B-Splines for Genomic Signatures

Filomena de Santis
Dipartimento di Informatica e Applicazioni
"R.M. Capocelli"
Università di Salerno
Baronissi, Salerno, ITALY

Gennaro Iaccarino
Dipartimento di Informatica e Applicazioni
"R.M. Capocelli"
Università di Salerno
Baronissi, Salerno, ITALY

Abstract *- A variety of data analysis concerned with genome sequences support the proposal that each living organism owns a genomic signature. Classical approaches to the genomic signature deciphering are based on the counting of oligonucleotides in DNA sequences. Statistical tools as well as graphical representations have been used to associate genomic signatures with oligonucleotides frequencies in the genome. Moreover, many experimental results also show that the core features characterizing the whole genome are preserved in short subsequences, validating the idea of a genomic signature. In this paper we present a new method to detect the genomic signature. The proposal stands on two ideas: first, give a simple mathematical representation of trinucleotide frequencies; second, deal it with B-spline functions. The result is a continuous function representing the genomic signature. The advantages of our method resides in maintaining the essential properties of the signature itself and providing a more formal instrument of analysis.*

Keywords: *B-Splines, Spline Functions, Genomic Signature, Oligonucleotides.*

1 Introduction

Classical approaches to the genomic signature deciphering are based on the counting of specific subsequences, such as dinucleotides, trinucleotides, and tetranucleotides, in a long DNA sequence; however, this sort of counting is well suited for words of length less or equal than four. For longer words a more tractable tool, the so-called Chaos Game Representation, (CGR for short), is used with the objective of depicting word frequencies in the form of fractal images for any length *n*. CGR experimental results have also shown that the core features characterizing the whole genome are preserved in short subsequences; this strongly holds up the idea of a genomic signature pertaining to each leaving organism, [8, 16, 20]. In this paper we present a new mathematical method to represent the genomic signature that allows us to achieve both the goals of providing a graphic visualization of the result and an analytical expression of it. The first achievement is based on the use of spline functions, and the second on the use of B-splines. Thus, for each genome the signature can be expressed as an elegant mathematical form, or equivalently as a 2D graphics, that also gives chances to exploit the genome with other analytical tools.

1.1 State of the Art

Whole genome sequencing, in addition to create the primary information on the structure and the function of genes , has opened a variety of new research fields devoted to the analysis of their structures. A comparative analysis of genome structures, indeed, allows to ask many questions of overall nature: for example, whether there exist short strings that do not appear at all in a genome or , quite the opposite, exhibit the main characteristics of the whole genome attesting the validity of the *genomic signature* hypothesis.

The genomic signature concept begun to raise early in the 1960s and 1970s, when the first characterizations of the DNA were tried using *in vitro* biochemistry to measure the doublet (nearest-neighbour two-base sequences) frequencies and determine the patterns of deviation from their random expectation. The technique of doublet frequency analysis, originally developed by Kormberg and coauthors [11], proved that the observed frequencies of the sixteen possible doublet sequences normalized to the component mononucleotides give the 'general design', that is to say an unique and characteristic pattern for each organism or class of organism. In [11, 13, 14, 15, 16, 17, 19, 20] a detailed study involving a broad phylogenetic range with the aim to detect predispositions and irregularity in the occurrences of dinucleotides, trinucleotides, and tetranucleotides within and between genomic sequences is presented. In particular, extremes of over-representation and under-representation of short oligonucleotides are identified by means of the evaluation of the *dinucleotide relative abundance* for each pair of dinucleotides XY:

$$\rho_{XY} = f_{XY} \; / \; f_X \; f_Y$$

where f_X and f_Y, respectively, denote the frequencies of the nucleotides X and Y in a DNA sequence S, and f_{XY} the frequency of the dinucleotide XY. The experimental tests done on many different genomes have shown that the dinucleotide relative abundance is almost invariant throughout a given genome, and thus the *profile of the relative abundances*, that is to say the set of the relative abundances for all the possible nucleotide pairs, is referred to as its *genomic signature*, specifically appropriate to be indicative of different organisms and/or class of organisms. Similar evaluations are available for characterizing the relative abundance of trinucleotides, tetranucleotides, and higher order oligonucleotides.

The first graphical approach to define a genomic signature using DNA sequences was introduced in 1990 by Jeffery [16] that represented the DNA sequence organization by using Chaos Game Representation (CGR) images. A CGR is plotted in a square whose four vertices are labelled by the nucleotides A, C, G, T, respectively; each point corresponds to one base of the sequence, and is plotted in the quadrant of the square labelled with that base. Positions in the quadrant depend on precedents bases. The major advantage of CGR is the use of a two-dimensional plot to provide a visual representation of primary DNA sequence organization for a sequence of any length, including whole genomes. Many considerations have been done on CGR method; in 1992 Hill et al. [13] proposed CGR images to compare and tabulate DNA sequences and in 1997, with Singh [14], compared and explored CGRs of mitochondrial genomes in order to classify specific whole genomes. On the contrary, in 1993 Goldman used the Markovian chain model to analyze the patterns shown in CGR images in order to demonstrate the unreliability of the method.

Since 1995, genomic signatures have been studied from a variety of perspectives, and many problems have been faced with its help. Recently, in [15, 20] a spectrum of genomic signatures, all based on frequencies of k-mers on whole genomes, has been proposed. In particular the FCGR, of k-order, is a matrix of frequencies extracted by the CGR image; namely, a $FCGR_k(s)$ is a $2^k \times 2^k$ matrix obtained as a grid on a CGR image of the sequence s. At last, Hsieh et al. in [15] guessed an universal whole genome signature composed by a pattern of duplications, based on words frequencies, and common to all genomes in the same classification.

2 Splines and B-Splines

In order to understand the mathematical background from which our method derives, we first briefly recall the notion of the so called Spline and B-spline functions. [9, 12, 19].

Definition 1. Given a strictly increasing sequence of real numbers, $x_1, x_2, \dots, x_n$, such that

$$a \equiv x_1 \le x_2 \le \cdots \le x_{n-1} \le x_n \equiv b \qquad \textbf{(1)}$$

a spline function $S_k(x)$ of degree *k,* with knots $x_1, x_2, \dots, x_n$, is defined in [a,b] according to the following two properties:

- In each interval (x_i, x_{i+1}) for i = 0,…,n $S_k(x)$ is given by some polynomial of degree *k* or less.

- $S_k(x)$ and its derivatives of orders *1,2,3,…,m-1* are continuous everywhere in [a,b].

Definition 2. A *normalized B-spline* of degree *k*, $B_{i,k+1}$, with respect to $x_i, \dots, x_{i-k+1}$ different nodes is defined as follows.

$$B_{i,k+1}(x) = (x_{i+k+1} - x_i)\, g[x_i, \dots, x_{i+k+1}] \qquad \textbf{(2)}$$

where the generic g(t) function values are:

$$g(t) = (t-x)_+^k \begin{cases} (t-x)^k & if\ x \le t \\ 0 & otherwise \end{cases} \qquad \textbf{(3)}$$

From (2) and (3) it results

$$B_{i,k+1}(x) = (x_{i+k+1} - x_i) \sum_{j=0}^{k+1} \frac{(x_{j+1} - x)_+^k}{\prod_{i=0}^{k+1} (x_{i+j} - x_{i+l})} \qquad \textbf{(4)}$$

A particular case of B-splines is defined in equidistant points $x_{i+1}=x_i+h$ for i=0,..,n-1. In this case the equation (4) becomes (5). [2,5,10].

$$6h^3 B_{i,4}(x) = \begin{cases} (x-x_i)^3 & if\ x \in [x_i, x_{i+1}] \\ h^3 + 3h^2(x-x_{i+1}) + 3h(x-x_{i+1})^2 - 3(x-x_{i+1})^3 & if\ x \in [x_{i+1}, x_{i+2}] \\ h^3 + 3h^2(x_{i+3}-x) + 3h(x_{i+3}-x)^2 - 3(x_{i+3}-x)^3 & if\ x \in [x_{i+2}, x_{i+3}] \\ (x_{i+4}-x)^3 & if\ x \in [x_{i+3}, x_{i+4}] \\ 0 & otherwise \end{cases} \qquad \textbf{(5)}$$

An analytical representation of the spline functions, used in the interpolation processes, can be obtained introducing *2·k* fictitious nodes

$$x_{-k} \le x_{-k+1} \le \ldots \le x_{-1} \le x_0 \equiv a, \quad b \equiv x_n \le x_{n+1} \le \cdots \le x_{n+k} \qquad \textbf{(6)}$$

associated to B-splines $B_{i,k+1}$ with *i = k,…,-1* and *i = n-k,…,n-1*. In this way, spline functions $S_k(x)$ are defined as follows.

$$s_k(x) = \sum_{i=-k}^{n-1} c_i B_{i,k+1}(x) \tag{7}$$

where real values c_i are the B-spline coefficients of s_k. and $B_{i,k+1}$ is calculated as in (5).

Splines constitute a class of piecewise polynomial functions satisfying certain conditions about the continuity of the function and its derivates. [12]. Splines are *kn.*

The linear systems arising in spline problems tend to be badly conditioned, and this may cause difficulties when attempting to solve these systems directly in order to obtain the required parameters. The numerical instability increases with the dimension of the linear system. B-splines have been introduced to overcome this problem. A B-spline is generally a basic spline function for which the approximation is defined as linear combination of them.

3 The Method

In this section we present a new method to explore genomic signatures using spline functions. The aim is to obtain the graphic representation of some functions which are technically derived by means of a fast interpolation process, but have the peculiar role of depicting the genomic signature of some organism. That is to say, the genomic signature can be represented by means of the analytical expression of a mathematical function, or a set of them, defined on the DNA genome sequence. More precisely, our method analyses 1 Kbs long sequences (1000 bases), calculates frequencies of trinucleotides and simply reports them on a Cartesian system with x-axes labeled by trinucleotides and y-axes labeled with the frequencies of each trinucleotide in the DNA genome sequence. The set of points just obtained in the coordinate plane is interpolated with spline functions and the result is a graphic 2D curve showing trinucleotide frequencies on a finite range. An analytical expression, or a set of them, can be derived from the linear systems arising in the spline interpolation, and its resolution with B-splines. Moreover, it can be used for simple mathematical, statistic and probabilistic analyses. Figure 1 and 2 show the genomic signatures for the Human beta-globin and the Rattus Norvegicus, calculated on 1 kbs long sequences.

Using equations (5) with *h=1*, and introducing *2k* fictitious nodes, the analytical expression of the genomic signature is given using equations (7).

4 Tests

Experiments on a phylogenetic tree have been done using the Hamming distance to test diversities between signatures. The results show that nodes at the same level in the tree have similar signatures and constant Euclidean and Hamming distances; whereas, nodes placed in different tree levels have different Euclidean and Hamming distances, moreover increasing with the distance in the tree. Our tests have been done by using *Matlab* applications. Fig 3 shows the phylogenetic tree used for testing.

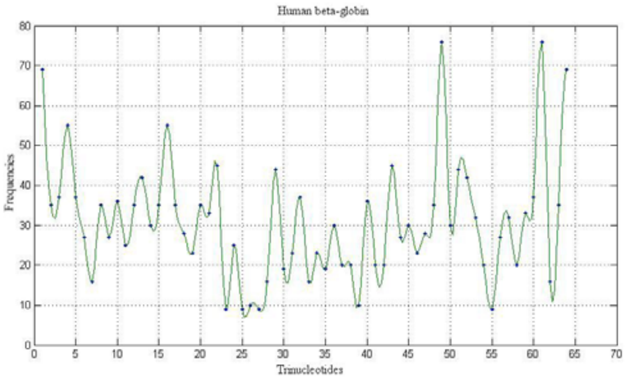

Fig. 1. The genomic signature of the Human beta-globin. The x-axes are labeled with the 64 trinucleotides ordered in lexicographic way; the y-axes represent the frequencies of each trinucleotide in the sequence. The interpolation by spline functions returns the function of the figure.

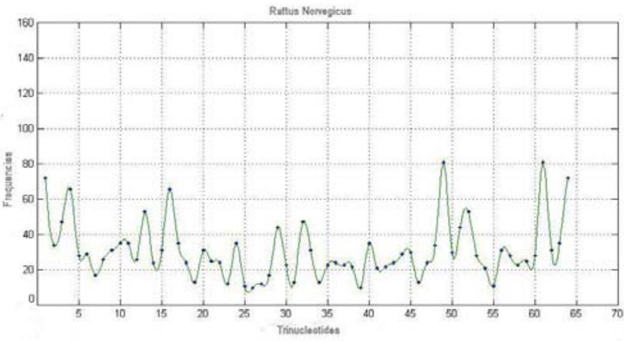

Fig. 2. The genomic signature of the Rattus Norvegicus. In this case frequencies are placed in a small range included among 5-80.

The following definitions can be useful in order to discuss the testing results.

Definition 3. Let $x_1, x_2, \ldots x_n$, be *n* values in a definite range, and *f(x)* and *g(x)* functions on this range; the *Euclidean distance* between *f(x)* and *g(x)* is a function *h(x)* defined as follows.

$$h(x)=\left|f(x)-g(x)\right| \quad \forall x=1,\ldots,n \tag{8}$$

Definition 4. Let $\alpha=\alpha_1,\ldots,\alpha_n$ and $\beta=\beta_1,\ldots,\beta_n$ be vectors of *n* elements; the *Hamming distance* between α and β is defined as follows.

$$H(\alpha,\beta)=\sum_{i=1}^{n}\left|\alpha_i-\beta_i\right| \tag{9}$$

Euclidean and Hamming distances allow us to make use of further measures to classify and compare our genomic signatures. As a matter of fact, valuable information has been carried out by simple statistics instruments such as the point to point mean, the standard deviation and the distance from the mean. Table 1 reports our test results for a group of organisms.

The tests we have done explored the phylogenetic tree of figure 3. For each node of it, we computed the genomic signature on a 1kbts sequence and compared the outcomes pair-wise. Figures 6-9 show the results. As long as the compared species get away with respect to each other in the tree levels, the outcomes clearly increase showing plain differences among their signatures. The graphic forms of the interpolating functions also exhibit valuable differences among species placed at different tree levels. Figures 4 and 5 are an example of the use of the Euclidean distance: nodes with similar signatures induce an Euclidean distance almost constant and near to zero; nodes with different signatures induce an Euclidean distance scarcely constant with many peaks far from zero. In the first case (figure 4) Balenoctera physalus and Balenoctera musculus have similar genomic characteristics and their signature are very similar and almost overlapping. In this case the Euclidean distance is constant and near to the zero. In the second case (figure 5) we compared genomic signatures of Prototheca wickerhamii and Drosophila yakuba, more distant in the tree levels. The result is a non constant function of the Euclidean distance, whose distance from the zero is valuable.

Eventually, it is also worthy noticing that values calculated on frequencies of trinucleotides between nodes increases when nodes move along the tree levels. This an important characteristic that allows to use this kind of representation for the genomic signature as a fast classification method or a genome recognition tool.

GENOMIC SIGNATURES IN STATISTIC COMPARISON

Compared Signatures	**HD**	**MEAN**	**SD**	**MD**
Balenoptera p. – Balenoptera m.	128	2	39,19	0,174
Balenoptera p. – Bos Taurus	236	3,687	2,833	0,354
Balenoptera p. – Phoca Vitulina	1412	22,062	27,87	3,484
Balenoptera p. – Didelphis v.	472	7	7	0,929
Balenoptera p. – Gellus gellus	584	9,125	7,757	0,969
Balenoptera p. – Ascaris suum	844	13,187	9,812	1,226
Balenoptera p. – Paramecium A.	1592	24,875	30,79	3,849
Phoca V. – Helicoerus g.	152	2,375	2,669	0,346
Rattus n. – Mus Miusculus	192	3	3,344	0,418
Rattus n. – Human	616	10	9,13	1,142
Cyprinus c. – Orossossoma l.	244	3,812	2,697	0,337
Cyprinus c. – Onchorhyncus m.	392	6,125	6,401	0,800
Cyprinus c. – Strongylacentrotus p.	764	11,937	14,02	1,752
Cyprinus c. – Ascaris suum	776	12,125	10,38	1,298
Cyprinus c. – Parameclum a.	1352	21,125	23,95	2,993
Drosophila y. – Anopheles g.	236	3,687	3,086	0,385
Drosophila y. – Apis mellifera l.	620	9,687	10,993	3,886
Drosophila y. – Artemia f.	836	13.063	14,195	1,774
Drosophila y. – Ascaris suum	844	13,188	16,682	2,085
Drosophila y. – paramecium a.	804	12,563	16,012	2,001
Strongylocent. p. – Paracentrotus l.	292	4,5625	5,8199	0,727
Podosopora a. – Schrizosacc. p.	972	15,118	14,74	1,842
Marchantia p. – Prototheca w.	1112	17375	14,724	1,840
Oncorhyncus m. – Ascaris suum	772	12,063	10,25	1,281
Oncorhyncus m. – Paramecium a.	1520	23,75	27,975	3,497

