Chapter 1

Introduction to Chaos Game Representation

1.1 Introduction

Computational biology deals with the use of mathematical tools to extract useful information from biological data. Representative problems in computational biology range from the assembly of high-quality DNA sequences from fragmentary ‘shotgun’ DNA sequencing to the prediction of gene regulation with data from mRNA micro-arrays and protein chips. Although efforts are continuously being made towards understanding the characteristics of genomes, any particular genome is too long and too complex for a person to directly comprehend its characteristics. This chapter gives an introduction to one such mathematical technique called Chaos Game Representation (CGR). CGR was originally proposed as a scale-independent representation for genomic sequences by Jeffrey in 1990 (Jeffrey, 1990). The technique, formally an iterative function system, can be traced further back to the foundations of statistical mechanics, in particular to Chaos theory (Bar-Yam, 1997).
1.2 Introduction to Chaos Game

The Chaos Game is an algorithm which produces pictures of fractal structures. In mathematics, the term chaos game, as coined by Michael Barnsley (1988), originally refers to a method of creating a fractal, using a polygon and a random point inside it. In a simple form, it proceeds as follows.

1. Plot three non-collinear points on a paper. Label the points as A, B and C.

2. Plot another point anywhere on the plane. This is the current point.

3. Now take a six sided die and roll it. If the number which appears on top is 1 or 2, then plot a point mid-way of the current point and A. If the number is 3 or 4, then plot the same towards B and if the number is 5 or 6 then plot the same towards C. The point which you have last plotted is the current point.

4. Again roll the die and repeat step 3, where the current point is the point is the point you have plotted last.

If these steps are repeated many times, one might expect a paper covered with random dots or perhaps a triangle filled with random dots. Such is not the case. What we obtain is seen in the figure, a triangle filled with a sequence of smaller and further smaller triangles. This figure is called the ‘Sierpinski Gasket’ after the mathematician who first defined it.
Introduction to Chaos Game Representation

Figure 1.1 - Sierpinski’s gasket

With four initial points the result is different. We will not obtain squares inside squares. What we obtain is a square uniformly filled with points.

Mathematically the chaos game is represented by an Iterated Function System (IFS). IFS is a finite collection of mappings $F_i: X \rightarrow X$ defined on a metric space $X$, with

$$x_i(n) = \sum_i F(x_i(n-1)).$$

Each equation gives the formula for computing the new values of $x_i$.

1.3 Chaos Game Representation of DNA sequences

The DNA sequence is composed of four nucleotides Adenine (A), Guanine (G), Cytosine (C) and Thymine (T). For our use, it is treated formally as a simple string comprising of the characters A, T, G and C. Suppose that each nucleotide is assigned a point as follows; A is (0, 0), T is (1, 0), G is (1, 1) and C is (0, 1). Given a DNA sequence, it can be visually represented in a CGR as follows. Plot the initial point in the centre of the square (0.5, 0.5) formed by the four points A, T, G and C as
Introduction to Chaos Game Representation

its vertices. The first nucleotide in the sequence is considered. Plot a point exactly midway between the current point and the vertex corresponding to the nucleotide. For the next nucleotide, take this second point as the current point and repeat the same procedure. i.e. The CGR of a nucleotide at position i of a sequence is exactly halfway between the previous point and the vertex corresponding to the present nucleotide.

Mathematically it is represented by an Iterated Function System, here a pair of linear equations defined by,

\[
\begin{align*}
    x_i &= 0.5(x_{i-1} + g_x(i)) \\
    y_i &= 0.5(y_{i-1} + g_y(i))
\end{align*}
\]

where \( g_x(i) \) is the x coordinate of the vertex corresponding to the nucleotide at position i and \( g_y(i) \) is the y coordinate of that vertex.

As an illustration consider the sequence ATGCGAGTGT
Figure 1.2 - CGR of ATGCGAGTGT
But if we continue plotting the same way for a genome region, the resulting figure is not a square filled with random dots. The CGRs for the complete genome of the bacterium Clostridium Tetani E88 and the mitochondrial genome of the plant Arabidopsis Thaliana is given as illustrations. Observe that the uniformly filled square for random probabilities strongly contrasts with the apparent structure displayed by the CGR for the DNA sequences. Also notice the difference between the patterns formed.

Figure 1.3 - Clostridium Tetani E88 - Complete Genome
Earlier we have shown that plotting points randomly in a square using Chaos Game, give a square randomly filled with dots, that is without any particular patterns. But plotting DNA sequences using CGR show visible patterns in the picture. What we see is the attractor formed by the iterated function system. The pictures have a complex structure which varies depending on the input sequence. H.J. Jeffrey (1990) proposed this method and visualised the patterns of different sequences. Intuitively, non-randomness in the picture corresponds to non-randomness in the sequence. It implies that the nucleotide sequences are following some kind of rule. Jeffrey noted that a pattern in one part of the picture was repeated in many places, but in varying magnitudes. The CGR thus exhibits the property of self-similarity which is very important in the study of fractals and chaotic dynamics. He noticed that there is a one to one correspondence between the sequence and the points in the CGR. Hence any visible pattern in the CGR corresponds to some pattern in the sequence of bases. It is
to be noted that adjacent nucleotides in the sequence may not be plotted adjacent to each other. He observed that the visible patterns represent global as well as local patterns in the sequence.

1.4 Frequency Chaos Game Representation

Jeffrey (1990) has observed that each point in the CGR corresponds to exactly one subsequence (starting from the first base). Though he mentioned the one to one correspondence, he did not give a method to reconstruct the sequence from the CGR. The point in a CGR corresponding to one base of a sequence is plotted in the quadrant of the square labelled with that base. This is because each quadrant comprises all points that are halfway between one corner and any other point within the square. Conversely, all points plotted within a quadrant must correspond to subsequences of the DNA sequence that end with the base labelling the corner of that quadrant. For example, any base G gives rise to a point in the G (upper-right) quadrant of the square; and every point in that quadrant corresponds to a base G in the DNA sequence. This association between points and subsequences continues recursively to sub-quadrants, sub-sub-quadrants etc.

A correspondence between the subsequence and the CGR points is described as follows. In a CGR whose side is of length 1, two sequences with suffix of length ‘k’, are contained within the square with side length $2^{-k}$. i.e. For a particular k-mer, its CGR co-ordinates will always be contained in a specific square with side length $2^{-k}$.

In the figure shown below, the CGR point which appears bold in the lower right quadrant is used to trace the sequence backwards. Since the point lies in the lower right quadrant the last nucleotide in the sequence is T. Subdivide this quadrant
Introduction to Chaos Game Representation

into four and we can see that our point lies in the upper right quadrant which stands for G. i.e. our sequence ends with GT. Further subdividing the current quadrant into four will make the point fall into the lower right quadrant which stands for A. Here we have divided the CGR square four times to obtain the last four nucleotides which is GAGT. This can be extended to any further resolution as desired. In theory, it is possible to reconstruct the entire sequence from the first base.

Figure 1.5 - Resolving a CGR

Another way to look at the above picture is that, whenever the pattern ‘GAGT’ appears in the sequence, a dot will be plotted somewhere in the corresponding square. That is, counting the number of points in the corresponding square will give the frequency of appearance of ‘GAGT’ in the sequence. Thus counting the CGR points in the squares of a $2^k \times 2^k$ grid gives the number of occurrences of all possible k-mers.
in the sequence. This type of representation is called a Frequency Chaos Game Representation (FCGR). The structure of FCGR was introduced by Deschavanne et al. (1999) and the name FCGR was proposed by Almeida et al. (2001). Note that those points on the grid square lines are not to be counted because they represent the length $k-1$ oligonucleotide at the beginning of the DNA sequence. These $k-1$ points can be omitted as long as the DNA sequence is much longer than $k$.

![CGR and FCGR of order 2](image)

<p>| | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>53</td>
<td>44</td>
<td>29</td>
<td>31</td>
</tr>
<tr>
<td>72</td>
<td>28</td>
<td>44</td>
<td>46</td>
</tr>
<tr>
<td>60</td>
<td>59</td>
<td>74</td>
<td>83</td>
</tr>
<tr>
<td>91</td>
<td>90</td>
<td>92</td>
<td>101</td>
</tr>
</tbody>
</table>

