The Spectrum of Genomic Signatures: from Dinucleotides to Chaos Game Representation

Yingwei Wang[1], Kathleen Hill [2], Shiva Singh [2], Lila Kari [1]
Department of Computer Science [1], Department of Biology [2]
University of Western Ontario, London, Ontario, N6A 5B7, Canada

September 14, 2004

Abstract

In the post genomic era, access to complete genome sequence data for numerous diverse species has triggered multiple avenues for examining and comparing primary DNA sequence organization of entire genomes. Previously, the concept of a genomic signature was introduced with the observation of species-type specific Dinucleotide Relative Abundance Profiles (DRAPs); dinucleotides were identified as the subsequence with the greatest bias in representation in a majority of genomes. Herein, we demonstrate that DRAP is one particular genomic signature contained within a broader spectrum of signatures. In this spectrum, an alternative genomic signature, Chaos Game Representation (CGR), provides a unique visualization of patterns in sequence organization. A genomic signature is associated with a particular integer order or subsequence length that represents a measure of the resolution or granularity in the analysis of primary DNA sequence organization. We quantitatively explore the organizational information provided by genomic signatures of different orders through different distance measures, including a novel Image Distance. The Image Distance and other existing distance measures are evaluated by comparing the phylogenetic trees they generate for 26 complete mitochondrial genomes from a diversity of species. The phylogenetic tree generated by the Image Distance is compatible with the known relatedness of species. Quantitative evaluation of the spectrum of genomic signatures may be used to ultimately gain insight into the determinants and biological relevance of the genome signatures.

1 Introduction

Although efforts are continuously being made toward understanding the characteristics of genomes, any DNA sequence in a genome is too long and too complex for a person to directly comprehend its characteristics. In 1990, Jeffrey proposed using Chaos Game Representation (CGR) to visualize DNA primary sequence organization (Jeffrey 1990). A CGR is plotted in a square, the four vertices of which are labelled by the nucleotides A, C, G, T, respectively. The plotting procedure can be described by the following steps: the first nucleotide of the sequence is plotted halfway between the centre of the square and the vertex representing this nucleotide; successive nucleotides in the sequence are plotted halfway between the previous plotted point and the vertex representing the nucleotide being plotted. 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 entire genomes.

CGRs of DNA sequences show interesting patterns. Various geometric patterns, such as parallel lines, squares, rectangles, and triangles can be found in CGRs. Some of the CGRs even show a complex fractal geometrical pattern which is very similar to the Sierpinsky Triangle (Mandelbrot 1982). These interesting features relevant to the DNA sequence organization attracted further research in CGR (Dutta & Das 1992, Hill, Schisler & Singh 1992, Oliver, Bernaola-Galvan, Guerrero-Garcia & Roman-Raldan 1993).

In 1993, Goldman analyzed the patterns shown in CGRs and concluded that "it is unlikely that CGRs can be more useful than simple evaluation of nucleotide, dinucleotide and trinucleotide frequencies"(Goldman 1993). According to this conclusion, CGR should be relegated to the status of a pictorial representation of nucleotide, dinucleotide and trinucleotide frequencies.

After this sobering conclusion, research on CGRs still continued but with less frequency. (Hill & Singh 1997) compared CGRs of mitochondrial genomes and explored the evolution of species-type specificity in DNA sequences. (Almeida, Carrico, Maretzek, Noble & Fletcher 2001) suggested that CGR is a generalization of Markov Chain probability tables that accommodates non-integer orders.

In parallel to CGR research, Karlin and Burge proposed the concept of genomic signature (Karlin & Burge 1995). The key observation behind the genomic signature concept is that dinucleotide relative abundance profiles (DRAPs)

of different DNA sequence samples from the same organism are generally much more similar to each other than to sequences from other organisms, and that closely related organisms generally have more similar DRAPs than distantly related organisms. It was concluded from these observations that the DRAP values constitute a genomic signature of an organism.

Since 1995, research on genomic signatures has been done from different perspectives (Karlin, Mrazek & Campbell 1997, Campbell, Mrazek & Karlin 1999, Deschavanne, Giron, Vilain, Fagot & Fertil 1999, Deschavanne, Giron, Vilain, Vaury & Fertil 2000, Gentles & Karlin 2001, Sandberg, Winberg, Branden, Kaske, Ernberg & Coster 2001, Edwards, Fertil, Giron & Deschavanne 2002, Hao, Lee & Zhang 2000). (Campbell et al. 1999) compared genomic signatures among prokaryote, plasmid, and mitochondrial DNA. (Deschavanne et al. 2000) showed that word usage in short fragments of genomic DNA (as short as 1 kb) is similar to that of the whole genome, thus providing a strong support to the concept of genomic signature. (Gentles & Karlin 2001) looked for the genomic signature of various eukaryotes. (Sandberg et al. 2001) proposed a method to classify sequence segments using genomic signatures. More recently, genomic signatures were used in phylogenetic analysis (Edwards et al. 2002).

In 1999, an interesting paper provided a link between CGRs and genomic signatures (Deschavanne et al. 1999). Experiments showed that variation between CGR images along a genome was smaller than variation among genomes. "These facts strongly support the concept of genomic signature and qualify the CGR representation as a powerful tool to unveil it."(Deschavanne et al. 1999)

In this paper, we discuss CGR and DRAP from the following perspectives: In Section 2 we challenge the idea that CGR is merely a representation of nucleotide, dinucleotide, and trinucleotide frequencies. The aim of Section 2 is to provide evidence supporting the claim that CGRs have more complex features worth further investigation. Section 3 shows that DRAP is a special case of CGR and cannot reproduce a CGR; based on this observation, we propose the idea of a spectrum of genomic signatures, and describe the common features and variations of different genomic signatures. Section 4 discusses various distance definitions between genomic signatures of two DNA sequences. *Order* is an integer number associated with a genomic signature to describe the granularity of this genomic signature. In Section 5 we design an experiment to quantitatively analyze the information provided by the genomic signatures of different orders

of a given DNA sequence. Section 6 presents our major conclusions.

2 What Determines the Pattern in a CGR?

The interesting patterns in CGRs inspired exploration of the underlying determinants of these patterns in different ways. (Hill et al. 1992) tried to use image analysis techniques to categorize and analyze CGRs. (Goldman 1993) used Markov Chain model simulation to explore these determinants.

(Goldman 1993) concluded that "the CGR gives no further insight into the structure of the DNA sequence than is given by the dinucleotide and trinucleotide frequencies" and "unless more complex patterns are found in CGRs, there is no justification for ascribing their patterns to anything other than the effects described in this paper." These conclusions had the effect that CGRs have subsequently been much less studied from this perspective. Actually, not much research on CGRs was done after (Goldman 1993) was published. In this section we first present arguments supporting our claim that CGRs give more insight into DNA structures than those given by nucleotide, dinucleotide, and trinucleotide frequencies, and then present our answer to the question "What determines the pattern in a CGR?"

2.1 Short Nucleotide Frequencies Cannot Solely Determine the Pattern in a CGR

The results reported in (Goldman 1993) are obtained through DNA sequence simulation based on Markov Chain model. We first briefly introduce the first-order and second-order Markov Chain model. In the first-order Markov Chain model, successive bases in a simulated sequence depend only on the preceding base. A 4×4 matrix P defines the probabilities with which subsequent bases follow the current base in a DNA sequence. If the base labels A, C, G, and T are equated with the numbers 1, 2, 3, and 4, then P_{ij}, the jth element of the ith row of P, defines the probability that base j follows base i. The row-sums of P must equal 1. Using this matrix, a simulated DNA sequence is obtained by selecting a first base randomly, according to the frequencies of the bases in the DNA under study; if this is base i, then the probabilities P_{i1}, P_{i2}, P_{i3}, and P_{i4} are used to select the next base, and so on until the simulated sequence is

of the same length as the original DNA sequence.

In the second-order Markov Chain model, each base depends on the previous two bases. All probabilities in the form of P_{XYZ}, which is the probability that base Z follows the dinucleotide XY, are used to simulate the original sequence.

The major result in (Goldman 1993) can be described as follows: For a DNA sequence a, we may construct a simulated sequence a' using the first-order Markov Chain model, such that a and a' have the same length and the same nucleotide and dinucleotide frequencies. We would then find that the CGRs of a and a' have similar patterns, suggesting the hypothesis that the patterns in a CGR are mainly determined by the nucleotide and dinucleotide frequencies of the sequence. If the CGRs constructed from a and a' are not similar, we can construct another simulated sequence a'' using the second-order Markov Chain model, such that a and a'' have the same length and the same nucleotide, dinucleotide, and trinucleotide frequencies. We would then find that the CGRs of a and a'' have similar patterns, suggesting the hypothesis that the patterns in a CGR are completely determined by the nucleotide, dinucleotide, and trinucleotide frequencies of the sequence.

Figure 1 shows counterexamples to the conclusion in (Goldman 1993). Six CGRs are presented in Figure 1. In the left column, the three CGRs are plotted from human DNA sequence, human mtDNA sequence, and *Neurospora crassa* DNA sequence (GenBank accession number: U01317, J01415, and AABX01000061). In the right column, the three CGRs are plotted from sequences constructed by simulating the length, the single-nucleotide, dinucleotide, and trinucleotide frequencies of the corresponding sequences in the left column. The striking point supporting our claim is that these two columns of CGRs are not similar at all.

How could this happen? The reason is that the sequences shown in the right column are constructed by simulating the sequences shown in the left column using an algorithm other than the Markov Chain model. While we do not go into the details of the algorithm that constructed these sequences, we use the following example to illustrate the sequences constructed by this algorithm. Suppose we have sequences X, Y, and Z, where Sequence X is a segment of the human genome and Sequence Y and Z are constructed by simulating X.

X: GATCACAGGTCTATCACCCTATTAACCACTCACGGGAGCT

Y: AAAAAAAAAAACCCCCCCCCCCCCGGGGGGGTTTTTTTTT

Z: AACACACACACAGAGATATATCCCGCTCTCTCTCGGGGTT

We can verify that Sequence Y has the same single nucleotide frequencies

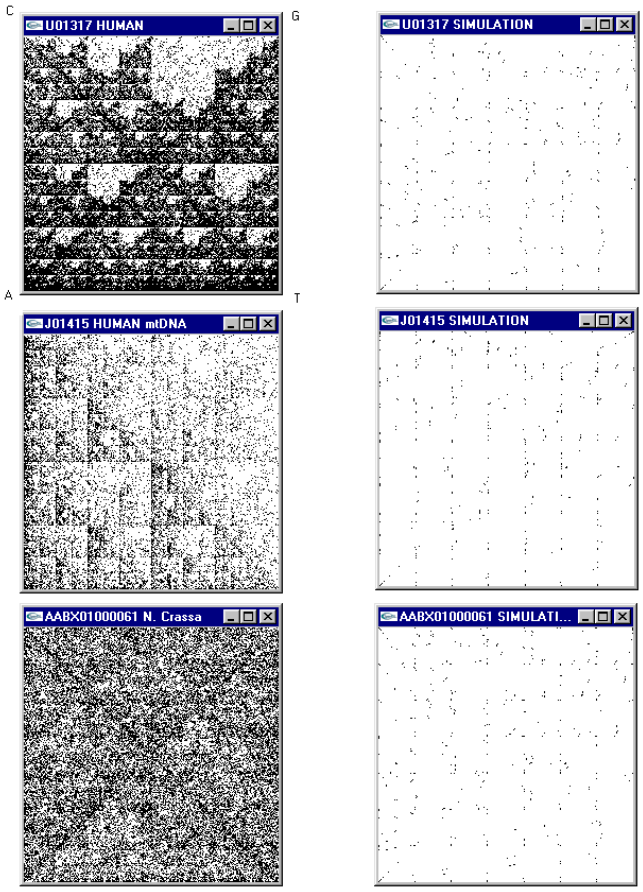

Figure 1: In each row, the two CGRs correspond to sequences with same lengths, same single-nucleotide, dinucleotide, and trinucleotide frequencies. The CGR in the left column corresponds to a naturally existing DNA sequence (human DNA sequence, human mtDNA sequence, and *Neurospora crassa* DNA sequence, from top to bottom). The CGR in the right column is artificially constructed to have a high degree of non-randomness. As a result, a CGR in the right column looks sparse because many points in the plot coincide.

as Sequence X. Sequence Z has both the same single nucleotide frequencies and similar dinucleotide frequencies as Sequence X. We also notice that the nucleotides in Sequence Y and Sequence Z are highly non-randomly arranged.

Because of this non-randomness, the CGRs of Sequence Y and Sequence Z have many points that coincide so that these CGRs "look" very sparse.

Although the sequences shown in the right column of Figure 1 are constructed for illustration purposes, and such sequences of this length (more than 16,000 bases) can hardly exist in natural DNA genomes, Figure 1 clearly shows that a sequence's nucleotide, dinucleotide, and trinucleotide frequencies cannot solely determine its CGR pattern.

2.2 What Determines the Pattern in a CGR?

Figure 1 shows that a CGR contains more information than nucleotide, dinucleotide, and trinucleotide frequencies. The next question is what determines the pattern in a CGR?

Suppose a DNA sequence has been plotted in a CGR, and the size of the CGR is 1×1. We divide the CGR by a $2^k \times 2^k$ grid. According to the characteristics of a CGR, each grid square corresponds to a length k oligonucleotide, and the number of points within a grid square is the number of occurrences of the corresponding length k oligonucleotide in this DNA sequence. The length $k-1$ oligonucleotide at the beginning of the DNA sequence corresponds to points on the grid lines instead of inside any grid square. If the DNA sequence is much longer than k, we can omit these $k-1$ points.

