Discrete Applied Mathematics 157 (2009) 2378–2387

Contents lists available at ScienceDirect

Discrete Applied Mathematics

journal homepage: www.elsevier.com/locate/dam

Prokaryote clustering based on DNA curvature distributions

L. Kozobay-Avraham [a,b], S. Hosid [a,b], Z. Volkovich [c], A. Bolshoy [a,b,*]

[a] Department of Evolutionary and Environmental Biology, University of Haifa, Haifa 39105, Israel

[b] Genome Diversity Center of Institute of Evolution, University of Haifa, Haifa 39105, Israel

[c] Software Engineering Department, ORT Braude College of Engineering, Karmiel 21982, Israel

ARTICLE INFO

Article history:
Received 22 January 2007
Received in revised form 20 June 2007
Accepted 11 June 2008
Available online 23 September 2008

Keywords:
Curved DNA
Clustering methods
k-means
PAM

ABSTRACT

Massive determination of complete genome sequences has led to the development of different tools for genome comparisons. Our approach is to compare genomes according to typical genomic distributions of a mathematical function that reflects a certain biological function. In this study we used comprehensive genome analysis of DNA curvature distributions in coding and non-coding regions of prokaryotic genomes to evaluate the assistance of mathematical and statistical procedures. Due to an extensive amount of data we were able to define the factors influencing the curvature distribution in promoter and terminator regions such as growth temperature, genome size, and $A+T$ composition. Two clustering methods, K-means and PAM, were applied and produced very similar clusterings that reflect genomic attributes and environmental conditions of the species' habitat.

1. Introduction

The term DNA curvature refers to a characteristic of DNA fragments, which are bent without application of any external forces. This property is also called intrinsic curvature or sequence-dependent DNA curvature. The presence of curved DNA was established by biological experiments in the early 1980s (see, [22,35,4,7,31]). Based on experimental results, some computational models, including our model [2,26], were developed to predict the magnitude of DNA curvature with high reliability. The existence of upstream curved sequences (UCS) was shown experimentally for many genes in prokaryotes [25]. Comprehensive genome analysis of DNA curvature in regulatory regions was performed by us [3,17,16,18], and others [12, 23].

It is well known that the genomes are annotated with quite dramatically varying degrees of quality. The most striking examples are the *A.* pernix and *P.* horikoshii, which are two of a small handful of genomes that are classified as hyperthermophiles. Naturally, one could suspect that part of the trends that might be seen is simply artifactual fluctuations, due to differing qualities in gene finding, rather than "real" differences. Our analysis brought us to the conclusion there is no systematic bias in the quality of Archaeal gene predictions (data not shown).

Offered approach was to detect predicted frequent occurrences of especially high local extremes of curvature distribution function upstream to starts of the coding sequences. Following detection of similarly outstanding curvature distribution downstream of genes in *Escherichia coli* and *Bacillus subtilis*, wide genomic comparisons (170 complete prokaryotic genomes) were performed, and we found that not only upstream, but also downstream, intergenic regions are significantly more curved than would be expected from their dinucleotide composition [18]. Putative influence of environmental and genomic factors as well as taxonomic factors on curvature distribution in promoter and terminator regions were indicated [3,17,16,

* Corresponding author. Tel.: +972 4 8240382; fax: +972 4 8240382.
E-mail address: bolshoy@research.haifa.ac.il (A. Bolshoy).

0166-218X/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.dam.2008.06.049

18]. The most prominent effect on DNA curvature distribution, in these regulatory regions, was presented by the growth temperature.

To avoid any misunderstanding we want to mention that everywhere in this manuscript the term "start of gene" means "start of translation or a position of a first codon", and "end of gene" means "end of translation or a position of a stop codon". "Start of gene" does not refer to the transcription +1 site! The reason for this usage is rather simple: transcription initiation site is not known for most of the more than 200 genomes used.

Cluster analysis and other statistical tests were performed on genomic curvature distributions data. For every genome, we predicted DNA curvature profiles: one typical profile for the start of translation (covering a promoter region and a beginning of a coding region) and another one around the end of genes (3′-end of coding sequence and a putative transcription terminator region). The profiles were predicted using the CURVATURE program [26]. Actually, instead of using raw curvature distributions for clustering, we used normalized data. Randomized sequences were constructed and curvature excess profiles in standard deviation units were calculated for each genome.

The normalization distinguishes our attitude to construction curvature profiles from the profiles used in other studies [24, 33,3].

Six different distances have been examined for the purpose of future clustering. The squared Euclidean distance between the genomes appears to be the most reasonable from a biological point of view. Consequently, this distance has been used for all further mentioned results of clustering.

2. Methods

2.1. Clustering methods

Clustering problems arise in many areas of bioinformatics. Clustering is an example of unsupervised learning when the number and type of classes are unknown, and available data samples are unlabeled. Groups (clusters) are constructed to achieve a relatively high similarity among the groups' elements in addition to a relatively low similarity between elements of different groups. Let us consider a subset $X = \{x_1, x_2, \ldots, x_m\}$ in the n-dimensional Euclidean space R^n. Consider a partition $\Pi = \{\pi_1, \ldots, \pi_k\}$ of the set, i.e.

$$\bigcup_{j=1}^{k} \pi_j = X, \qquad \pi_i \cap \pi_j = \varnothing \quad \text{for } i \neq j. \tag{1}$$

For a real-valued function q whose domain is the set of subsets of X the quality of the partition is defined as

$$Q\,(\mathbf{\Pi}) = \sum_{j=1}^{k} q(\pi_j). \tag{2}$$

In fact, clustering problem is another instance of a global optimization problem of finding a partition

$$\Pi^{(0)} = \{\pi_j^{(0)}, j = 1, \ldots, k\}, \tag{3}$$

which optimizes $Q(\Pi)$. Often function q is associated with a "dissimilarity measure", or a distance-like function $d(x, y)$. The term a distance-like function is used, since this function is not required to satisfy all requirements to a distance function: the function is not necessarily symmetrical or necessarily satisfies the triangle inequality. Function q can be constructed by means of $d(x, y)$ as such. We introduce k centroids (medoids) $(c_1, \ldots, c_k)$ as a prescribed subset of R^n. This set defines a partition of X as

$$\pi_i = \{x \in X : d(c_i, x) \leq d(c_j, x), \text{ for } i \neq j\}. \tag{4}$$

(Ties are broken arbitrarily.) On the other hand, for a given partition the centroids set is defined as

$$c(\pi_i) = \arg\min_{c_i} \left\{ \sum_{x \in \pi_i} d(c_i, x) \right\}. \tag{5}$$

Thus,

$$q(\pi_i) = \sum_{x \in \pi_i} d(c(\pi_i), x) \tag{6}$$

and the mentioned optimization problem is reduced to finding an appropriate centroids' set as a solution to the task

$$C = \arg\min_{c_1, \ldots, c_k} \left\{ \sum_{i=1}^{k} \sum_{x \in \pi_i} d(c_i, x) \right\}. \tag{7}$$

Dissimilarity measures can be chosen in different ways, particularly as information distances of divergences (see, for example [28,15]). We consider the following six cases for two vectors $x = \{x_i, i = 1, \ldots, n\}$ and $y = \{y_i, i = 1, \ldots, n\}$: