Biological sequence analysis by vector-valued functions: revisiting alignment-free methodologies for DNA and protein classification

Susana Vinga

Instituto de Engenharia de Sistemas e Computadores: Investigação e Desenvolvimento (INESC-ID)
R. Alves Redol 9, 1000-029 Lisboa, Portugal
Tel. (+351) 213 100 300 Fax: (+351) 213 145 843
E-mail: svinga@algos.inesc-id.pt

Departamento de Bioestatística e Informática, Faculdade de Ciências Médicas – Universidade Nova de Lisboa (FCM/UNL)
Campo dos Mártires da Pátria 130, 1169-056 Lisboa, Portugal
Tel. (+351) 218 803 052 Fax: (+351) 218 851 920

Abstract

Biological sequence analysis is at the core of bioinformatics, bringing together several fields, from computer science to probability and statistics. Its purpose is to computationally process and decode the information stored in biological macromolecules involved in all cell mechanisms of living organisms – such as DNA, RNA and proteins – and provide prediction tools to reveal their structure, function and complex relationship networks.

Within this context several methods have arisen that analyze sequences based on alignment algorithms, ubiquitously used in most bioinformatics applications. Alternatively, although less explored in the literature, the use of vector maps for the analysis of biological sequences, both DNA and proteins, represents a very elegant proposal to extract information from those types of sequences using an alignment-free approach.

This work presents an overview of alignment-free methods used for sequence analysis and comparison and the new trends of these techniques, applied to DNA and proteins. The recent endeavors found in the literature along with new proposals and widening of applications fully justifies a revisit to these methodologies, partially reviewed before (Vinga and Almeida, 2003).

Keywords bioinformatics; sequence analysis; vector maps; DNA, proteins.
Running heads revisiting alignment-free sequence analysis

1 Introduction

Bioinformatics is an emerging and relatively new scientific multidisciplinary field that has developed very fast in the past years due to the high increase of biological data generation, particularly of genetic datasets.

Biological sequence analysis (Durbin *et al.*, 1998) is at the core of bioinformatics, being its oldest sub-discipline. Most molecular processes involve macromolecules, e.g. DNA and proteins, which can be represented as sequences. Although some of the initial paradigms are changing and new integrative techniques are being developed, it is accepted to be true that sequence determines structure that in turn determines molecular function and the overall biological role of the cell’s molecules. The recent genome sequencing projects have created an enormous quantity of data which gave rise to an urgent need of new techniques and algorithms for analyzing the massive amounts of information produced. Even with the recent explorations of higher-level integrative data, such as microarrays and genetic regulatory networks, sequence analysis and comparison is still a crucial area of research since almost all tasks depend on algorithms that process and investigate strings, from searching for similar sequences in databases to classification problems. Interestingly, this discipline is also posing new problems and challenges to statisticians and computer scientists, with the development of new algorithms and conjectures that are directly inspired by open problems in biology.

Bioinformatics and biological sequence analysis have undergone a considerable development in the past years and are expected to have an extremely high impact in the future, namely in the medical sciences and in molecular biology, continuing to cross several other disciplines and expanding on other methodologies and fields (Kanehisa and Bork, 2003).

Alignment-based methodologies are probably the most widely used tools in sequence analysis problems. They consist on arranging two sequences one on top of the other as to highlight their common symbols and sub-strings. The fundamental idea inherent to this procedure is that sequences that share the same substrings might have the same function or be related by homology. Several algorithms solved the problem of optimally aligning two sequences, either globally, such as Needleman-Wunch (Needleman and Wunsch, 1970), or locally, such as Smith-Waterman (Smith and Waterman, 1981), as to obtain a maximum score under specific scoring schemes. Other programs have been developed so far and are now ubiquitously used. These include BLAST (Altschul *et al.*, 1990; Altschul *et al.*, 1997) and FASTA (Pearson and Lipman, 1988; Pearson, 1990), for calculating dissimilarities between sequences and searching databases, and CLUSTAL (Thompson *et al.*, 1994), for multiple alignments, allowing the comparison of several sequences simultaneously.

All these methods are computational heavy but enable the researcher to compare and to search for specific traits and substrings (Gusfield, 1997). One of the major goals of alignment is to identify shared similarities that might indicate a past common ancestry or homology relationships. In fact, by alignment, one can collect evidence of events that putatively took place during molecular evolution. Furthermore the efficient implementation of the dynamic programming associated with these algorithms has created an “alignment” paradigm, overthrowing in part other possibilities for comparing and analyzing sequences.

With the explosive growth of data from new high-throughput techniques, the alignment paradigm should be reviewed. Not only it is computationally expensive but also presents several intrinsic problems. For example, it is very difficult to align complete genomes and it is a method that deals inadequately with reshuffling, to name some of the disadvantages. Other drawback is the input parameters and scoring schemes selection, which are somewhat arbitrary and not consensual between applications.

On the other hand several alignment-free techniques have emerged, based on vector representation of the sequences, providing valuable tools for their analysis and comparison. In this context, vector maps are vector-valued functions which transform sequences onto *n*-dimensional vectors, i.e., $\mathbb{R}^n$ space. The analysis is subsequently taken on this image space, benefiting of all the results and methods at hand for these kind of sets, given by linear algebra and related fields.

It is noteworthy the improvement and growth of this methodologies in the past two decades and it is apparent a burst of alignment-free techniques in more recent years, outdating a three-year-old review by the author (Vinga and Almeida, 2003). In fact, several new methods have appeared since then, making this revisit timely. The previously referred review can be considered as a complement to this present chapter.

The pioneer works on alignment-free techniques for sequence comparison date back to the 60's, when the first approaches to classify and analyze DNA sequences considered the base composition alone. The first attempts to classify proteins also started by analyzing their aminoacid composition. A lot has evolved from those first naïve approaches: several studies have consolidated alignment-free techniques for sequence analysis, departing from a mononucleotide composition approach to higher-order relations. This step unquestionably enriched all the comparisons since more information was extracted. Another significant progress was the introduction of normalization and/or transformation of the composition vectors in order to correct for biases, reduce the background frequencies, and implement more sophisticated models. Meanwhile, machine learning and bioinformatics had high impact on the type of analysis at hand of the research, greatly widening the scope of the available tools for sequence analysis.

It is worth mentioning that neither methods nor models are ultimate and, from a scientific point of view, the perfect procedure would merge and integrate several techniques. Hence, this chapter does not intent to overestimate alignment-free algorithm results over alignment-based methodologies, but highlight some of its advantages and possible applications. In fact, very recently, these two paradigms were optimally integrated in a step-wise algorithm: alignment-free, fast and efficient algorithms to pre-filter relevant relationships, followed by more time consuming techniques, but applied to a smaller subset. Presently, at least one paper that combines both approaches for multiple alignments was published (Edgar, 2004) which forestalls a burst on new hybrid techniques in the future.

In this review the main focus will be the formal and mathematical representation of sequences and the corresponding description of the analysis and comparison methodologies. For more information about the problems molecular and cell biology poses and additional background, other resources are recommended, e.g. (Lodish *et al.*, 2004).

A warning note to the reader: the borders of alignment-free and vector maps techniques are not unambiguously defined – published articles associate and integrate distinct methods – which makes the organization of this chapter somehow arbitrary. Inevitably some reviewed papers will cross those classification borders, and the reader is invited to try and understand exactly were to fit a specific work.

Another important aspect is related with nomenclature: each paper reviewed adopted a specific and sometimes contrasting terminology. In order to maintain the coherence of this chapter, we decided to alter and make uniform whenever possible all the definition and representations, although changing the original proposals.

Although non complete, given the high number of published papers so far, reaching so many applications, this work is intended to show new trends in this exciting field. If this mission is accomplished, the author might be forgiven for the inevitable omissions a review work always entails.

2 Background

This section introduces some background concepts and definitions related with the main functions available to transform sequences onto vectors. Once defined the sequence representation and its properties, the next issue is to find good measures of dissimilarity in the image space and subsequently analyze and classify the vectors thus obtained (Figure 1). Both problems, to some extend segregated, had had innumerous solutions in the past, for which the fields of linear algebra, information theory, complexity measures and stochastic processes have provided useful background and methodologies.

The following sections intention is to present an overview of the several types of representation and the possible ways we can deal with the resulting information, i.e. how can we compare and analyze the *n*-dimension vectors *F*(*X*) obtained from sequences *X*. Some important notions will be recalled and the nomenclature used throughout the remaining sections will also be defined.

2.1 Words in sequences

A sequence *X* can be represented as a succession of *N* symbols from a given alphabet $\mathcal{A}$, of length *r*, i.e., $X \equiv s_1 s_2 \cdots s_i \cdots s_N$, $i = 1, \ldots, N$ and $s_i \in \mathcal{A}$. For DNA the alphabet $\mathcal{A}$ is composed by the nucleotide symbols representing the 4 bases $\mathcal{A} = \{A, C, G, T\}$ and for proteins each symbol of this alphabet represents one of the aminoacids. For natural language texts, such as English or Portuguese, $\mathcal{A}$ is simply the set of all possible characters in each idiom.

A segment of *L* symbols, with $L \leq N$, is designated an *L-tuple* (in some references is also defined as *L-word*, *L-plet*, or *L-mer*). The set $\mathcal{W}_L$ consists of all possible *L-tuples* that can be obtained from the alphabet $\mathcal{A}$ (with *r* symbols), and has *K* elements (Eq.1).

$$\begin{aligned} &\mathcal{W}_L = \{w_{L,1}, w_{L,2}, \cdots, w_{L,K}\} \\ &K = r^L \end{aligned} \quad \textbf{(1)}$$

The identification of *L-tuples* in the sequence *X* can then be object of counting occurrences with overlapping (Eq.2). Computationally, the counting is usually performed by taking an *L*-wide sliding window that is run through the sequence, from position *1* to *n*–*L*+1, the maximum total number of counts one can obtain.

$$c_L^X = \left(c_{L,1}^X, \cdots, c_{L,K}^X\right) \quad \textbf{(2)}$$

Similarly, the word frequencies f_L^X estimate the probabilities $p_L^X = \left(p_{L,1}^X, p_{L,2}^X, \cdots, p_{L,K}^X\right)$ of finding specific words $w_{L,i}$, $i = 1, \ldots. K$. The vector of frequencies f_L^X is the relative abundance of each word (Eq.3).

$$f_L^X = \frac{c_L^X}{\sum_{j=1}^K c_{L,j}^X} \Leftrightarrow f_{L,i}^X = \frac{c_{L,i}^X}{n - L + 1} \quad \textbf{(3)}$$

Sometimes, for convenience, the frequency vector *f* is indexed by the *L*-tuple it represents instead of its relative order *i*, i.e., $f_{w_L} \equiv f_{s_{i_1} \cdots s_{i_L}}$ represents the frequency of the *L*-tuple $w_L \equiv s_{i_1} \cdots s_{i_L}$. The difference between the two nomenclatures should be evident from the context where they appear.

For example, for DNA sequences, $\mathcal{A}=\{A, T, C, G\}$, $r=4$, a three letter word, $L=3$, could be $w_3=GTC$. For the sequence $X=GTGTGA$, where $n=6$, the vector p_3^X is estimated by the relative frequencies of all trinucleotides. The frequencies, determined by sliding a 3 letter window $n–L+1=4$ times, would be:

$$W_L = \{GTG, TGT, TGA, AAA, AAC, \ldots\}$$

$$c_3^X = (2,1,1,0,0,\ldots)$$

$$f_3^X = (0.5, 0.25, 0.25, 0, 0, \ldots)$$

The vectors c_3^X and f_3^X have length $K=4^3=64$, the zero coordinates corresponding to missing words in *X*, in this case absent trinucleotides. Under the alternative representation one might write $f_{GTG}^X = 0.5$.

2.2 Vector valued functions

A vector map or mapping is a vector-valued function, i.e., a function that assumes values on the space $\mathbb{R}^n$. Given a sequence $X \equiv s_1 s_2 \cdots s_N$, $i = 1, \ldots, N$ from an alphabet $\mathcal{A}$, it is possible to define a vector map $F : \mathbf{S} \rightarrow \mathbb{R}^n$, which transforms *X* (taken from the space of all possible sequences **S**) onto an *n*-dimensional vector $x = (x_1, x_2, \ldots, x_n)$:

$$F(X) \in \mathbb{R}^n \tag{4}$$

The following Figure 1 depicts the transformation steps, from sequences to *n*-dimensional spaces and the subsequently pos-processing methods reviewed in this chapter.

Fig.1

Figure 1: Vector maps of biological sequences. This figure represents the transformation of a biological sequence *X* onto an *n*-dimensional vector. This might be represented as $F : \mathbf{S} \rightarrow \mathbb{R}^n$. This chapter will review several methodologies that study and analyze sequences in this image space.

It is possible to individualize several mappings from sequences to vectors. The following sections will describe the maps based on *L*-tuple composition and those based on iterated function systems, namely Chaos Game Representation (CGR). Other representations based on Markov chain models, time series, information theory and complexity will also be indicated.

2.2.1 *L*-tuple composition

One of the simplest tasks, initiated in the beginning of bioinformatics endeavors, was to assess the composition of the sequences in terms of their nucleotide or aminoacid content, for DNA and proteins respectively, which corresponds to extract their 1-tuple frequencies. This evaluation was subsequently refined, by considering high order tuples.

Some authors consider these approaches as linguistic methods, because of the obvious analogies with the study of natural languages texts in which symbols and words are defined and compared.

As mentioned above, this type of mapping is simply given by the frequencies of each *L*-tuple, over an alphabet with *r* symbols:

$$\begin{array}{l} F : \mathbf{S} \to \mathbb{R}^n \\ F\left(X\right) = f_L^S = \left(f_{L,1}, \ldots, f_{L,n}\right) \end{array}, \text{with } n = r^L = K \qquad \textbf{(5)}$$

The vectors obtained can subsequently be object of normalization and filtering procedures, in order to extract relevant information and correct for several biases.

2.2.2 Chaos Game Representation (CGR) and iterated function systems

Chaos game representation (CGR) was first presented in 1990 (Jeffrey, 1990) as a method for mapping DNA sequences onto vector spaces. It is derived from iterated function systems and is related with fractal geometry and chaos theory (Edgar, 1990). There are several applications of this method in bioinformatics, such as the investigation of patterns in DNA, the extraction of Markov models transition tables (Almeida *et al.*, 2001) and the calculation of entropies (Vinga and Almeida, 2004; Vinga, 2005). The algorithm itself is closely linked to binary representations of sequences, conceiving a useful property shown below. The CGR generalization for higher-order alphabets, named Universal sequence maps (USM), was presented in another paper (Almeida and Vinga, 2002).

The CGR iterative algorithm maps each symbol onto a 2-dimensional vector, following an iterative procedure where each point is calculated using the previous one. The CGR mapping $x_i \in \mathbb{R}^2$ of an *N*-length DNA sequence $X \equiv s_1 s_2 \cdots s_N$, $i = 1, \ldots, N$, $s_i \in \mathcal{A} = \left\{A, C, G, T\right\}$ is given by the following equation 6:

$$\begin{cases} x_0 \sim Unif\left(0,1\right)^2 \\ x_i = x_{i-1} + \dfrac{1}{2}\left(y_i - x_{i-1}\right), i = 1, \ldots, N \end{cases} \quad where\ y_i = \begin{cases} \left(0,0\right) & if \quad s_i = 'A' \\ \left(0,1\right) & if \quad s_i = 'C' \\ \left(1,0\right) & if \quad s_i = 'G' \\ \left(1,1\right) & if \quad s_i = 'T' \end{cases} \qquad \textbf{(6)}$$

In the original proposal of CGR (Jeffrey, 1990), the starting point was taken as (0.5,0.5), although none of the properties are altered by choosing a random initial position x_0. The algorithm assigns each of the DNA bases to a vertex in the square $[0,1]^2$ and, in each step, goes half the distance towards the corner representing the following symbol in the original sequence. The plot of all the points x_i originates a fractal-like pattern, as exemplified in Fig. 2a.

One of the most interesting properties of CGR is its ability to group points that represent the same suffix in the original sequence. In fact, the same suffix of any length will be mapped in the same sub-square, even if the original symbols are far apart in the sequence, as exemplified in Fig.2b.

Fig.2a)	Fig.2b)

Figure 2: CGR example and suffix property. On the left an example of CGR algorithm is shown, with the image obtained for the human beta globin region on chromosome 11 (HUMHBB), constituted by 73308 bases. In the right panel the suffix property is exemplified: sequences sharing the same suffix are on the sub-square labeled with that substring.

CGR can also be interpreted as a binary representation of the sequence. For example, from the coordinates of one symbol in base 2 it is possible to extract all the previous symbols in the sequence that originated that mapping. As an example, a point in base 10 with coordinates x_i = $(0.65625, 0.21875)_{10}$ = $(0.10101, 0.00111)_2$ corresponds to the succession of symbols (1,0)→(0,0)→(1,1)→(0,1)→(1,1), i.e. the suffix GATC. (For more properties of CGR/USM see (Vinga, 2005).)

In this case we have a collection F_i of vector maps, one per symbol in a given context $i = 1,\ldots,N$, which can be rationalized as:

$$\begin{array}{l} F_i : S \to \mathbb{R}^n \\ F_i\left(S\right) = \left(x_{i1}, x_{i2}\right) \end{array}, \text{ with } n = 2 \qquad \textbf{(7)}$$

The literature in CGR has grown significantly in the past decade, as described on the following sections, which anticipates its importance in the future as a general model for sequence analysis.

2.2.3 Markov chain models

Markov chains are probabilistic models that have extensively been applied in biological sequence analysis, constituting a rich and fruitful approach to study DNA and proteins (Durbin *et al.*, 1998; Robin *et al.*, 2005).

Markov chains are special cases of stochastic processes (Kulkarni, 1995), where the probability of a sequence $X \equiv s_1 s_2 \cdots s_i \cdots s_N$ is defined conditionally to the last *L* symbols. This constitutes a "memoriless" property: given the present (the last symbols or suffix in the sequence), the probability of the following symbol is independent form the past. Formally, the probability of observing a given sequence $s_1 s_2 \cdots s_i \cdots s_N$ with $s_i \in \mathcal{A}$ can be expressed with conditional probabilities as:

$$\begin{aligned} & p\left(X_N = s_N, \ldots, X_i = s_i, \ldots, X_2 = s_2, X_1 = s_1\right) = \\ & = p\left(s_1 s_2 \cdots s_i \cdots s_N\right) = p\left(s_N \left| s_1 s_2 \cdots s_{N-1}\right.\right) \cdot p\left(s_{N-1} \left| s_1 s_2 \cdots s_{N-2}\right.\right) \cdot \ldots \cdot p\left(s_2 \left| s_1\right.\right) \cdot p\left(s_1\right) \end{aligned}$$

The Markov property allows the simplification of the last expression. For an *L*-order Markov chain, the current *L*-tuple is sufficient to calculate the probability of the next symbols, i.e., the probabilities are independent from the symbols occurring before the position *i*–*L*+1:

$$p\left(s_{i+1}\left|s_1\cdots s_i\right.\right)=p\left(s_{i+1}\left|s_{i-L+1}\cdots s_i\right.\right)$$

Under a zero-order Markov chain model the probabilities are simply $p\left(s_{i+1}\left|s_1\cdots s_i\right.\right)=p\left(s_{i+1}\right)$, i.e., the probabilities of each symbol do not depend on the current state.

When studying *L*-tuple frequencies, the sequence is usually modelled as a (*L*-2)-order Markov chain, named the *maximal order model*. Under this assumption, the probability of an *L*-tuple is simplified, since $p\left(s_L\left|s_1s_2\cdots s_{L-1}\right.\right)=p\left(s_L\left|s_2\cdots s_{L-1}\right.\right)$, giving the following probability:

$$\begin{aligned}p\left(s_1s_2\cdots s_{L-1}s_L\right)&=p\left(s_L\left|s_1s_2\cdots s_{L-1}\right.\right)\cdot p\left(s_1s_2\cdots s_{L-1}\right)\\&=p\left(s_L\left|s_2\cdots s_{L-1}\right.\right)\cdot p\left(s_1s_2\cdots s_{L-1}\right)\\&=\frac{p\left(s_1s_2\cdots s_{L-1}\right)\cdot p\left(s_2\cdots s_{L-1}s_L\right)}{p\left(s_2\cdots s_{L-1}\right)}\end{aligned}\qquad\textbf{(8)}$$

This model will appear in some applications described in the following sections.

The estimation of probabilities defined above allows us to calculate the expected number of occurrences of a specific word or *L*-tuple and also its statistical significance. In fact, these models can recognize under and over-represented motifs, which might indicate that the regions they represent are important from a biological point of view. There is already a large study of words in sequences and their statistical properties, which are out of the scope of this chapter. For a comprehensive introductions see (Reinert *et al.*, 2000; Robin *et al.*, 2005) and references therein.

Markov chain models are closely connected with vector maps, since the estimation of the transition probabilities typically involves the calculation of the *L*-tuple frequencies. In fact, as mentioned above, it is usually taken $\hat{p}^X\left(s_1\cdots s_L\right)=f^X_{s_1\cdots s_L}$. Since CGR maps generalize *L*-tuple frequencies, it is straightforward to extend Markov models to these iterative function systems. For example, the maximum likelihood estimation of the transition probability of having an "*A*" given that the current symbol is a "*G*", *p*(*A*|*G*), i.e., a first order Markov chain, involves the calculation of 2-tuples counts or frequencies $\hat{p}\left(A\,|\,G\right)=\frac{c_{GA}}{c_{G+}}=\frac{f_{GA}}{f_{GA}+f_{GC}+f_{GT}+f_{GG}}$. This estimation can be calculated directly from the CGR, due to the suffix property exemplified in Fig. 2b. The 2-tuples involved in the last expression are in specific regions of the map. It is then natural that CGR is shown to be generalization of any-order Markov models (Almeida *et al.*, 2001). In a recent exploit, several methods were tested to build finite memory predictive models (Tino and Dorffner, 2001) and those based on CGR were among the best, by partitioning the CGR map in arbitrary regions. In fact, high accuracy is attained if the geometry of the prediction is allowed to vary, without the constrain of choosing specific squares of size 2^{-L}.

2.2.4 Time series, complexity and entropy

Time series constitute another possible type of sequence representation. The proposal for these methodologies is central in the biophysicists' literature, allowing to study, among other questions, the long and short-range correlations present in DNA. Briefly, this type of approach considers a sequence as a time series and uses models currently available in the field of stochastic processes and dynamical systems to unravel the correlation structure of DNA. In several applications the type of algorithms is similar to other approaches, namely by considering some Markov dependencies up to a certain level. Information theory also plays a role, with entropy measures and periodicity evaluation of the series obtained.

Although slightly out of the scope of this survey, it is important to mention important references in this field. A valuable repository of bibliography related with correlations in DNA and related themes is available in the extremely well curated website http://www.nslij-genetics.org/dnacorr, with some methods reviewed in (Li, 1997) and references therein. Departing from a simple random walk model for DNA other articles created more complex interactions. In a recent report, Dehnert and colleagues have proposed a model based on a discrete autoregressive process of order *p*, DAR(p), showing that it is possible to create a dissimilarity measure from the vectors of the DAR parameters (Dehnert *et al.*, 2005). Using this technique, phylogenetic trees can be reconstructed allowing inter and intra-species discrimination.

The notion of complexity and entropy is also extensively presented elsewhere (Cover and Thomas, 1991; Li and Vitanyi, 1997) and only briefly exemplified in the reviewed applications.

