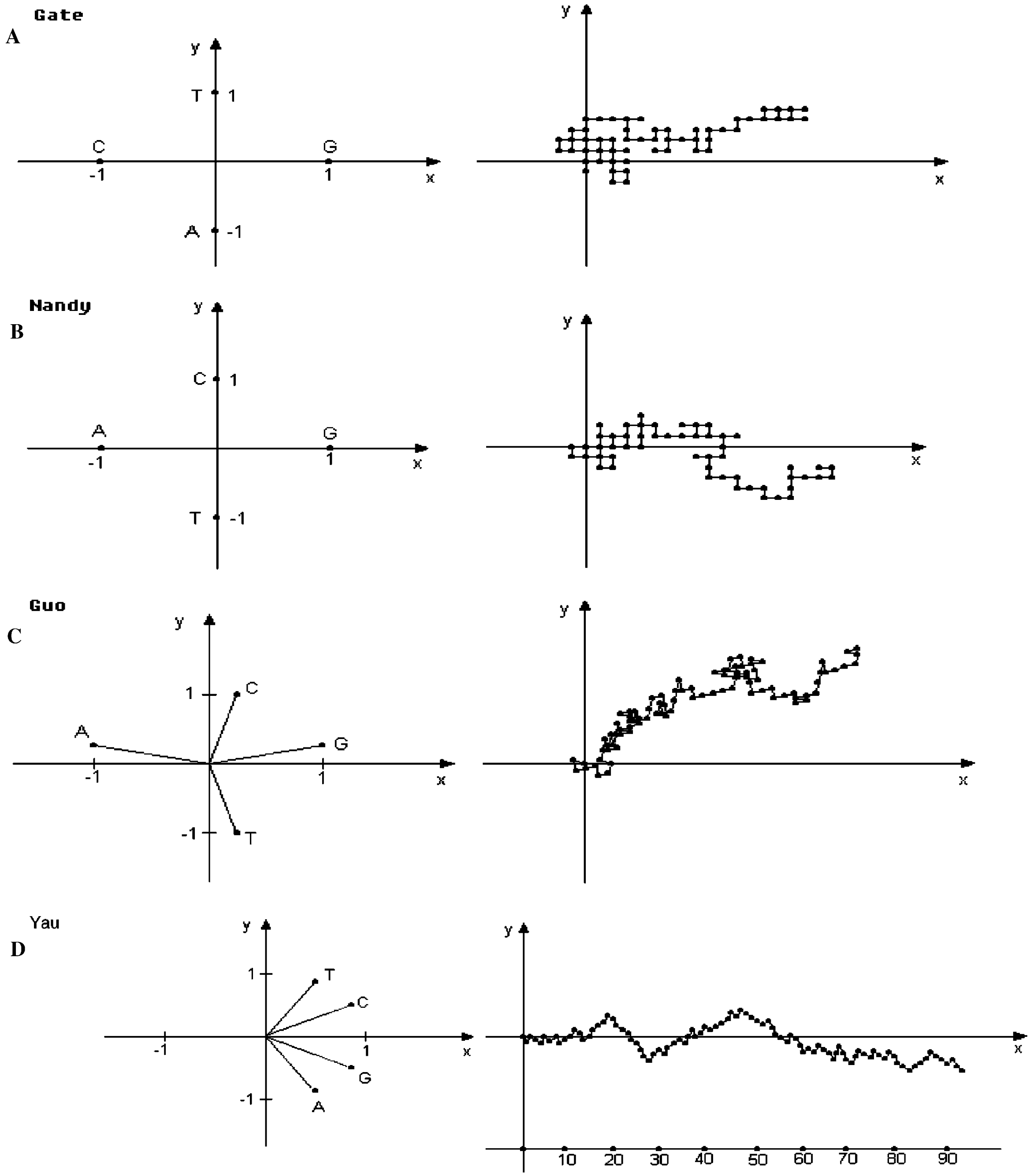

Fig. 1. Graphic representation of DNA sequence.

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.