$\times$ 1000 bp

**Figure 1.6 - CGR and FCGR of order 2**

Figure 1.6 shows a CGR and its corresponding FCGR of order 2. Here the CGR is divided into $2^2 \times 2^2$ squares and the number of points in each square is written in the corresponding FCGR matrix. The values in this FCGR gives various dimer frequencies (multiplied by 1000) of the given sequence. It is also possible to calculate oligonucleotide frequencies of non-integer lengths by resolving the CGR using grids of sizes other than powers of two. (Almeida et al, 2001)
Thus CGR, which was primarily meant only to be a visualization technique of nucleotide sequences, was shown to give rise to a fast algorithm for computing oligonucleotide frequencies of any length. Instead of being a graphical representation like CGR, an FCGR is a numerical matrix. The method, thus provides a graphical representation as well as a storage tool.

1.5 Review of CGR in sequence analysis

After its introduction in 1990, the potential of CGR to analyze sequences generated much interest among researchers. The observation, that the visible patterns in CGR represent global as well as local patterns in the sequence, was relevant to the DNA sequence organization. This attracted immediate further research (Basu et al, 1992; Hill et al, 1992 and Oliver et al, 1993). Hill et al. examined the CGRs of coding sequences of 7 human globin genes and 29 relatively conserved alcohol dehydrogenase genes from phylogenetically divergent species. The results showed that, CGRs of human globin cDNAs were similar to one another and to the entire human globin gene complex. Moreover, Adh CGRs were similar for genes of the same or closely related species but were different for relatively conserved Adh genes from distantly related species. The paper suggested that dinucleotide frequencies may account for the self-similar pattern that is characteristic of vertebrate CGRs and the genome-specific features of CGR patterns. Three years after the original proposition, Goldman (Goldman, 1993) interpreted that the frequency of dots in the CGR quadrants was nothing more than the oligonucleotide frequencies. CGR research received a setback when he asserted that simple Markov Chain models based solely on di-nucleotide and tri-nucleotide frequencies can completely account for the complex patterns exhibited in CGRs of DNA sequences. He concluded that the CGR
Introduction to Chaos Game Representation

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 oligonucleotide frequencies. Jeffrey (1990) had earlier plotted CGR of
Human Beta Globin Region on Human Chromosome 11 and the most noticeable
feature in it was the repeated (self-similar) pattern of sparse 'double scoop' shaped
regions, the largest of which is at the top of the G quadrant. Goldman pointed out that
the double scoop is nothing more than the relative rarity of CG dinucleotides. He
claimed that a four state discrete time Markov Model could easily simulate the
“double scoop” pattern and other features obtained in the CGR. According to this
conclusion, CGR should be relegated to the status of a pictorial representation of
nucleotide, dinucleotide and trinucleotide frequencies. These sobering conclusions
had the effect that CGRs have subsequently been much less studied from this
perspective.

The use of CGR for the study of the entropy of genomic sequences was noted
by Roman-Ronald et al (1994) and Oliver et al (1993). Oliver et al divided the
square into $4^\text{th}$ smaller squares as in the case of an FCGR and counted the point
density in each square. A histogram of the densities was prepared after determining
appropriate intervals. Shannon’s formula was applied to the probability distribution
histogram, thus obtaining an entropic estimate of the DNA sequence. The entropic
profile of the sequence was drawn by considering entropies at various resolution
levels. Oliver et al. showed that the entropic profiles clearly discriminate between
random and natural DNA sequences. The paper also illustrates that the entropic
profile show a different degree of variability within the genome and between
Introduction to Chaos Game Representation

genomes. The paper observes that vertebrate nuclear genomes show more variable entropic profiles than bacterial and mitochondrial ones.

The original proposition of CGR was meant for genomic sequences only. In later works it was more generalized and was shown to represent other biological sequences such as proteins (Basu et al., 1997; Pleißner et al., 1997) and also sequences of arbitrary finite number of symbols (Tino, 1999). Basu et al. used concatenated amino acid sequences of proteins belonging to a particular family. A new method of CGR was used with a 12 sided polygon in place of the CGR square. Each vertex of the polygon represented a group of amino acid residues leading to conservative substitutions. The CGR was partitioned into grids, and an estimation of the percentages of points plotted in the different segments allowed quantification of the nonrandomness of the CGR patterns generated. The CGRs of different protein families exhibited distinct visually identifiable patterns.