A practical CGR media, such as paper, screen, or memory space, always has a limited resolution. If the resolution of a CGR is $\frac{1}{2^k}$ and we divide the CGR by a $2^k \times 2^k$ grid, we cannot distinguish points inside a same grid square. We may paint a whole grid square with a gray scale to express the number of points in the grid square. The more points in the grid square, the darker the shade of grey.

To conclude, if a CGR's resolution is $\frac{1}{2^k}$ and the DNA sequence is much longer than k, this CGR is completely determined by all the length k oligonucleotide occurrences or frequencies. These frequencies reveal all information in a CGR.

The above discussion suggests another method of plotting a CGR: Suppose we want to plot a DNA sequence in a CGR on a media, and the resolution of this media is $\frac{1}{2^k}$. In other words, the image area in this media is divided into $2^k \times 2^k$ squares. We first count the number of occurrences for each length k oligonucleotide in this DNA sequence. Then we paint each square with a proper

shade of gray according to the number of occurrences of the corresponding length k oligonucleotide in this DNA sequence.

Consequently, if $k = 3$, a CGR is totally determined by nucleotide, dinucleotide, trinucleotide frequencies, and the conclusions in (Goldman 1993) are correct.

However, if $k > 3$, a CGR cannot be totally determined by nucleotide, dinucleotide, and trinucleotide frequencies. Longer oligonucleotide frequencies may influence the pattern of the CGR. Figure 1 shows that sometimes longer oligonucleotide frequencies may have a big impact on the CGR pattern.

Now let us consider (Goldman 1993) again: The contribution of this paper was that it introduced the Markov Chain model to simulate a sequence. The CGRs of the simulated sequences indicate that nucleotide, dinucleotide and trinucleotide frequencies are able to produce the patterns in CGRs. However, while it is true that nucleotide, dinucleotide and trinucleotide frequencies are able to produce the patterns in CGRs, it was erroneous to conclude that these frequencies can solely determine the patterns in CGRs. Finally, we believe that, under a different interpretation, the Markov Chain model can still prove to be useful, and we will use the Markov Chain model again in Section 5.

3 A Spectrum of Genomic Signatures

3.1 DRAP and FCGR (the Frequency matrix extracted from a CGR)

Both DRAP and CGR have been proposed as genomic signatures. Before we examine the relationship between these two kinds of genomic signatures, we first discuss their definitions to clarify any ambiguity.

For a sequence s, the dinucleotide relative abundance profile DRAP(s) is an array $\{\rho_{XY} = f_{XY}/f_X f_Y\}$, XY stands for all possible dinucleotide combinations, where f_X denotes the frequency of the mononucleotide X in s and f_{XY} the frequency of the dinucleotide XY in s. We use s' to denote the reverse complement of s. DRAP(ss') is the dinucleotide relative abundance profile computed from the sequence s concatenated with its reverse complement s'.

CGRs, in their original form, are not convenient to be stored and processed in a computer. Thus we introduce another form of CGR: FCGR (the Fre-

quency matrix extracted from a CGR). The structure of FCGR was introduced in (Deschavanne et al. 1999) and the name FCGR was proposed in (Almeida et al. 2001). For convenience, in this paper we introduce the concept of order of a FCGR and define a kth-order FCGR as follows.

A kth-order FCGR for sequence s, denoted as $FCGR_k(s)$, is a $2^k \times 2^k$ matrix. To obtain this FCGR, we first plot a CGR from s, then divide this CGR by a $2^k \times 2^k$ grid so that each grid square corresponds to an element in the matrix, count the number of points *inside* each grid square, and use the number of points as the matrix element corresponding to the grid square. We do not count those points on the grid square lines because they represent the length $k-1$ oligonucleotide at the beginning of the DNA sequence, and we can omit these $k-1$ points as long as the DNA sequence is much longer than k. Note that, instead of being a graphical representation, like CGR, a FCGR is a numerical matrix.

A FCGR can also be constructed directly from a sequence instead of plotting a CGR first and then converting the CGR into a FCGR. We can construct a FCGR directly by counting the number of occurrences of each length k oligonucleotide in the sequence and putting this number into the appropriate place of the FCGR matrix according to the correspondence between a length k oligonucleotide and a CGR grid square.

A first-order FCGR and a second-order FCGR are shown as follows, where N_w is the number of occurrences of the oligonucleotide w in sequence s.

$$FCGR_1(s) = \begin{pmatrix} N_C & N_G \\ N_A & N_T \end{pmatrix}$$

$$FCGR_2(s) = \begin{pmatrix} N_{CC} & N_{GC} & N_{CG} & N_{GG} \\ N_{AC} & N_{TC} & N_{AG} & N_{TG} \\ N_{CA} & N_{GA} & N_{CT} & N_{GT} \\ N_{AA} & N_{TA} & N_{AT} & N_{TT} \end{pmatrix}$$

The definition of $FCGR_{k+1}(s)$ can be obtained by replacing each element N_X in $FCGR_k(s)$ with 4 elements

$$\begin{matrix} N_{CX} & N_{GX} \\ N_{AX} & N_{TX} \end{matrix}$$

A kth-order FCGR can also be applied to a sequence concatenated with its reverse complement, and such a FCGR is described as $FCGR_k(ss')$.

If we have a CGR of resolution $\frac{1}{2^k}$, we can transform the gray shade of each square into a number, and obtain a kth-order FCGR; If we have a kth-order FCGR, we can plot a CGR of resolution $\frac{1}{2^k}$ according to the method introduced in Section 2. In conclusion, a kth-order FCGR is equivalent to a CGR of resolution $\frac{1}{2^k}$.

A kth-order FCGR contains 4^k numbers of length k oligonucleotide occurrences. If $k = 5$, 4^k is over one thousand; if $k = 10$, 4^k is over one million. It is almost impossible for a person to comprehend so many numbers at one time. Because a kth-order FCGR is equivalent to a CGR of resolution $\frac{1}{2^k}$, one can comprehend the major features present in the large number of elements in the FCGR matrix by visually checking the equivalent CGR of resolution $\frac{1}{2^k}$.

Finally, the relationship between DRAPs and 2nd-order FCGRs is that DRAPs are deducible from 2nd-order FCGRs but 2nd-order FCGRs are not deducible from DRAPs.

3.2 A Spectrum of Genome Signatures

Both second-order FCGR and DRAP have 16 elements, and each element corresponds to one dinucleotide. An element of a second-order FCGR involves only the frequency of the dinucleotide. On the contrary, as the name "relative abundance" suggests, an element of a DRAP is the ratio of the dinucleotide frequency to the frequencies of the two single nucleotides composing this dinucleotide. We call an element of a DRAP a *relative frequency*. Thus we may call DRAP a second-order *relative FCGR*, and denote it as $\mathrm{rFCGR}_2(s)$.

$$rFCGR_2(s) = \begin{pmatrix} \rho_{CC} & \rho_{GC} & \rho_{CG} & \rho_{GG} \\ \rho_{AC} & \rho_{TC} & \rho_{AG} & \rho_{TG} \\ \rho_{CA} & \rho_{GA} & \rho_{CT} & \rho_{GT} \\ \rho_{AA} & \rho_{TA} & \rho_{AT} & \rho_{TT} \end{pmatrix}$$

Normally, in a DRAP all the elements are organized in an array; in a second-order relative FCGR the same elements are organized in a matrix. Organizing the elements of a DRAP as a second-order relative FCGR not only reveals the similarity of a DRAP and a FCGR, but also enables us to define a *kth-order relative FCGR* ($k \geq 2$).