It is noteworthy the relation between the methods in this category, with overlapping areas and techniques. For example, several recent studies highlight the correspondence between time series, random walks and CGR (Wu, 2004; García and José, 2005). Other work used multifractal analysis of specific number representations of sequences, also related to CGR, distinguishing between coding and non-coding regions in whole genomes (Zhou *et al.*, 2005).

2.3 Comparing vectors: metrics and dissimilarities

In the last section several methods to represent sequences as *n*-dimensional vectors were described, through the definition of specific vector-valued functions. This section overviews the procedures used to compare the vectors obtained by assessing their dissimilarity or distance, thus estimating the similarity of the original sequences they represent. This is accomplished by defining appropriate metrics and dissimilarities in the image space $\mathbb{R}^n$, a problem that can be segregated from the sequence representation itself.

Formally, a metric space (**S**, *d*) is a set **S** together with a non-negative function $d:\mathbf{S}\times\mathbf{S}\rightarrow\mathbb{R}_0^+$ satisfying, for all $X,Y,Z\in\mathbf{S}$, the properties of positivity, symmetry and triangle inequality:

$$\begin{aligned}&d\left(X,Y\right)=0\Leftrightarrow X=Y\\&d\left(X,Y\right)=d\left(Y,X\right)\\&d\left(X,Y\right)\leq d\left(X,Z\right)+d\left(Z,Y\right)\end{aligned}\qquad\textbf{(9)}$$

Such a function *d* is called a *metric* and measures the distance between pairs of points *X* and *Y* in **S**. In the present context, the set **S** represents the collection of all possible sequences and *d*(*X*,*Y*) is therefore a measure of their distance.

There are also functions *d*(*X*,*Y*) that do not obey to all three properties described in Eq.9 (for example, in some of them the triangle inequality is not verified) but have nonetheless high importance in several applications. The function *d*(*X*,*Y*) in these generalizations will be referred to as *dissimilarity* between two sequences.

When analyzing the vectors obtained from the maps, this context provides tools to assess and compare the sequences. As mentioned in the introduction, this chapter follows a previous review of alignment-free methods (Vinga and Almeida, 2003), where several dissimilarly measures and metrics were described in detail. For sake of clarity, the most widespread used distances therein described are briefly recalled in Table 1; the reader is advised to search for the full references in the original review (Vinga and Almeida, 2003).

Table 1 Definition of dissimilarity measures between sequences *X* and *Y*, *d*(*X*,*Y*). See also (Vinga and Almeida, 2003) for a complete description and full references. The measures are based on the comparison of the vectors of counts, frequencies, or methods where the resolution *L* of the tuple is not required.

Dissimilarity	Equation
Euclidean	$d_L^E(X,Y)=\left(c_L^X-c_L^Y\right)^T\cdot\left(c_L^X-c_L^Y\right)=\sum_{i=1}^K\left(c_{L,i}^X-c_{L,i}^Y\right)^2$
Weighted Euclidean d2	$d^2(X,Y)=\sum_{L=l}^u\sum_{i=1}^K\rho_i\left(c_{L,i}^X-c_{L,i}^Y\right)^2$
Standard Euclidean	$d_L^{SE}(X,Y)=\left(c_L^X-c_L^Y\right)^T\cdot\left[\mathrm{diag}\left(s_{11},\cdots,s_{KK}\right)\right]^{-1}\cdot\left(c_L^X-c_L^Y\right)=\sum_{i=1}^K\frac{\left(c_{L,i}^X-c_{L,i}^Y\right)^2}{s_{ii}}$
	$d^{SE*}=\sum_{L=l}^u d_L^{SE}$
Mahalanobis	$d_L^M(X,Y)=\left(c_L^X-c_L^Y\right)^T\cdot\mathbf{S}^{-1}\cdot\left(c_L^X-c_L^Y\right)=\sum_{i=1}^K\sum_{j=1}^K\left(c_{L,i}^X-c_{L,i}^Y\right)\cdot s_{ij}^{inv}\cdot\left(c_{L,j}^X-c_{L,j}^Y\right)$
	$d^{M*}=\sum_{L=l}^u d_L^M$
Linear correlation coefficient	$d_L^{LCC}(X,Y)=\frac{K\sum_{i=1}^K f_{L,i}^X\cdot f_{L,i}^Y-\sum_{i=1}^K f_{L,i}^X\cdot\sum_{i=1}^K f_{L,i}^Y}{\left[K\sum_{i=1}^K\left(f_{L,i}^X\right)^2-\left(\sum_{i=1}^K f_{L,i}^X\right)^2\right]^{1/2}\cdot\left[K\sum_{i=1}^K\left(f_{L,i}^Y\right)^2-\left(\sum_{i=1}^K f_{L,i}^Y\right)^2\right]^{1/2}}$
Kullback-Leibler	$d_L^{KL}(X,Y)=\sum_{i=1}^K f_{L,i}^X\cdot\log_2\left(\frac{f_{L,i}^X}{f_{L,i}^Y}\right)$

Cosine	$d_L^{COS}(X,Y)=\theta_{XY}$, where $\cos(\theta_{XY})=\frac{\left(c_L^X\right)^T\cdot c_L^Y}{\left\lVert c_L^X\right\rVert\cdot\left\lVert c_L^Y\right\rVert}=\frac{\sum_{i=1}^{K}c_{L,i}^X\cdot c_{L,i}^Y}{\sqrt{\sum_{i=1}^{K}\left(c_{L,i}^X\right)^2}\cdot\sqrt{\sum_{j=1}^{K}\left(c_{L,j}^Y\right)^2}}$
Evolutionary	$d_L^{EVOL}(X,Y)=-\ln\left[\left(1+\cos\theta_{XY}\right)/2\right]$
CGR/USM	$d^{USM}(a,b)=-\log_2\left(\max_i\left\lvert a_i-b_i\right\rvert\right)$
Kolmogorov complexity	$d^{KC}(X,Y)=1-\frac{K(X)-K(X\mid Y)}{K(XY)}$

2.4 Sequence analysis and classification

One of the main objectives of mapping the sequences onto vector maps, using the methodologies described above, is to obtain a representation more amenable to extract information, either to analyze one particular sequence or to compare and classify a set of several sequences. The goal of all these methods is to infer key features of the sequences as to extract relevant biological information, using solely the vectors.

The definition of metrics and dissimilarities in this space allows precisely the subsequent classification of the sequences. In fact, several authors use those functions to create dendrograms or trees, using specific clustering algorithms. Theses techniques are nowadays so numerous that it would impractical to review them in detail. Consequently, classification, clustering analysis and grouping algorithms are not described in this chapter: for a comprehensive introduction (Everitt *et al.*, 2001) and (Gordon, 1999) are recommended. Most of the methods use as input a dissimilarity matrix with all pairwise comparisons between the sequences, which is then object of agglomeration, for which exist widely available implementation in standard statistical packages. More specific clustering methodologies applied in the phylogenetic and evolutionary framework are also reviewed elsewhere (Snel *et al.*, 2005).

The following sections contain examples of how the above mentioned methodologies can provide tools to efficiently group and analyze both DNA and proteins.

3 Alignment-free methodologies

The previous sections briefly described the formalism of vector-maps based analysis of sequences. This section presents some new applications of those methods, concentrating on the new developments of this area in recent years. In particular, a brief overview of the progresses will be carried out. All the algorithms can be classified as alignment-free methods because they are not based on alignment to compare and analyze sequences. The recent interest in this area is demonstrated by the wealth of articles and reviews published (Bolshoy, 2003; Vinga and Almeida, 2003), some of them revisited in the next sections.

3.1 Transforming L-tuple frequencies: variations of a theme

Departing from the simple frequency idea, several methods have been envisaged to deal with more complex problems and data. One of the most straightforward approaches is to pre- or pos-process the original composition vectors to extract meaningful and discriminant information. This is accomplished by specific normalization steps, filtration procedures and/or feature selection, using algebraic and statistical tools and properties.