Deshavanne et al (1999) showed that subsequences of a genome exhibit the main characteristics of the whole genome, attesting to the validity of a genomic signature concept. The short oligonucleotide composition of a particular genome is more or less same throughout the entire genome. This property of the nucleotide sequence of an organism is known as the genome signature of that particular organism. His experiments showed that variation between CGR images along a genome was smaller than variation among genomes. He claimed that these facts strongly support the concept of genomic signature and qualify the CGR as a powerful tool to unveil it.

The measure generated on the attractor of the CGR (which is an Iterated Function System) provides more information on the sequence (Guiterrez et al., 1998;
Hao, 2000). Guitierrez et al. (Guitierrez et al., 2001) mapped a DNA symbolic sequence onto a singular measure on the attractor of a particular Iterated Function System model. A multifractal analysis of this measure is performed and singularities were interpreted in terms of mutual information and statistical dependency among subsequence symbols.

It was Almeida et al. (Almeida et al., 2001) who demonstrated that CGR may be upgraded from a mere representation technique to a sequence modeling tool. He showed that the distribution of points in the CGR has two properties: it is unique, and the source sequence can be recovered from the coordinates such that distance between positions measures similarity between the corresponding sequences. The frequency of various oligonucleotide combinations, the ‘genomic signature’, can be determined by dividing the CGR space with a grid of appropriate size and counting occurrence in each quadrant. In order to obtain the frequency matrix of oligonucleotide length \( n \), a \( 2^n \times 2^n \) grid must be used. Almeida et al showed that Markov chain models are in fact particular cases of CGRs contrary to the claim by Goldman (Goldman, 1993). The frequency matrices extracted from CGR is called a Frequency Chaos Game Representation (FCGR) and can now be reordered in the more useful Markov Chain model (MCM) format (Goldman, 1993; Almagor, 1983; Avery, 1987). Almeida showed that the conversion from FCGR to MCM is straight forward only if the number of quadrants \( k \) satisfy the condition, \( k = 2^{2n} \), where \( n \geq 1 \) is an integer. i.e. The FCGR represents an MCM only when, \( k = 2^{2n} \) is satisfied. In other words, they showed that the distribution of points in CGR is a generalization of Markov chain probability tables that accommodates non-integer orders. Unlike MCM, FCGR is not constrained to represent sequences with an integer number of bases. This fundamental
characteristic of CGR is illustrated by Almeida et al. for *E.coli* thrA where the frequency of oligonucleotides with a fractionary length has been computed by dividing the CGR plane with a $10 \times 10$ grid ($k = 100$ violates condition in the above equation). Almeida et al. also suggested a global distance measure to measure the dissimilarity between the sequences. The measure was based on a weighted Pearson correlation coefficient $r_w$ between the FCGRs. Let the two sets of FCGR quadrants be $x$ and $y$ with $x_i$ and $y_i$ representing the frequency in the $i^{th}$ quadrant. The weighted Pearson correlation coefficient is calculated as follows:

$$nw = \sum_{i=1}^{N} x_i y_i$$

$$xw = \frac{\sum_{i=1}^{N} x_i^2 y_i}{nw}$$

$$yw = \frac{\sum_{i=1}^{N} y_i^2 x_i}{nw}$$

$$sx = \frac{\sum_{i=1}^{N} (x_i - \bar{x}_w)^2 x_i y_i}{nw}$$

$$sy = \frac{\sum_{i=1}^{N} (y_i - \bar{y}_w)^2 x_i y_i}{nw}$$

$$rw_{x,y} = \frac{\sum_{i=1}^{N} \frac{x_i - \bar{x}_w y_i - \bar{y}_w}{\sqrt{sx} \sqrt{sy}} x_i y_i}{nw}$$

The advantage of using weighted correlation coefficient is that, the importance of each quadrant is made proportional to its magnitude. Hence a quadrant with a significantly high occurrence of a particular oligonucleotide is given more importance while determining similarity. The distance between the sequences is defined to be $d = 1 - r_w$ and this value ranges between 0 and 2. Note that the distance 0 corresponds to perfect correlation between the sequences, i.e. the sequences are similar. Almeida
et al. also recognized the property of CGR in finding out local similarity. The paper notes that two sequences with the same last nucleotide cannot be further than 0.5 distance apart. Also, two sequences with the same last two nucleotides cannot be further than 0.25 distance apart. The presence of similar nucleotides upstream will further shorten this distance. Note that each similar pair of nucleotides halves the distance between the sequences prior to it. This method of finding regions of local similarity was not further explored by the scientific community. The work of Almeida et al. thus positioned CGR as a powerful sequence modelling tool that has the advantages of computational efficiency and scale independence.