Table 1. Results of a group of comparisons done among the genomic signatures in the phylogenetic tree nodes. HD stands for Hamming Distance, MEAN for point-to-point Mean, SD for Standard Deviation and MD for Distance from the Mean

5 Conclusions

We have presented a simple and computationally easy procedure to detect the genomic signature in DNA sequences. Web services, applets or client-server can be implemented to return genomic signature of 1 Kbs sequences, or compare them. Moreover, any general purpose computer can effortlessly compute and display the genomic signature permitting its availability to a large range of users. As shown in [10], we can also use mathematical and numerical procedures on genomic signature in order to solve other important problems in bioinformatics such as the characterization of the horizontal transfer in genomes, or the mutation in species.

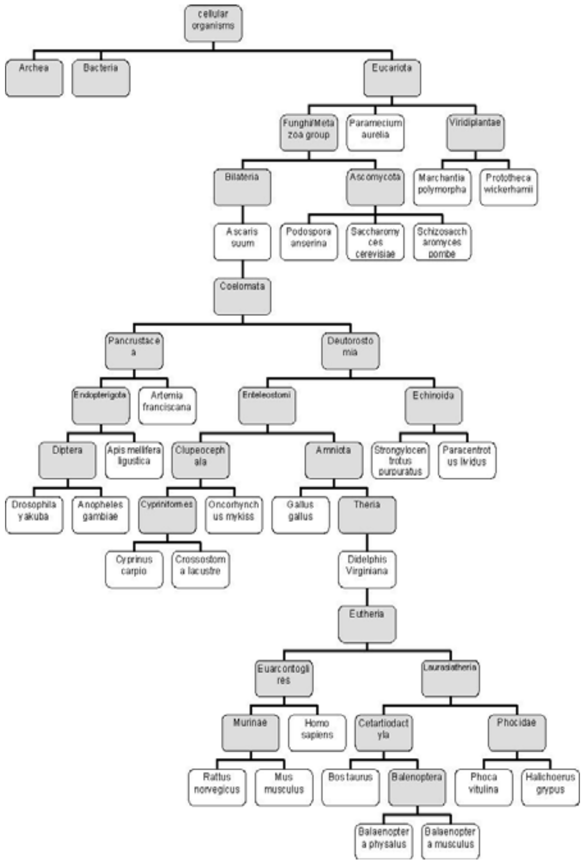

Fig. 3. The phylogenetic tree used for testing. Gray nodes represent the phylogenetic classification, and include sub-trees. White nodes represent the tested sequence.

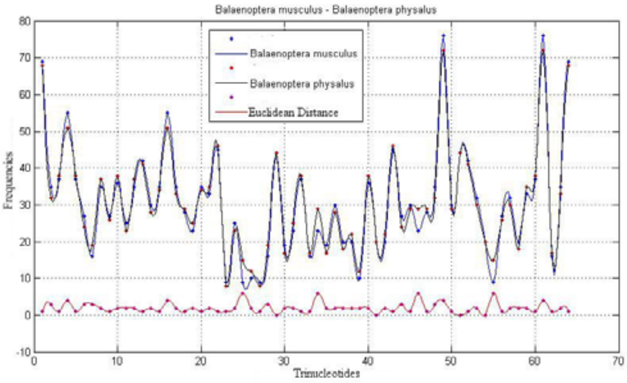

Fig. 4. Test between near nodes in the phylogenetic tree. Genomic signatures are similar, and the Euclidean distance is constant in a low level.

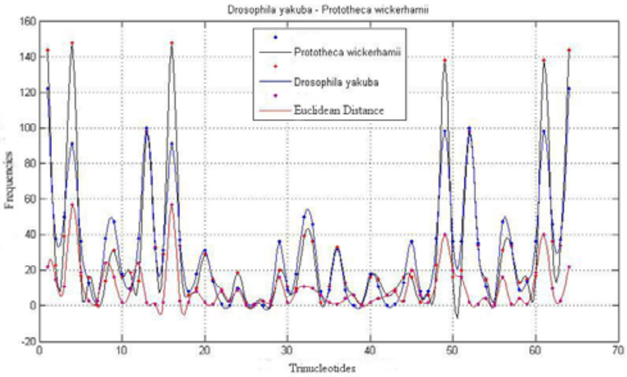

Fig. 5. Test between distant nodes in the phylogenetic tree. Genomic signatures are different ,and the Euclidean distance is not constant and far from zero.