As referred to in the previous review, Blaisdell's paper introduced the notion of sequence similarity measure without pre-alignment using Markov models and all *L*-tuple counts (Blaisdell, 1986), associated with Euclidean distances $d_L^E\left(X,Y\right)$ on this space. This metric was further extended by weighting the vectors, and named d2-distance in a subsequent paper (Torney *et al.*, 1990). Their distributional statistical study was accomplished later (Lippert *et al.*, 2002). This last step would provide the tools to estimate the statistical significance of the comparisons, deducing the asymptotical distributions of d2 for random sequences. For example, depending on the word type and the sequence length, the asymptotical distribution is shown to be Poisson, Normal or Gumbel. With this extension, the transformation by weighting of the original *L*-tuple frequency vectors introduced greatly improved the results that could be obtained.

In another pioneer work dating back to 1990 (Pietrokovski *et al.*, 1990), the raw *L*-tuple DNA frequencies are normalized by their expected values. Accordingly, an alignment-free measure of dissimilarity between two sequences based on contrast values is proposed, based on the difference between observed and expected frequencies of *L*-tuples. These expected frequencies are calculated under a maximal order Markov model, using the information of the (*L*-1) and (*L*-2)-tuple counts (Eq.8). For example, to calculate the expected frequency of a given 4-tuple $S \equiv s_1 s_2 s_3 s_4$ under the maximal Markov model of order 2, $E\left[f_S^X\right]$, the observed frequency $f_{s_1s_2s_3}^X$, $f_{s_2s_3s_4}^X$, $f_{s_2s_3}^X$ are used such that:

$$E\left[f_{s_1s_2s_3s_4}^X\right] = \frac{f_{s_1s_2s_3}^X \cdot f_{s_2s_3s_4}^X}{f_{s_2s_3}^X} \qquad \textbf{(10)}$$

The corresponding *contrast values* $q_L^X = \left(q_{L,1}^X,\ldots,q_{L,i}^X,\ldots,q_{L,K}^X\right),\; i=1,\ldots,K=4^L$, are the deviations of the expected $E\left[f_{L,i}^X\right]$ vs. the observed $f_{L,i}^X$ frequencies and define the contrast *L*-vocabulary (CV):

$$q_{L,i}^X = f_{L,i}^X - E\left[f_{L,i}^X\right] \qquad \textbf{(11)}$$

High q_L values correspond to over-represented *L*-tuples and low q_L to under-represented ones, under the specified Markov model.

When comparing two different sequences *X* and *Y*, the measure of dissimilarity is defined, for a given resolution *L*, as the correlation coefficient $d_L^{CV}\left(X,Y\right)$ between the two contrast *L*-vocabularies:

$$d_L^{CV}\left(X,Y\right) = \frac{\sum_{i=1}^{K} q_{L,i}^X q_{L,i}^Y}{\sqrt{\sum_{i=1}^{K}\left(q_{L,i}^X\right)^2 \cdot \sum_{i=1}^{K}\left(q_{L,i}^Y\right)^2}} \qquad \textbf{(12)}$$

The value $d_L^{CV}\left(X,Y\right)$ can be interpreted as the cosine of the angle between the contrast vectors *q*(*X*) and *q*(*Y*) – thus measuring a type of correlation between the vectors – where each coordinate or dimension corresponds to one oligomer. Values close to 1 indicate similar sequence attributes, values around 0 signify that no relation is present and values near -1 mean that the same words or *L*-tuples are over and under-represented in the opposite pattern in both sequences. The authors further extend this measure averaging $d_L^{CV}\left(X,Y\right)$ for several *L*-tuples (from di- to pentanucleotides), creating a combined version of the previous one, shown to be more informative:

$$d^{CV*}=\frac{1}{4}\sum_{L=2}^{5}d_L^{CV}\left(X,Y\right)$$

In this work (Pietrokovski *et al.*, 1990) the basic idea of a dissimilarity measure not requiring alignment was presented and the major advantages of this approach were described. One property there highlighted was the computational convenience (computationally inexpensive) when confronting with alignment-based methods. In addition, the comparison of short sequences of contrasting lengths is made possible with this approach (as with all angle-based metrics), since the measure implemented is unaffected by the overall counting, apart from sampling biases. At this point it should be apparent the relation between this measure and those presented in Table 1, namely $d_L^{COS}\left(X,Y\right)$ and $d_L^{LCC}\left(X,Y\right)$, applied in subsequent studies. The results confirmed the possibility of application of these methodologies for the quick screening of databases and the taxonomic classification of sequences.

In this work some of the problems of alignment-based methods were also remarked, they are best applicable when the sequences are similar, since detection of distance relatedness poses both computational and conceptual problems. The linguistic interpretation given to this approach is related with the natural languages equivalent, where each idiom can be characterized by a vocabulary (words) and we can distinguish texts (or sequences) in terms of their origin and further characterize them through their resemblance solely based on the different frequencies of the corresponding substrings.

Following this report, the same author comprehensively reviewed linguistic methods for sequence comparison, defined as those in which the basis for sequences analysis is their oligomer composition (Pietrokovski, 1994). It should be noted that there are other possible ways of calculating the deviations, namely by considering other types of Markov models, hence originating different expected values of the frequency vectors. In addition, other types of distances are also available. For example, if the mean values of q_j are subtracted the Pearson's correlation coefficient $d_L^{LCC}\left(X,Y\right)$ is obtained. In this review, several applications are presented which show the potential of the measure, ranging from identifying similar (homologous) and dissimilarity (taxonomically and functionally) sequences, locating eukaryotic promoters and identifying imported sequences in the mitochondrial yeast genome.

These notions of linguistic properties and contrast vocabularies of genomes were the rationale for studying in depth the oligonucleotide bias in *Bacillus subtilis* (Rocha *et al.*, 1998). In this work is stated that over- and under represented words or *L*-tuples in a genomic datasets, defined by their contrast values, might indicate phenomena of positive/negative selection. The authors studied 1 to 8-tuple frequency vectors, normalized by their expected values, finding biases indicating that over-represented *L*-tuples play an important role as biological signals. The analysis was further extended to other organisms and Kendall-τ association measures between the corresponding vocabularies were compared. Interestingly, the analysis of over and under-representation of short oligonucleotides in DNA sequences had been carried out before (Burge *et al.*, 1992), leading afterwards to the genomic signature concept described in the following section. The analysis of the relation between nucleotide biases in virus and their phylogeny had also been explored in a previous report (Bronson and Anderson, 1994).

All these works show that processing the frequency vectors is a good solution for analyzing data and correcting for a wide sort of biases, present in the original mapping.

Another alternative concerning the transformation would be to select relevant information from specific *L*-tuples, thus filtering superfluous information. Instead of using the complete f_L vector, a reduced version can also be employed to characterize the sequences. The idea was pursued in some recent works.

The definition of compositional spectrum of one sequence *X* relative to a set *W* (Kirzhner *et al.*, 2002) is related with the this idea of processing the *L*-tuple frequencies. The method is based on randomly selecting *n* different *L*-tuples, which constitute the set *W*, and finding all occurrences of those *L*-tuples in the sequence *X*, but allowing for errors. This means that imperfect matching is permitted in the frequency vector, considering the Hamming distance, and that only a fraction of all possible *L*-tuples is taken into account. The compositional spectrum is simply the histogram of these imperfect or fuzzy word occurrences. It is possible to define a dissimilarity measure between compositional spectra, the CS-distance, by the Spearman's rank correlation coefficient, which allows the large-scale comparison and clustering of genomic sequences (Kirzhner *et al.*, 2003; Kirzhner *et al.*, 2005). All the results show robustness and reproducibility across several simulations (given the random choice of the set *W*) and the classifications are consistent with other reports. Furthermore, these CS-distances are larger for inter-species comparisons than for intra-species, which corroborates a signature concept described in the following section.