Anh et al. (2002) considered the problem of matching a DNA fragment to an organism using its entire genome. The authors used Recurrent Iterative Function System (RIFS) another iterative function system which has resemblance to CGR. Their hypothesis was that the multifractal characteristic of the probability measure of a complete genome, as captured by the RIFS, is preserved in its reasonably long fragments. The RIFS of the fragments of various lengths were compared with that of the original sequences using Euclidean distance as a distance measure. The hypothesis is supported by results obtained on five randomly selected genomes.

Wang et al. (Wang et al, 2005) made a detailed and comprehensive study on various genomic signatures. The papers first concern was to prove that while nucleotide, di-nucleotide and tri-nucleotide frequencies are able to influence the patterns in CGRs these frequencies cannot solely determine the patterns in CGRs. Their work generated a new sequence which simulated the dinucleotide frequency of another sequence. The CGR of the original sequence and the simulated sequence were seen not to be same. The same procedure was repeated for trinucleotide
frequencies and then also the CGRs of the original sequence and the simulated sequence did not match. These were counter examples to the result claimed by Goldman (1993). It was shown that the CGR of a sequence was not solely dependent on oligonucleotide frequencies. They showed that frequencies of oligonucleotides of all lengths are needed to determine the CGR absolutely. The second part of this paper by Wang et al. concerns various genomic signatures. In parallel to CGR research, Karlin and Burge proposed the concept of genomic signature (Karlin and Burge, 1995) which says 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 those of sequences from other organisms. In addition, closely related organisms generally have more similar DRAPs than distantly related organisms. Wang et al. (2005) demonstrated that DRAP is one particular genomic signature contained within a broader spectrum of signatures. He claimed that CGR, which provides a unique visualization of patterns in sequence organization, is another alternative genomic signature within this spectrum. In his opinion, DRAP can be considered as a second order FCGR, where the relative frequency of nucleotides is plotted instead of the usual frequency. Note that relative frequency is defined as the original frequency divided by the product of frequencies of the component monomers. Expanding this, he generalized DRAP by defining trinucleotide relative FCGR. The trinucleotide relative frequency is defined as trinucleotide frequency divided by the product of frequencies of the component monomers. Based on these the paper proposes that various kinds of genomic signatures exist, and they can be considered as members of a spectrum of genomic signatures. The paper notices that, before computing the FCGR the sequence has to be concatenated with the reverse complement strand to nullify strand bias. Another thing is that, different organisms
will be having sequences of varying length and hence FCGRs have to be standardized by nullifying the effect of sequence length, in order to effectively compare between two of them. Thirdly, the paper also proposes some distance measures between genomic signatures of two DNA sequences. Two geometric distances which he proposes are the usual Euclidean distance and Hamming distance between two standardized FCGRs. Another geometric distance the paper proposes is the Image distance, an innovation in this paper, which is computed using two concepts: neighbourhood of an integer and density in that neighbourhood. Yet another distance he mentions is a statistical one called the Pearson distance based on weighted correlation coefficient introduced by Almeida et al. (2001), which we have already mentioned. He further evaluated the phylogenetic tree produced by these various distances by comparing it with the phylogenetic tree obtained using CLUSTALW.

Dufraigne et al. (2005) used the property of genome signature to detect horizontal transfer of genes between various organisms. Since DNA transfers originate from species with a signature different from those of the recipient species, the analysis of local variations of signature along recipient genome may allow for detecting exogenous DNA. First the entire genome is scanned with a sliding window while calculating the corresponding local signature. Then, the signature of each window is evaluated by measuring its deviation from the signature of the whole genome. If the signature of a window is markedly different from that of the whole genome similar signature is searched for in a database of genomic signatures to find the putative origin of that particular fragment. Deschavanne et al. analyzed a total of 22 prokaryote genomes in this way. It has been observed that atypical regions make up ~ 6% of each genome on the average. Most of the claimed Horizontal Transfers as
Introduction to Chaos Game Representation

well as new ones were detected using this method. The origin of putative DNA transfers is looked for among ~12000 species. Donor species are proposed and sometimes strongly suggested, considering similarity of signatures.