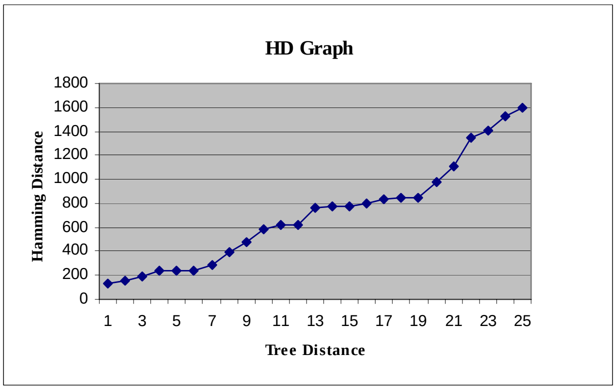

Fig. 6. Graphics for the variation of the Hamming distance along the phylogenetic tree levels. The Hamming distance increases as long as the node distance increases.

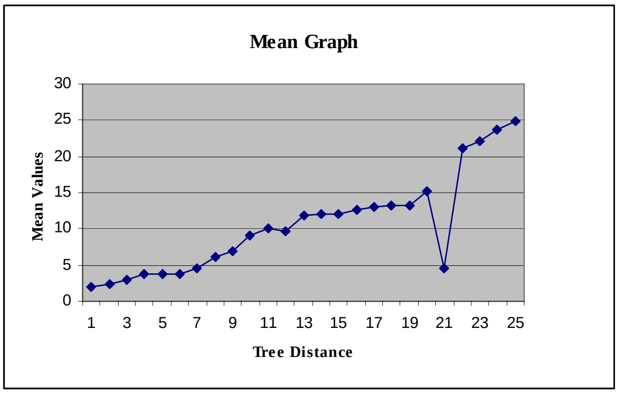

Fig. 7. Graphics for the mean variation along the phylogenetic tree levels.. The frequency value increases as long as the node distance increases.

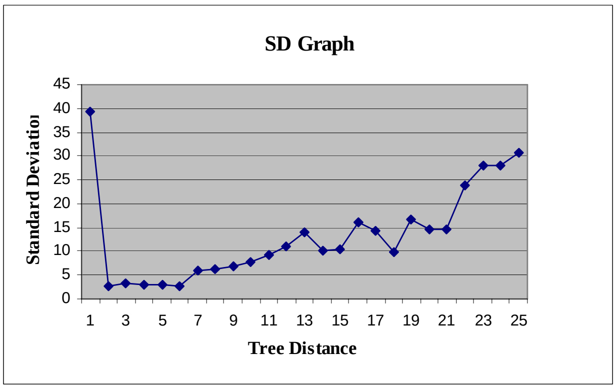

Fig. 8. Graphics for the variation of the Standard Deviation along the phylogenetic tree levels. The frequency mean value increases as long as the node distance increases.

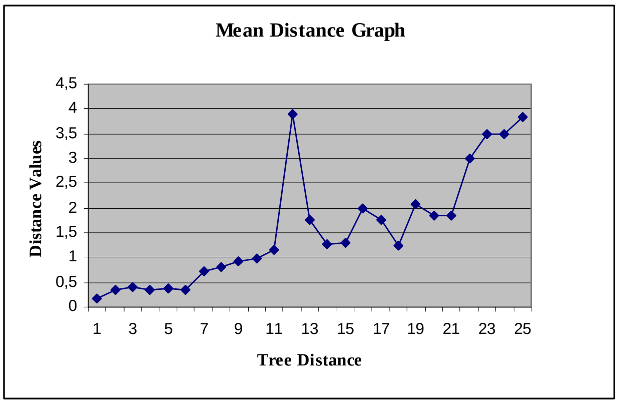

Fig. 9. Graphics for the variation of the Distance from the mean along the phylogenetic tree levels. The frequency mean value increases as long as the node distance increases.