Other possible extension that takes into account the variability and imperfect consensus of sequences was proposed by Torres and colleagues (Torres and Nieto, 2003), where a 12-dimensional space is used to represent a codon. More specifically, each of the 3 group of 4 coordinates, represents a base (*U*=(1,0,0,0), *C*=(0,1,0,0); *A*=(0,0,1,0); *G*=(0,0,0,1). If insufficient knowledge of the chemical structure is available, the binary vectors will represent a fuzzy set, with values between 0 and 1. For example *p*=(0.3 , 0.4 , 0.1 , 0.2 , 0 , 1 , 0 , 0 , 0 , 0 , 1) will represent (*C*,*A*,*U*), where the first codon position can have a probabilistic interpretation as: "*C* to the extent 0.40, *U* to the extent 0.30, *A* to the extent 0.10 and *G* to the extent 0.20". A metric on this space is further proposed, i.e., defining a distance between sequences *X* and *Y*, based in the differences between the vectors p^X and p^Y, similar to the Manhattan or city-block distance:

$$d^{codon}\left(X,Y\right)=\frac{\sum_{i=1}^{12}\left|p_i^X-p_i^Y\right|}{\sum_{i=1}^{12}\max\left\{p_i^X,p_i^Y\right\}} \qquad \textbf{(13)}$$

This is proven in fact to be a distance (Nieto *et al.*, 2003), thus defining a metric space in $\mathbb{R}^{12}$, and is explored to measure the differences between genomes *Mycobacterium tuberculosis* and *Escherichia coli*. All the codons are then mapped into a fuzzy set of frequencies of the genome.

Interestingly, this might be interpreted as Position-Specific Scoring Matrices (PSSM) for that particular region: in fact, by assuming that each entry is a matrix m_{ij} with dimension 4x3, these measure the probability of occurrence of each symbol s_i in each codon position j. This result could be further applied to compare PSSM extracted from different groups, creating a quantitative dissimilarity measure in this fuzzy set.

Another result where the vectors are pre-processed is given by the filtering of relevant substrings, ignoring the major non-characteristic data that might be considered background noise. In particular, some of the sub-strings might represent relevant features, so it would make sense to ignore part of information available (the complete frequency vector) and look for specific subsets, as mentioned above.

For example, alignment-free analysis can be achieved by extracting, from a sequence set, the shortest unique substrings (Haubold *et al.*, 2005), which are strings that occur only once and which cannot be further reduced in length without losing the property of uniqueness. For their detection, generalized suffix trees can be used, which makes for an efficient algorithm with running times that are linear in the combined lengths of the input sequences. In this work the authors analytically deduce the probability distribution of the occurrence of these strings in random models in order to measure their statistical significance and expected values in real sequences. The methodology is further applied to the *Caenorhabditis elegans*, human and mouse genomes (with approximate sizes 100Mbp, 2.8Gbp, 2.49Gbp) and the shortest unique substrings were found out to be respectively, 10, 11 and 11. This result cannot be expected since the human genome is 28 times larger that that of *C. elegans* and there is a small probability of observing these shortest unique strings. Furthermore they were shown to occur clustered close to genes, more that what would be expected just by chance. This method is proposed as a starting point for developing signature oligos and also for comparative genomics and taxonomy studies.

The notion of unique *L*-tuple was also explored in other two different contexts. In (Chen *et al.*, 2002) the idea of identifying unique segments of the human genome was efficiently applied to single nucleotide polymorphism (SNP's) mapping. Due to the high increase of databases, alignment-based algorithms are unpractical and very time-consuming to perform this task. Instead of aligning whole genomes, fixed-length unique sequence markers, or UniMarkers (UM), might be used to assign the genomic positions of SNP sites. UM's appear only once in the genome thus allowing to locate SNP's much faster that alignment-based methods. The authors find an optimum length of 15-tuple UM's to position the SNP's sequences. Interestingly, the UM's search was performed using a method that could be implemented with CGR maps, by searching unique sub-strings through bit wise operations and binary trees. This method can be easily generalized to Expressed Sequence Tags (EST) mapping. UM's generally represent a set of genome-wide, high resolution

genetic markers. The authors also highlight the potential use of this technique as a filter and anchor to improve the efficiency of alignment-based algorithms.
The use of UniMarkers was later explored in comparative genomics applications, more specifically to create synteny maps (Liao *et al.*, 2004), i.e., detect regions that are orthologous between two genomes. This fast alignment-free method was applied for mapping evolutionary conserved segments between two large genomes and further confronted with NCBI (National Center for Biotechnology Information) assignments, achieving high accuracy. Instead of performing a whole-genome alignment, with all the problems this procedure implies, the UM method overcomes some of those previous drawbacks, showing robustness to segment shuffling (without the need of having contiguity between orthologous). For example, the identification of common ancestry is performed by counting the relative numbers of their shared UM's, not by analyzing their order. Furthermore by considering large regions of the genome, the innumerous local similarities present are avoided; the computational and memory load is low, which renders feasible whole comparisons on a personal computer.
As seen, subsets of the *L*-tuple frequencies can be efficient descriptors of the original sequences, since they automatically filter noise and extract discriminant features.

Another recent article (Chen *et al.*, 2005) studied the distribution of the 6-tuple frequencies and analyzed the histograms obtained for several organisms. They obtained uni- and bimodal distributions and proposed models for the processes that might give rise to that behavior. The distribution of repeated and rare *L*-tuples was investigated in (Fofanov *et al.*, 2004), calculating the percentages of "used" oligonucleotides and providing analytical deduction of the probabilities of finding specific repetitions. Rare *L*-tuples can also be combined to distinguish individual gene sequences, and could be used as target for DNA probes (Gibbs *et al.*, 2005). All these results illustrate the vast number of applications obtained from transformed *L*-tuple vectors.

3.2 Entropy and complexity

Information theory has provided useful background and tools for sequence analysis, such as the definition of the Kullback-Leibler discrepancy between probability vectors. This definition was shown to be pertinent in several articles, highlighted in this section. The entropy concept is usually related with complexity theory and algorithms for compressing strings. The unpredictability of sequences might be measured with these methods, which have also shown to be useful for sequence comparison and classification.

In a recent paper a probabilistic measure for alignment-free comparison was proposed (Pham and Zuegg, 2004). In this work, the authors extended the Kullback-Leibler discrepancy or relative entropy between distributions (Table 1) to first order Markov-chains parameters. Considering a first-order Markov Model (MM) with parameters λ_1 and λ_2 estimated from two sequences X_1 and X_2, with lengths N_1 and N_2 respectively, we can define the dissimilarity $d^{MM}(X_1, X_2)$ from the comparison of the log-likelihoods $\mathrm{P}\left(X_i \middle| \lambda_j\right)$, $i, j = 1, 2$:

$$d^{MM}\left(X_1,X_2\right)=1-\exp\left[D_s\left(X_1,X_2\right)\right]$$
$$D_s\left(X_1,X_2\right)=\frac{D\left(X_1,X_2\right)+D\left(X_2,X_1\right)}{2} \tag{14}$$
$$D\left(X_1,X_2\right)\simeq\frac{1}{N_2}\log\frac{P\left(X_2|\lambda_1\right)}{P\left(X_2|\lambda_2\right)} \text{ and } D\left(X_2,X_1\right)\simeq\frac{1}{N_1}\log\frac{P\left(X_1|\lambda_2\right)}{P\left(X_1|\lambda_1\right)}$$

By combining the two formulations, they classified six DNA sequences, taken from the threonine operons of *Escherichia coli* and *Shigella flexneri*, and also applied the dissimilarity measure to the search of a database, attaining good levels of sensitivity and selectivity.