Cenac et al (2006) considered a possible representation of a DNA sequence in a quaternary tree, based on CGR, in which one can visualize repetitions of subwords. A CGR-tree was created, which turns a sequence of letters into a Digital Search Tree (DST), obtained from the suffixes of the reversed sequence.

Fertil et al (2005) created a workspace, named GENSTYLE, for nucleotide sequence analysis based on CGR. In addition to visualization of genomic signature, the toolbox provides for comparing different signatures for the purpose of building phylogenetic tree. The origin of short DNA fragments can be searched for using this tool. The homogeneity of the signature along an entire genome could be studied which can lead to detecting Horizontal Transfers as mentioned by Dufraigne et al (2005). The software further provides for measuring similarity and differences among sequences using statistical methods such as Principal Component Analysis.

1.6 Central idea of the thesis

This thesis is an attempt to explore and enhance the potential of Chaos Game Representation as a tool for Genome sequence analysis and comparison. We demonstrate for the first time the potential of CGR for making alignment-based comparisons of whole genome sequences. A fast algorithm for identifying all local alignments between two long DNA sequences using the sequence information contained in CGR points is developed and demonstrated. Another focus of the thesis is the use of CGR as a tool to explore the concept of genomic signature and use it for
deducing phylogenetic relationships. A number of studies have demonstrated that genome signature is a phylogenetic signal which means that genome signatures of evolutionarily related organisms tend to resemble each other. In this thesis, using the different oligonucleotide frequency profiles obtained by FCGR as different representations of the genome signature, we classify different groups of organisms based on similarity of the genome signature. We find that different representations of the genome signature lead to different resolutions of the levels of classification. We apply the tool to investigate the bacterial origin of the eukaryotic organelles - mitochondria and chloroplast - by comparing the genome signatures of the organelles with those of bacteria. This leads us to formulating an alternate hypothesis for the origin of mitochondria.

This work adds to the repertoire of sequence analysis applications of Chaos Game and positions CGR as a powerful tool for genome sequence analysis and comparison.

1.7 Organization of the rest of the thesis

The potential of CGR in making alignment-based comparisons of whole genome sequences is explored in the next chapter. In this chapter local alignments between two long DNA sequences are identified using the sequence information contained in CGR points. An algorithm is developed so as to compute the length of aligned sequence from the distance between corresponding CGR points of the pair of sequences. The algorithm is made faster by reducing the complexity from $O(n \times m)$ to $O(n)$. This is done by anchoring the alignment using the FCGR matrix.

The third chapter describes the investigations of the phylogenetic signal contained in genome signatures using FCGR. The chapter begins with a description
of the traditional methods of molecular phylogeny and outlining their limitations. Phylogenetic relationships based on similarity of genome signature are determined for different groups of organisms and different representations of the genome signature.

The fourth chapter deals with the application of FCGR to investigate a specific problem namely, the evolutionary origin of the eukaryotic organelles, mitochondria and chloroplasts. The genome signature tree shows a major discrepancy from the established hypothesis that the bacterial ancestor of mitochondria is a member of the group alpha proteobacteria. We find that the genome signatures of mitochondria are closer to cyanobacteria than to most alpha proteobacteria. The unique capability of cyanobacteria to perform both oxygenic photosynthesis and aerobic respiration prompts a more parsimonious hypothesis that a single endosymbiotic uptake of a cyanobacterium could have led to the birth of both the organelles. Other arguments such as timing of evolutionary and geological events, selectional advantages conferred by combined photosynthesis and aerobic respiration and structural and functional similarity of cyanobacterial membranes to both the organellar membranes are brought together so as to demonstrate the plausibility of this alternate hypothesis. This chapter underlines the necessity to take a re-look at established phylogenetic relationships based solely on amino acid sequence similarities.

Summary and future directions are given towards the end of the thesis.

1.8 References


   GENSTYLE: exploration and analysis of DNA sequences with genomic signature,
   *Nucl Acids Res*, 33:W512-W515


   2163—2170

   genomic signature, *Trends Genet*. 11: 283--290

   160(4): 457–470

   1709–2713