6 References

[1] B.E. Blaisdell, A.M. Campbell, S. Karlin. Similarity and Dissimilarity of Phage Genomes. Proc. Natl. Acad. Sci. USA, Vol. 93, 1996, pp. 5854-5859.

[2] B.E. Blaisdell, S. Karlin, I. Ladunga. Heterogeneity of Genomes: Measures and Values. Proc. Natl. Acad. Sci. USA, Vol. 91, 1994, pp. 12837-12841

[3] L. Brocchieri, S. Karlin, J. Mrazek, A.M. Campbell, A.M. Spormann. A Chimeric Prokaryotic Ancestry of Mitochondria and Primitive Eukaryotes. Proc. Natl. Acad. Sci. USA, Vol. 96, 1999, pp. 9190-9195

[4] C. Burge, A.M. Campbell, S. Karlin. Over- and Under-representation of Short Oligonucleotides in DNA Sequences, Proc. Natl. Acad. Sci. USA, Vol. 89, 1992, pp. 1358-1362

[5] C. Burge, S. Karlin. Dinucleotide Relative Abundance Extremes: a Genomic Signature, Trends in Genetics, Vol. 11, 1995, pp. 283-290

[6] A.M. Campbell, S. Karlin, J. Mrazek. Compositional Biases of Bacterial Genomes and Evolutionary Implications. Journal of Bacter., Vol. 179, 1997, pp. 3899-3913

[7] A.M. Campbell, S. Karlin, J. Mrazek, Genome Signature Comparisons among Prokaryote, Plasmid, and Mitochondrial DNA, Proc. Natl. Acad. Sci. USA, Vol. 96, 1999, pp. 9184-9189

[8] P. Deschavanne, G. Dufraigne, B. Fertil, A. Giron and J. Vilian. Genomic Segnature is preserved in short DNA Fragments. Proc. In the IEEE International Symposium on Bioinformatics and Biomedical Engineering, pp. 161-167, 8-10 November, 2000.

[9] A. Dold, B. Eckmann. Spline Functions. Lecture Notes in Mathematics, Karlsruhe. Springer Verlag Ed. 1975.

[10] C. Dufraigne, B. Fertil, S. Lespinants, Alain Giron and P. Deshavenne. Detection and characterization of horizontal transfers in prokaryotes using genomic signature. Nucleic Acid Research, Vol. 33, n. 1, 2005.

[11] N. Goldman. Nucleotide, Dinucleotide and Trinucleotide Frequencies Explain Patterns Observed in Chaos Game Representation of DNA Sequences. Nucleic Acid Research, Vol. 21, pp. 2487-2491, 1993.

[12] T. N. E. Greville. Theory and Applications of Spline Functions. Academic Press, 1969.

[13] K. A. Hill, N. J. Schislerh, S. M. Singh. Chaos Game Representation of coding regions of human globin genes and alcohol dehydrogenase genes of phylogenetically divergent species. Journal of Molecular Evolution, Vol. 35, pp. 261-269, 1992.

[14] K. A. Hill, S.M. Singh. The evolution of species-type specificity in the global DNA sequence organization of gmitochondrial genomes. Genome, Vol. 40, pp. 342-356, 1997.

[15] L. Hsieh, T. Chen, C. Chang and H.C. Lee. A Universal Signature in Whole Genomes. Proc. In IEEE Computer System Bioinformatics (CBS04), pp. 20-30, 2004.

[16] H. Jeffery. Chaos Game Representation of Gene Structure. Nucleic Acid Research. 18, pp. 2163-2170, 1990

[17] J. Josse, A.D. Kaiser, A. Kornberg. Enz. Synth. of Deoxyrib. Acid-VIII.Frequencies of Nearest Neighbor Base Sequences in Deoxyrib.Acid. Journal of Biol. Chem., Vol. 236, 1961, pp. 864-875.

[18] S. Karlin, I. Ladunga. Comparison of Eukaryotic Genomic sequences. Proc. Natl. Acad. Sci. USA, Vol. 91, 1994 , pp. 12832-12836

[19] H.R. Schwartz. Numerical Analysis, a comprehensive introduction. John Wiley and Sons Ed. 1989.

[20] Y. Wang, K. Hill, S. Sing and L. Kari. The spectrum of Genomic Signature: from Nucleotides to Chaos Game Representation, 2005.