The Kullback-Leibler discrepancy was also recently explored in another article (Wu *et al.*, 2005), under a more general formulation of dissimilarities, the Cressie-Read family of discrepancies of parameter λ. Following previous work (Wu *et al.*, 2001), the authors directly applied a symmetric version of this measure for λ=0, which corresponds exactly to the symetrized version of the Kullback-Leibler discrepancy, or SK-LD, and study its dependence from mutational rates, window sizes and word lengths of sequences under study. This objective was fulfilled by performing simulations of DNA sequences with distinct known mutation percentages from an original set and by studying the dissimilarities obtained when several window lengths and *L*-tuple resolutions are considered. The SK-LD is defined as a minimum of the dissimilarities taken for all the *W* length windows:

$$d_L^{SKLD}\left(X,Y\right)=\min_W\left(\frac{I_L^W\left(X,Y\right)+I_L^W\left(Y,X\right)}{2}\right)$$
$$I_L^W\left(X,Y\right)=\frac{1}{\left(W-L+1\right)+0.5}\sum_{i=1}^{4^L}\left(c_{L,i}^X+\varepsilon_L\right)\log\frac{c_{L,i}^X+\varepsilon_L}{c_{L,i}^Y+\varepsilon_L} \tag{15}$$

where the pseudo-counts $\varepsilon_L=0.5\times4^{-L}$ avoid the possibility of having infinity values.

The results are also compared with alignment algorithms, represented by BLAST, and other metrics, quantitatively assessing their relative accuracy. The SK-LD method is further applied to four different experiments. The first comprises the classification of six threonine operons, showing good agreement with previous methods. The second case consisted on searching databases with a query sequence, thus assessing the better sensitivity and selectivity of SK-LD. The third experiment is the classification of shuffled ORF (open reading frames) sequences, demonstrating that this measure is much better that BLAST in presence of genome rearrangements. Finally, using the results obtained in the last experiment, they further suggest the use of this dissimilarity for selecting oligo probes for use in gene expression microarray design.

Very recently, a different approach to sequence classification was proposed, based on word ranks (Goldberger and Peng, 2005). The key idea was to sort the abundance of all the *L*-tuples present in two different sequences *X* and *Y*, i.e. the vectors f_L^X and f_L^Y, and compare the order or rank $R\left(w_i\right)$ of each word w_i in those sorted vectors. For two similar texts the plot of the ranks $\left(R_X\left(w_i\right), R_Y\left(w_i\right)\right)$ would be a straight diagonal line: each word occupying the same relative position, i.e., the same abundance order, in the two sequences. By analyzing the dispersion of the points from the diagonal the authors define a measure of dissimilarity between the sequences or an *information based similarity index*:

$$d_L^{rank}\left(X,Y\right)=\frac{1}{K-1}\sum_{i=1}^{K}\left|R_X\left(w_i\right)-R_Y\left(w_i\right)\right|\frac{H_X\left(w_i\right)-H_Y\left(w_i\right)}{\sum_{i=1}^{K}\left(H_X\left(w_i\right)-H_Y\left(w_i\right)\right)} \qquad \textbf{(16)}$$

The absolute difference of ranks $\left|R_X\left(w_i\right)-R_Y\left(w_i\right)\right|$ is proportional to the Euclidean distance from a given point to the diagonal line. This term is then weighted by the Shannon entropy $H\left(w_i\right)=-p_{w_i}\ln p_{w_i}$ that measures the relative abundance of the word w_i: the more frequent tuples will have a higher contribution for the overall dissimilarity between X and Y. This measure can be interpreted as a weighted city-block dissimilarity on the rank order. In a way, it filters the repetitions because its main contribution is from the overall ranking or order of one particular word or tuple, ignoring the high discrepant values of their absolute frequencies.

The authors have applied this dissimilarity to several types of texts from different alphabets: in particular, they have used the results to classify the authorship of English (e.g. Shakespeare plays) and Chinese literature (Yang *et al.*, 2003b), the type of cardiac pathology derived from interbeat interval time series (Yang *et al.*, 2003a) and the genomic classification of the SARS Coronavirus (Goldberger and Peng, 2005).

The results are very promising, showing that the combination of information-theory and rank-order statistics might have a wide range of applications in the future for large-scale analysis of genomic databases.

Several methods cut across different sections described above. For example, in a recent paper (Gangal and Sharma, 2005) the classification of sequences was performed by using CGR, along with non-linear time series descriptions based on the Lyaponov exponent and on the Tsallis entropy. The information was used to train an SVM model and correctly discriminate between promoter and non-promoter regions.

Another type of analysis related with entropy and complexity concepts was developed recently. In these studies, the characterization of the *linguistic complexity* of genomes is related with the notion of self repetitiveness. This quantity is calculated by using a sliding window and assessing the ratio of the number of all present L-tuples over the total number of possible combinations, given by 4^L. This means that in highly repetitive regions, this fraction will be low: only a small percentage of all possible substrings are used. In contrast, the windows with more distinct L-tuples will have a higher ratio, corresponding to higher variability. This approach was used in several studies, e.g. (Crochemore and Verin, 1999; Troyanskaya *et al.*, 2002). The notion here defined is closely connected with the entropy definition. In fact, high complexity zones correspond to high topological entropy and low complexity zones, as measured by the relative variability of sub-strings, are related with lower entropy or high level of repetitiveness. This alignment-free methodology was shown to be useful to determine new biological features in *S. cerevisiae* yeast chromosomes, filtering regular regions (Crochemore and Verin, 1999). The authors also propose a measure of dissimilarity between windows using the Jaccard index, i.e., the ratio between the common factors and the total number present in both sequences. In (Troyanskaya *et al.*, 2002) the calculation of the linguistic complexity is efficiently computed in linear time using suffix trees and the algorithms were applied to the complete genome of *Haemophilus influenzae*.

These methods are clearly related with complexity and entropy as referred to before when calculating the Kolmogorov complexity of mitochondrial DNA (Li *et al.*, 2001) and the complexity based on compression methods, with programs such as GenCompress (Chen *et al.*, 1999) and, more recently, other measures of dissimilarity based on the Burrows-Wheeler transform (Mantaci *et al.*, 2005).

These results illustrate the relation between entropy and complexity concepts, but are still dependent on the resolution of the tuples considered. In the works reviewed a specific value for *L* is implicit and required in all the analysis. Other "resolution-free" approaches were also pursued. The CGR/USM maps might be classified in this category and can be associated with the genomic signature concept described below. Other type or resolution-free methodologies is based on the complexity of strings and their compressibility.

Following the work of Li and colleagues (Li *et al.*, 2001) that introduced a dissimilarity measure based on the Kolmogorov complexity $d^{KC}(X,Y)$ to compare whole mitochondrial unaligned genomes, another metric was proposed for phylogenetic tree construction (Otu and Sayood, 2003). In this work a distance between sequences $d^{LZ}(X,Y)$ is defined, based on the Lempel-Ziv (LZ) complexity of *X*, *c*(X), which is related to the number of steps required by a process that builds the sequence *X*. The distance between two sequences *X* and *Y* is obtained by comparing the LZ complexity of *X* (and *Y*) with their concatenation *XY* (or *YX*):

$$d^{LZ}(X,Y) = \max\left\{c(XY) - c(X), c(YX) - c(Y)\right\} \qquad \textbf{(17)}$$

The key idea is that two similar sequences will have common regularities, thus leading to one partly "explaining" the other in terms of processing steps. Among the results presented in this work are the proofs of the distance properties of $d^{LZ}(X,Y)$ (Eq.9) and the tests on simulated sequences, showing the better performance of this metric when compared with maximum likelihood and parsimony methods. They further apply the new method to real mitochondrial genome data, achieving phylogenetic trees in agreement with several other studies. One important aspect of this methodology is that the whole genome is used, instead of partial data. Furthermore, it is an alignment-free, automatic algorithm, thus avoiding several problems encountered when performing multiple alignments, such as the presence of gene rearrangements, inversions, transpositions and translocations and unequal length of the sequences.

The above defined relative complexity measure (RCM) was subsequently applied to construct phylogenetic trees for fungi from the mitochondrial cytochrome b gene, the 18S rDNA gene and the ITS-1 (internal transcribed spacer) and ITS-2 regions of the rDNA gene complex (Bastola *et al.*, 2004). This study showed the potential of this approach as a reliable and robust method to infer phylogenetic relationships.