By expanding the definition of a DRAP, we can define a trinucleotide relative abundance profile as $\{\rho_{XYZ} = f_{XYZ}/\ f_X f_Y f_Z\}$ for all trinucleotide XYZ, and further define even longer oligonucleotide relative abundance profiles in the same way. Similarly, we can express a trinucleotide relative abundance profile as a 3rd-order relative FCGR. In the general situation, we can express a length k oligonucleotide relative abundance profile as a kth-order relative FCGR. These relative FCGRs can also be visualized in a CGR.

Based on these discussions, we propose the idea that various kinds of genomic signatures exist, and they can be considered as members of a spectrum of genomic signatures. All genomic signatures in the spectrum have some common features: each genomic signature is a numerical matrix and can be visualized in a CGR; a positive integer number called *order* determines its granularity; if the order is k, the numerical matrix has $2^k \times 2^k$ elements. Each element in the matrix is mapped to a length k oligonucleotide as described in the definition of FCGR.

For now, we know that the spectrum of genomic signatures contains two major categories: FCGR and relative FCGR. Note that a second-order relative FCGR is equivalent to a DRAP. Besides choosing FCGR or relative FCGR, other choices also determine the variations in the spectrum of genomic signatures.

One choice is about whether we concatenate the DNA sequence with its reverse complement before we count the frequencies. By concatenating a DNA sequence with its complement we eliminate some unnecessary biases in the sequence. In the original definition of a DRAP, a DNA sequence is always concatenated with its reverse complement before the frequencies are counted, but this is not the only choice. When we construct a FCGR or a relative FCGR, we can either concatenate the DNA sequence with its reverse complement or not do so.

Another issue concerns the standardization of FCGRs. The sum of all elements in a FCGR is proportional to the length of the DNA sequence in question. Because a genomic signature is supposed to be only associated with a species, we need to eliminate the factor of DNA sequence length from FCGRs so that we are able to compare the FCGRs obtained from DNA sequences of different lengths and further explore the usefulness of FCGRs. The procedure of eliminating the sequence length factor from the FCGR definition, called herein standardization, will be described in the sequel.

Suppose A is a kth-order FCGR. We know A is a $2^k \times 2^k$ matrix, and we use $a_{i,j}$ $(1 \leq i, j \leq 2^k)$ to denote the elements in this matrix. We can standardize A, and the standardized FCGR, denoted by $\overline{A}$, is defined as follows:

$$\overline{A} = \frac{4^k}{\sum_i \sum_j a_{i,j}} * A$$

Suppose the elements in $\overline{A}$ are denoted as $b_{i,j}$, we have the following property:

$$\sum_{i=1}^{2^k} \sum_{j=1}^{2^k} b_{i,j} = 4^k$$

This property means that in a standardized kth-order FCGR the sum of all elements is the number of elements so that the average value of the elements of the matrix is 1.

Standardized FCGRs make sense when different organisms are involved, but non-standardized FCGRs are still useful in some cases. For example, we can plot a CGR of resolution $\frac{1}{2^k}$ from a non-standardized kth-order FCGR, but we cannot plot a CGR from a standardized kth-order FCGR.

In this section, we have presented the common features and variations inside the spectrum of genomic signatures. We conjecture that different genomic signatures could be used for different purposes.

4 The Distance Between Genomic Signatures of Two DNA Sequences

A distance between the genomic signatures of two DNA sequences is an important measure in evaluating the difference between them. Such a distance measure can be used in phylogenetic analysis.

There are many different ways to define such a distance measure. Different distance definitions have different features and can serve different purposes.

4.1 Geometric Distances

According to geometry, we can define the Euclid distance and Hamming distance between two points in a multi-dimensional space. We may consider a $2^k \times 2^k$ matrix as a point in a 4^k-dimensional space, and define Euclid distance and Hamming distance as follows:

Definition 1 (Euclid Distance) *Let us assume that matrices* $\overline{A} = (a)_{2^k \times 2^k}$ *and* $\overline{B} = (b)_{2^k \times 2^k}$ *are standardized kth-order FCGRs. We define the Euclid distance between* $\overline{A}$ *and* $\overline{B}$ *as:*

$$dE(\overline{A}, \overline{B}) = \frac{\sqrt{2^k}}{4^k} * \sqrt{\sum_{i=1}^{2^k} \sum_{j=1}^{2^k} (a_{i,j} - b_{i,j})^2}$$

The constant $\frac{\sqrt{2^k}}{4^k}$ is used to adapt the standard Euclid distance so that the distance values for different k values will be in the same range. We omit the detailed discussion about why this constant should be $\frac{\sqrt{2^k}}{4^k}$; briefly speaking, this constant is related to the definition of standardized FCGR.

Definition 2 (Hamming Distance) *Let us assume that matrices* $\overline{A} = (a)_{2^k \times 2^k}$ *and* $\overline{B} = (b)_{2^k \times 2^k}$ *are standardized kth-order FCGRs. We define the Hamming distance between* $\overline{A}$ *and* $\overline{B}$ *as:*

$$dH(\overline{A}, \overline{B}) = \frac{1}{4^k} * \sum_{i=1}^{2^k} \sum_{j=1}^{2^k} |a_{i,j} - b_{i,j}|$$

(Campbell et al. 1999) suggested that for two sequences f and g the distance between two DRAPs should be defined as

$$\delta(f, g) = 1/16 \sum_{XY} |\rho_{XY}(f) - \rho_{XY}(g)|$$

Actually, this DRAP distance is a special case of Hamming distance.

As we observed, there are several ways in which one can define the distance between two FCGRs. One of the desirable properties of such a distance would be that the distance between two DNA sequences of the same genome should be small. The Hamming distance satisfies this property for DRAP and low-order FCGRs ($k \leq 3$). However, for higher-order ($k > 7$) FCGRs, the Hamming distance could be pretty big, close to the maximum distance value, because longer oligonucleotide frequencies are not always stable. In these cases, the Hamming distance lost its discrimination power.

We define thus another distance to measure the similarity between two FCGRs according to the image similarity between the two CGRs visualized from these FCGRs. This distance is called *Image distance*, and it is an expansion of the Hamming distance.

The following definitions are needed to define Image distance.

Definition 3 (Neighborhood) *Suppose we have a matrix $A = (a)_{n\times n}$, a positive integer R, and a pair (i, j) where $1 \leq i, j \leq n$. A neighborhood of radius R, centered at (i, j), denoted as $\Theta_R(i, j)$, consists of all integer pairs (s, t), where $1 \leq s, t \leq n$, $s \in [i - R, i + R]$, and $t \in [j - R, j + R]$.*

Definition 4 (Density) *For a matrix $A = (a)_{n\times n}$, we define a density matrix $(density_R(A))_{n\times n}$, where for any (i, j), $1 \leq i, j \leq n$*

$$density_R(A)_{i,j} = \frac{\sum_{(s,t)\in\ \Theta_R(i,j)} a_{s,t}}{\sum_{(s,t)\in\ \Theta_R(i,j)} 1}$$

Now we define the Image distance between two FCGRs:

Definition 5 (Image Distance) *Let us assume that matrices $\overline{A} = (a)_{2^k\times 2^k}$ and $\overline{B} = (b)_{2^k\times 2^k}$ are standardized kth-order FCGRs. We define the Image distance between $\overline{A}$ and $\overline{B}$ as:*

$$dI_R(\overline{A}, \overline{B}) = \frac{1}{4^k} * \sum_{i=1}^{2^k}\sum_{j=1}^{2^k} \left|density_R(\overline{A})_{i,j} - density_R(\overline{B})_{i,j}\right|$$

In the above definition, the Image distance is related to the neighborhood radius R. If $R = 0$, $density_R(A)_{i,j} = a_{i,j}$; $dI_0(A, B)$ is the Hamming Distance of two FCGRs. If $R > 0$, the frequency values within the neighborhood are averaged in the distance computation. The value of R can be chosen by a trial-and-error method so that the distance values fit for a specific purpose.

The following formula gives an upper bound of the maximal Image distance value for a specific R, but we omit the proof of this formula:

$$dI_R(\overline{A}, \overline{B}) \leq 2 \times \left(\frac{2R + 1}{R + 1}\right)^2.$$

4.2 Statistical Distance

According to statistics, the distance between two sets of samples can be obtained by calculating the Pearson's correlation coefficient between these two sets of samples. We may consider a $2^k \times 2^k$ matrix as a set of samples, and define a distance measure accordingly. (Almeida et al. 2001) suggested that the distance between two FCGRs could be based on a slightly modified Pearson's correlation coefficient.

Definition 6 (Pearson Distance) *Suppose FCGR A is expressed as an array* x_i $(1 \leq i \leq n)$*, and FCGR B is expressed as an array* y_i $(1 \leq i \leq n)$*. The Pearson distance* $dP(A, B)$ *can be defined through the following steps:*

$$
\begin{aligned}
nw &= \sum_{i=1}^{n} x_i \cdot y_i \qquad \overline{x}w = \frac{\sum_{i=1}^{n} x_i^2 \cdot y_i}{nw} \qquad \overline{y}w = \frac{\sum_{i=1}^{n} x_i \cdot y_i^2}{nw} \\
sx &= \frac{\sum_{i=1}^{n} (x_i - \overline{x}w)^2 \cdot x_i \cdot y_i}{nw} \qquad sy = \frac{\sum_{i=1}^{n} (y_i - \overline{y}w)^2 \cdot x_i \cdot y_i}{nw} \\
rw_{x,y} &= \frac{\sum_{i=1}^{n} \frac{x_i - \overline{x}w}{\sqrt{sx}} \cdot \frac{y_i - \overline{y}w}{\sqrt{sy}} \cdot x_i \cdot y_i}{nw} \qquad dP(A, B) = 1 - rw_{x,y}
\end{aligned}
$$

One advantage of the Pearson distance is that we do not need to standardize FCGR matrices before the distance calculation.

4.3 Evaluation of Distances

It is difficult to conclude that one distance measure is better than the others because different distance definitions have different features and can serve different purposes. Here we just evaluate these distance definitions from a specific perspective: generating phylogenetic trees.

We chose 26 mitochondrial DNA sequences; each sequence is from a specific organism. These sequences are described in Table 1.

We construct a 10th-order FCGR for each sequence. For each distance definition we calculate the distances among all these FCGRs. We notice that for Hamming distance, most of the distance values are close to the maximum Hamming distance value, showing that when the order is 10 the Hamming distance cannot properly discriminate FCGRs. Thus we do not use the Hamming distance in the next step.

After excluding the Hamming distance, we use the software package PHYLIP to generate phylogenetic trees from the distance matrices corresponding to the Euclid distance, Image distance, and Pearson distance, respectively.

Due to space limitation, we omit the distance data and just show the phylogenetic trees obtained by this method in Figure 2 (a), (b), and (c). As a term of reference, we also use CLUSTALW to directly generate a phylogenetic tree for

Table 1: Mitochondrial DNA sequences description

Accession No.	Organism	Common name
X15917	*Paramecium aurelia*	protozoa
M61734	*Podospora anserina*	fungus
U02970	*Prototheca wickerhamii*	alga
X54421	*Schizosaccharomyces pombe*	yeast A
M62622	*Saccharomyces cerevisiae*	yeast B
M68929	*Marchantia polymorpha*	plant
X54253	*Ascaris suum*	roundworm
X69067	*Artemia franciscana*	shrimp
J04815	*Paracentrotus lividus*	urchin A
X12631	*Strongylocentrotus purpuratus*	urchin B
X03240	*Drosophila yakuba*	fruit fly
L06178	*Apis mellifera*	honeybee
L20934	*Anopheles gambiae*	mosquito
X52392	*Gallus gallus*	chicken
X61010	*Cyprinus carpio*	carp
M91245	*Crossostoma lacustre*	loach
L29771	*Oncorhynchus mykiss*	trout
Z29573	*Didelphis virginiana*	opossum
X61145	*Balaenoptera physalus*	whale A
X72204	*Balaenoptera musculus*	whale B
X72004	*Halichoerus grypus*	seal A
X63726	*Phoca vitulina*	seal B
J01394	*Bos taurus*	cow
X14848	*Rattus norvegicus*	rat
V00711	*Mus musculus*	mouse
J01415	*Homo sapiens*	human

these 26 sequences, shown in Figure 2 (d). By checking the phylogenetic trees shown in these figures, we notice the following:

1. The tree generated from Euclid distances distinguishes all vertebrates from other organisms. This tree also distinguishes all invertebrates from other organisms with the exception of yeast B.

2. The tree generated from Image distances distinguishes all vertebrates from other organisms, but the invertebrates and other organisms are mixed together.

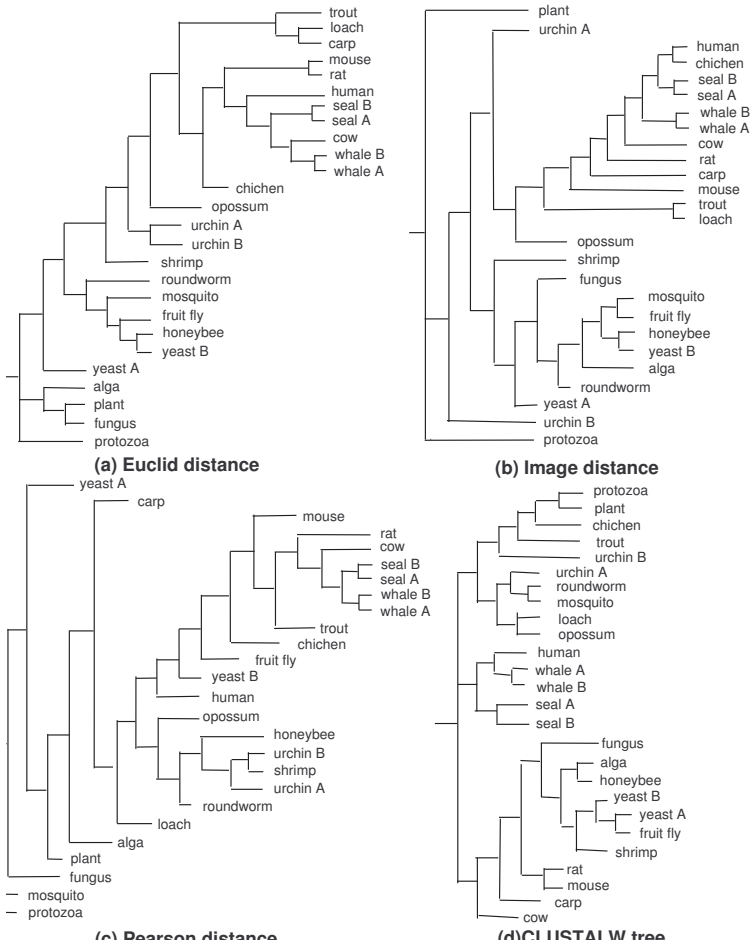

Figure 2: The phylogenetic trees constructed from the Euclid distances, Image distances, and Pearson distances among FCGRs of the 26 mitochondrial DNA sequences, and the phylogenetic tree constructed by CLUSTALW directly from the same 26 sequences.

3. The tree generated from Pearson distances distinguishes most vertebrates from other organisms with the exception of four organisms; invertebrates and other organisms are mixed together.

4. The tree generated by CLUSTALW cannot distinguish vertebrates, inver-

tebrates, and other organisms. These major categories are mixed together.

5. The trees generated from Euclid distances, Image distances, and Pearson distances are more compatible with the phylogenetic relatedness of the species than the tree generated by CLUSTALW. This result shows FCGRs must contain major phylogenetic information and a genomic signature study is a promising approach in phylogenetic analysis.

6. Among the trees generated from the three kinds of distances, the tree generated from the Euclid distance is the one most compatible with known relatedness of species. This result suggests that Euclid distance portrays a species type specific sequence structure.

5 Quantitative Analysis for Genomic Signatures of Different Orders of a DNA Sequence

In Section 3, we proposed the concept of a spectrum of genomic signatures. This section discusses the comparison of the genomic signatures of a same species within a spectrum.

Logically, lower order FCGRs are deducible from higher order FCGRs, while higher order FCGRs are not deducible from lower order FCGRs. Empirically, although in Secton 4 we only showed the phylogenetic trees generated by the highest (10th) order FCGR in our scope, in our experiments we found that the phylogenetic trees generated using higher order FCGRs are more compatible with known relatedness of species. The experiments reported in (Qi, Wang & Hao 2004) confirmed this result.

Even though the above evidences support the hypothesis that higher order FCGR are preferable, no definite conclusion can be drawn. Generating phylogenetic trees is only one among many possible applications of genomic signatures. It may well be that higher order FCGRs do not perform well in other application areas. In addition, with the increase of the order k, the number of elements in

kth-order FCGR matrix is increasing exponentially. The exponential increase of FCGR matrix elements not only increases the time cost and space cost during data processing, but also reduces the conciseness of the genomic signature in the sense that it makes the values of FCGR elements less comprehensible to human beings.

In the remainder of this section we propose a quantitative method of exploring the relationship between different order FCGRs of DNA sequences of the same species. The purpose of this method is to provide additional insight into what may be the "optimal" order of FCGR to be used in various practical applications.

The information gain provided by the use of a $(k+1)$th-order FCGR as opposed to a kth-order FCGR may not be always the same for different values of k. We investigate in the following this variation in information gain. The goal is to choose a value for k that is as small as possible (to minimize computational costs), while maximizing at the same time the information amount provided by the respective kth-order FCGR.

DRAP has been proposed as a genomic signature (Karlin & Burge 1995), suggesting that dinucleotide frequencies contain the major information about genomic organization; nucleotide, dinucleotide, and trinucleotide frequencies were considered as the sole determinants of a CGR pattern (Goldman 1993), suggesting that short oligonucleotide frequencies contain major information about genomic organization.

On the other hand, in Section 2 we brought forth arguments supporting the claim that short oligonucleotide frequencies cannot, in fact, totally determine the patterns in a CGR.

We believe that these two conclusions are not contradictory. Instead, they are compatible with each other because they describe different aspects of the relationship between CGR patterns and oligonucleotide frequencies. Although short oligo-nucleotide frequencies cannot totally determine a CGR's pattern, they are able to determine the major patterns in a CGR. We shall now attempt to bring forth an experiment that supports this hypothesis. Ideally, we should

compare FCGRs of different orders of a same DNA sequence and show that the higher the order, the smaller the distance between two consecutive FCGRs becomes. This would prove that the additional information obtained by increasing the order (granularity) becomes gradually smaller, and the highest order only brings information about the "details" of the genome organization.

The problem is that we cannot calculate the distance between FCGRs of different orders. Thus we design an experiment to calculate a measure which is similar to the distance between FCGRs of different orders in the range from 1 to 10. The 26 mitochondrial DNA sequences described in Table 1 are used in this experiment.

An explanation is in order regarding to our choice of the range for the values of k, namely from 1 to 10. The fact that the kth order FCGR matrix becomes sparse with the increase of k would suggest that the use of high values of k is not advisable. The value $k = 10$ is an upper bound, empirically established. Indeed, our experiments show that as long as the DNA sequence is long enough (for example, longer than 10,000 base pairs), its (k)th-order FCGR matrix is not very sparse for values of k that are less than 10. In addition, if a DNA sequence is very short (for example, shorter than 10,000 base pairs), we cannot obtain a stable genomic signature regardless of the value of k. This made $k = 10$ a good empirical choice for the order of the FCGR.

The procedure of this experiment is as follows: For each sequence and each number k between 1 and 10, we construct a simulated sequence which has the same length and the same (or very similar) kth-order FCGR with the original sequence using the $(k-1)$th-order Markov Chain model, in which each base depends on the previous $(k-1)$ bases. After the simulated sequence has been constructed, we calculate the Image distance defined in Section 4 between the 10th-order FCGR of the original sequence and the 10th-order FCGR of the simulated sequence. By experience, we choose $R = 20$ for Image distance calculation between 10th-order FCGRs.

Now we formally describe the above procedure. We use $L(s)$ to denote the length of sequence s; use $sim(A, L)$ to denote the length L sequence constructed

by simulating FCGR A. For a specific sequence s and an integer k, we calculate the following Image distance:

$$dI_{20}(FCGR_{10}(s), FCGR_{10}(sim(FCGR_k(s), L(s)))) * 1000 \quad (1)$$

By multiplying with 1000, we need only deal with integer numbers instead of decimal numbers. According to Section 4, the upper bound of the above distance is 7,624.

Why do we think this distance defined in (1) describes the difference between a kth-order FCGR and a 10th-order FCGR? This is a distance between two sequences, one being the original sequence, and the other being the one constructed by (k-1)th-order Markov Chain model to simulate the kth-order FCGR of the original sequence. Thus we only need to make sure that this simulated sequence can represent the kth-order FCGR.

According to the discussion in Section 2, we can obtain a same kth-order FCGR from many different sequences, and these sequences could be quite different. It is hard to say which sequence can represent the kth-order FCGR.

Because we are trying to compare the information provided by FCGRs of different orders, we wish we could find a sequence that does not contain more information than the kth-order FCGR does. The sequence constructed using the $(k-1)$th-order Markov Chain model is suitable for this purpose. The simulated sequence is constructed as randomly as possible with only one restriction: the kth-order FCGR of the simulated sequence is the same with the kth-order FCGR of the original sequence. Any special arrangement of the sequence on top of this restriction will influence the higher-than-k-order frequencies of this sequence, and thus not desirable.

The above discussion explained why we design the experiment in this way. Now let us consider what we expect from this experiment.

When $k = 10$, if we can perfectly simulate the original sequence, the new sequence should have the exactly same 10th-order FCGR so that the distance defined in (1) should be 0. Because the simulation techniques are imperfect, this Image distance could be a small amount other than 0.

When k goes from 1 to 10, we expect that the distance defined in (1) decreases rapidly, showing that the short oligonucleotide frequencies provide the major organizational information of a sequence. Table 2 gives the result of this experiment. For each k value, we have a group of 26 distances corresponding to the 26 DNA sequences. To describe the general tendency and variation of these data groups, statistical measures, such as average and standard deviation, can be used. In this experiment, because we only concern those distances that are larger than the average distance, we use the difference between the maximal distance and the average distance instead of the standard deviation to describe the variation. These two measures, the average distance and the difference between the maximal distance and the average distance, describe the general status of all the 26 distances. If the average distance is small and the difference between the maximal distance and the average distance is also small, we are sure that the kth-order FCGR and the 10th-order FCGR are uniformly similar for all DNA sequences. Figure 3 figuratively shows the average distance and the difference between the maximal distance and the average distance shown in the last two lines of Table 2.

From Table 2 and Figure 3 we we observe that:

1. In the *average* row of Table 2, the values drop from 284 to 221 when k goes from 1 to 2. When k continues to increase from 2 to 10, the values in this row decreases much slower. This observation suggests that a 2nd-order FCGR is at an optimal point in terms of "performance/cost" ratio: k is very small while the information amount provided by the FCGR is relatively large. If an application requires thus a "concise" genomic signature from the spectrum (a genomic signature whose matrix has a small number of elements), a 2nd-order FCGR is a reasonable choice.

2. In the $k = 1$ column of Table 2, some values are much larger than the average of this column; the difference between the maximal in this column and the average in this column is 191. In the $k = 2$ column, this difference drastically drops to 49. This observation suggests that in the $k = 2$

Table 2: FCGR Image distances between the original sequences and the new sequences simulated at different orders

GenBank	order k									
Accession No.	1	2	3	4	5	6	7	8	9	10
X15917	387	245	195	126	87	60	40	23	9	5
M61734	197	133	111	79	53	37	24	15	8	4
U02970	202	159	135	91	69	46	30	18	8	5
X54421	282	224	171	142	109	75	49	27	9	8
M62622	388	239	182	112	65	39	21	11	5	3
M68929	220	115	87	66	46	32	21	14	7	2
X54253	275	213	182	153	106	72	44	23	11	8
X69067	272	214	191	163	127	90	56	30	12	7
J04815	352	245	211	170	129	93	59	31	14	6
X12631	359	231	213	177	133	94	60	31	10	8
X03240	262	216	188	148	104	70	42	21	10	7
L06178	242	192	156	112	86	56	33	17	8	4
L20934	276	203	178	150	104	74	44	22	13	7
X52392	292	242	207	161	121	86	56	28	13	9
X61010	289	261	216	177	129	90	59	29	13	9
M91245	310	244	223	183	131	92	60	32	12	6
L29771	288	219	213	174	127	92	61	29	9	12
Z29573	272	222	198	164	119	82	52	28	10	5
X61145	259	237	210	173	124	89	57	31	12	10
X72204	275	239	217	166	125	88	56	33	16	6
X72004	284	270	221	179	127	87	58	32	14	16
X63726	287	258	212	185	129	90	57	30	10	10
J01394	275	210	198	163	128	88	57	29	13	11
X14848	278	240	204	163	120	85	56	29	17	9
V00711	266	245	204	163	122	86	54	28	11	8
J01415	289	231	208	174	125	89	56	29	14	6
average	284	221	190	151	109	76	49	26	11	7
max-average	191	49	33	32	24	18	12	7	6	9

column the values are uniformly smaller. When k continues to increase, the difference between the maximal and the average drops much slowly. This observation shows that in terms of variation, a second-order FCGR is also at an optimal point where k is very small while the variation is also very small.

3. The values in Table 2 enable us to evaluate the similarity between CGRs

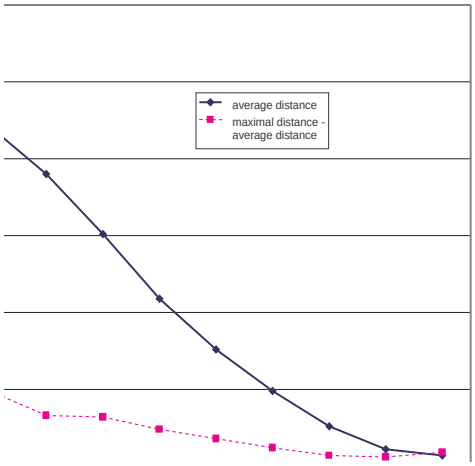

Figure 3: The solid line represents the average distance between original and simulated sequences of different orders from Table 2; the dotted line represents the difference between the maximal distance and the average distance in the same table.

without visual checking. We visually checked all CGR images involved and verified the following regularity: if the distance amount is less than 220, the two CGR images are similar in major patterns; if the distance amount is greater than 320, the two CGR images have different major patterns; if the distance amount is less than or equal to 320 and greater than or equal to 220, the two CGR images may or may not have major pattern differences.

4. If the simulation technique were ideal, when $k = 10$ the distances values would be 0s. In this experiment, when $k = 10$ the distance values are not 0s. These small distance values are noise caused by the imperfect simulation technique. When $k = 9$, the noise is also very strong so the distance values are not reliable. Due to the noise, for some sequences the distance value when $k = 10$ is even greater than the distance value when

$k = 9$.

5. A second-order FCGR is a much better choice than a first-order FCGR to serve as a genomic signature. Using higher order FCGRs will further reduce the distance amount shown in Table 2. For example, using a 5th-order FCGR, the average distance shown in Table 2 is further reduced by half (from 221 to 109), but the price is that the number of elements in the FCGR matrix increases from 16 to 1024. For modern computers, the time cost and space cost for 1024 elements are not an issue. However, for a human observer a 16-element matrix and a 1024-element matrix have completely different stories. A human observer may be able to check the 16 elements of a 2nd-order FCGR and interpret their meanings as occurrences of dinucleotide DNA sequences, but this observer is unable to do the same thing for a 1024-element matrix. A 5th-order FCGR does provide more information than a 2nd-order FCGR, but the extra information is not large enough given that the price is a significant loss in the "conciseness" of the FCGR. The user may tradeoff among these factors according to the concrete application at hand..

To conclude, a higher-order FCGR describes a DNA sequence more precisely, but more computational cost is needed because the number of elements in a FCGR matrix increases exponentially.

6 Conclusion

In this paper we propose a spectrum of genomic signatures and discuss different aspects of this idea.

First, we challenge the idea that CGR is merely a graphical representation of nucleotide, dinucleotide, and trinucleotide frequencies. Our counterexamples show that nucleotide, dinucleotide, and trinucleotide frequencies cannot totally determine the patterns in a CGR. Then we reveal the underlying determinants of CGR Patterns: if a CGR's resolution is $\frac{1}{2^k}$ and the DNA sequence is much

longer than k, this CGR is completely determined by all the numbers of length k oligonucleotide occurrences.

Secondly, based on the observation that DRAP and CGR are related, we propose the idea that all genomic signatures can be considered as members of a spectrum. All genomic signatures in this spectrum have common features, and each kind of genomic signature in this spectrum has its own characteristics.

Thirdly, we discuss various distance definitions between genomic signatures of two DNA sequences, and define Image distance to measure the pattern differences between two FCGRs. A distance between genomic signatures of two DNA sequences reflects the difference between the two organisms. The distance can be used in phylogenetic analysis and other applications.

Fourthly, we quantitatively analyze the information provided by the genomic signatures of different orders of a given DNA sequence with an experiment based on Image distance. This experiment shows that a 2nd-order FCGR is at an optimal point regarding the choice of order in the following sense: This genomic signature has a small number of elements, while the information amount it provides is relatively large. If we want to find a "concise" genomic signature with small number of matrix elements, a 2nd-order FCGR seems thus a reasonable choice.

In conclusion, we explore the relationship between different genomic signatures and propose that these genomic signatures can be considered as members of a spectrum of genomic signatures. We quantitatively analyze genomic signatures from the information gain perspective. Further topics of exploration include the role of the Image Distance in constructing phylogenetic trees, especially in determining the divergence time, as well as the possible use of genomic signatures in describing features of various taxonomic categories.

We thank the Editor, Prof. Allan Campbell, and the anonymous referee for their comments which greatly contributed to the clarity of our paper.

References

Almeida, J. S., Carrico, J. A., Maretzek, A., Noble, P. A. & Fletcher, M. (2001), 'Analysis of genomic sequences by chaos game representation', *Bioinformatics* **17**, 429–437.

Campbell, A., Mrazek, J. & Karlin, S. (1999), 'Genome signature comparisons among prokaryote, plasmid, and mitochondrial DNA', *Proc. Natl. Acad. Sci. USA* **96**, 9184–9189.

Deschavanne, P., Giron, A., Vilain, J., Vaury, A. & Fertil, B. (2000), Genomic signature is preserved in short DNA fragments, *in* 'IEEE International Symposium on Bioinformatics and Biomedical Engineering (BIBE'00)', pp. 161–167.

Deschavanne, P. J., Giron, A., Vilain, J., Fagot, G. & Fertil, B. (1999), 'Genomic signuture: Characterization and classification of species assessed by chaos game representation of sequences', *Mol. Biol. Evol.* **16**, 1391–1399.

Dutta, C. & Das, J. (1992), 'Mathematical characterization of chaos game representation', *Journal of Molecular Biology* **228**, 715–719.

Edwards, S. V., Fertil, B., Giron, A. & Deschavanne, P. J. (2002), 'A genomic schism in birds revealed by phylogenetic analysis of DNA stirngs', *Systematic Biology* **51**, 599–613.

Gentles, A. J. & Karlin, S. (2001), 'Genome-scale compositional comparisons in eukaryotes', *Genome Research* **11**, 540–546.

Goldman, N. (1993), 'Nucleotide, dinucleotide and trinucleotide frequencies explain patterns observed in chaos game representations of DNA sequences', *Nucleic Acids Research* **21**, 2487–2491.

Hao, B., Lee, H. & Zhang, S. (2000), 'Fractals related to long dna sequences and complete genomes', *Chaos, Solutons and Fractals* **11**, 825–836.

Hill, K. A. & Singh, S. M. (1997), 'The evolution of species-type specificity in the global DNA sequence organization of mitochondrial genomes', *Genome* **40**, 342–356.

Hill, K. A., Schisler, N. J. & Singh, S. M. (1992), 'Chaos game representation of coding regions of human globin genes and alcohol dehydrogenase genes of phylogenetically divergent species', *Journal of Molecular Evolution* **35**, 261–269.

Jeffrey, H. J. (1990), 'Chaos game representation of gene structure', *Nucleic Acids Research* **18**, 2163–2170.

Karlin, S. & Burge, C. (1995), 'Dinucleotide relative abundance extremes: a genomic signature', *Trends in Genetics* **11**, 283–290.

Karlin, S., Mrazek, J. & Campbell, A. M. (1997), 'Compositional biases of bacterial genomes and evolutionary implications', *Journal of Bacteriology* **179**, 3899–3913.

Mandelbrot, B. (1982), *The Fractal Geometry of Nature (2nd edition)*, W. H. Freeman and Co., San Francisco, California.

Oliver, J. L., Bernaola-Galvan, P., Guerrero-Garcia, J. & Roman-Raldan, R. (1993), 'Entropic profiles of DNA sequences through chaos-game-derived images', *Journal of Theoretical Biology* **160**, 457–470.

Qi, J., Wang, B. & Hao, B. (2004), 'Whole proteome prokaryote phylogeny without sequence alignment: A k-string composition approach', *Journal of Molecular Evolution* **58**, 1–11.

Sandberg, R., Winberg, G., Branden, C., Kaske, A., Ernberg, I. & Coster, J. (2001), 'Capturing whole-genome characteristics in short sequences using a naive bayesian classifier', *Genome Research* **11**, 1404–1409.