3.3 Genomic signatures: from L-tuple composition to CGR maps

The previous sections overviewed some alignment-free methodologies based on vector maps and several dissimilarities in the image space. This section also describes articles using the same type of algorithms but where the notion of genomic signature was applied to some extend.

The introduction of the term genomic signature dates back to the mid 90's (Karlin and Burge, 1995) following a previous work on oligonucleotide over and under-representation (Burge *et al.*, 1992). The initial results showed that the set of dinucleotide (or 2-tuples) normalized frequency vectors constituted a signature of one organism. This means that there are significant differences between intra and inter-species, as assigned by the odds ratio between 2-tuple frequencies normalized by expected values. The odds ratio $\rho_{s_i s_j}$ represents the dinucleotide bias of the 2-tuple $s_i s_j$, and is expressed as a function of its frequency $f_{s_i s_j}$ and the value expected under a zero order Markov model, $f_{s_i} \cdot f_{s_j}$, which is simply the product of the frequency of symbols $s_i, s_j \in \mathcal{A} = \{A, T, C, G\}$:

$$\rho_{s_i s_j} = \frac{f_{s_i s_j}}{f_{s_i} \cdot f_{s_j}} \qquad \textbf{(18)}$$

This formula is modified to accommodate double-stranded DNA, leading to a symetrized version, computed from the frequencies of the sequence concatenated with its inverted complementary version. This new version is denoted ρ^* and is obtained by using the frequencies $f^*[A] = f^*[T] = (f[A] + f[T]) / 2$ and $f^*[C] = f^*[G] = (f[C] + f[G]) / 2$. A wide collection of results is presented, showing that these odds ratios can be effectively applied to discriminate between sequences from different organisms, defining species-specific properties of DNA mechanisms.

The authors subsequently defined a dissimilarity based on ρ^*: the dinucleotide relative abundance distance or δ-distance between two sequences *X* and *Y* (Karlin and Ladunga, 1994).

$$d_2^{\delta}(X,Y) = \frac{1}{16} \sum_{i=1}^{16} \left| \rho_i^{*X} - \rho_i^{*Y} \right| \qquad \textbf{(19)}$$

where the sum extends over all possible 4^2 dinucleotides.

The relative abundance distance measure is extensively applied to large DNA sets, demonstrating that the δ-distance always differ more for genomic interspecific sequence comparisons than for genomic intraspecific sequence comparisons, indicating congruence over different genome sequence samples.

The impact of these results is reflected by several articles comparing the effective goodness of classification by using this 16-vector set. In fact, when comparing ρ* for different organisms this 16-dimensional vector is found to exhibit a remarkable stability within the same genome and can distinguish between sequences from different organism, justifying the introduction of a "signature" concept (Campbell *et al.*, 1999; Karlin, 2001)

Other studies dealing with the genomic signature concept have further been developed, e.g., the study of how pervasive are genomic signatures for the same species (Jernigan and Baran, 2002). The overall vales ρ* (taken for the whole sequence) have been compared with those obtained for small windows, using the δ-distance above defined (Eq.19). The results show that signatures are preserved intra-species even for short window lengths sizes. Additionally, the convergence properties of this measure were quantitatively defined and the stability of genomic signature in windows ranging in size *n* from 50kb down to 125 bases was assessed, showing that the global signature is locally persistent in all the scales scanned. In addition, it is shown that the intra-genomic δ-distance is approximately log-normally distributed, i.e., obeys a log-linear model of the form $\log\delta^* = \alpha - \beta\log n$. This fact amply corroborates the genomic signature concept previously introduced.

Departing from the original proposal, Hooper and Berg introduced a variation of the original genomic signature concept (Hooper and Berg, 2002), by using only the dinucleotides frequencies in genes. The *genomic 3:1 signature* therein defined is computed by taking the frequencies of the 2-tuples composed by the 3rd codon position and the 1st of the succeeding codon in the sequence. This frequency is the least restricted by aminoacid preference and codon usage, therefore being the most flexible and coding-independent dinucleotide bias on the gene level. Several microbial genomes were compared, including *E. coli* and *B. subtilis*, and genes with significant deviations from their genomic signatures were found, which might indicate horizontal transfer between the two genomes. The differences between inter and intra-species comparisons were statistical significant, strengthening the conclusions of previous studies.

These results show the importance of 2-tuple normalized frequencies as DNA sequence descriptors. A natural development of these results was the extension of the procedure to higher-order tuples, gaining more discriminatory precision.

It is noteworthy that Deschavanne and colleagues (Deschavanne *et al.*, 1999) have implemented the genomic signature concept using CGR maps. In fact, as previously referred to, the chaos game representation allows the depiction of all *L*-tuples in the same image, thus making possible the generalization of the signature concept to higher order oligonucleotides, with important phylogenetic implications (Hill and Singh, 1997). In particular, an 8-tuple resolution was applied to scan several organisms, representing Eukaryote, Eubacteria and Archaebacteria. Principal component analyses (PCA) of the vectors thus obtained correctly discriminate between domains of life. In fact, by observing the CGR images one can immediately recognize patterns and characteristic features, in a very appealing and straightforward format. Euclidean distances between those image matrices were further used to create a dissimilarity measure between genomes and classify the sequences. One interesting result observed in this pioneer work was that short sequences images are somewhat the blurred versions of the complete genome data. This shows that different parts of the genome follow the patters of the whole sequence, i.e. exhibits intra-species stability, which amply corroborates the genomic signature concept. This important generalization was explored in other works, showing the interconnection between *L*-tuple counts and chaos games.

In another work the genomic signature concept was implicitly used for the classification of short-sequences (Sandberg *et al.*, 2001). This work starts by defining frequencies profiles for the genomes by calculating their *L*-tuple frequencies f_L. These profiles represent a feature of one particular sequence and use the results previously obtained that inter-species differences are larger than intra-species comparisons. It is then possible to classify a new sequence with a naïve Bayesian classifier to assign sequences to classes, using maximum *a posteriori* estimates. This is equivalent to the classification of a new sequence in the class defined by the "closest" oligonucleotide frequency profile. Another important achievement of this work was the quantitative analysis of the classification accuracy for different resolutions (*L*) and for distinct percentage deletions of the training set (the original sequences from which the profiles are calculated). For example, bacterial sequences with only 400 nucleotides can be classified with accuracies above 85%, for optimal 8 and 9-tuple frequencies. Moreover, it was confirmed that inter-genomic differences are greater that the intra-genomic values. This property has motivated the use of these tools to infer horizontal transfer (HT) between bacterial genomes, by considering sliding windows of fixed length and calculating their profiles separately, searching for heterogeneities.

In another study from the same authors the dependence and relationship between the genomic signature and several bias factors was further examined (Sandberg *et al.*, 2003). In particular, the correlation between the *L*-tuple frequencies and the G+C content, synonymous codon choice and aminoacid usage was investigated. By using a non-normalized version of the 3, 6 and 9-tuple frequencies along with Euclidean distances between those vectors for 57 organisms, the correlation between those factors was quantified.

Following these results, the genomic signature concept has recently been explored to detect and characterize horizontal transfers in prokaryotes (Dufraigne *et al.*, 2005). Interestingly, CGR maps were again used as generalization of *L*-tuple frequencies, in one key application of iterated function systems to genome analysis. The notion of *local signature* is defined in this work by using the 4-tuple normalized frequencies through specific sliding windows. In this way it is possible to create a "profile", where atypical windows and/or positions with distinct compositions will be detected, i.e., the most heterogeneous ones. These atypical local signatures may point out DNA transfers and, in addition, might provide clues about their possible origin. The extensive study of 22 genomes, including *E. coli*, *B. subtilis* and *H. influenza*, and the comparison with other results further corroborate this statement.

The previous concepts lead to a central paper where the *spectrum of genomic signatures* was defined (Wang *et al.*, 2005). The authors consider Karlin's definition as a particular case obtained for 2-tuples, named dinucleotide relative abundance profiles (DARP), and extend it to higher order nucleotides using CGR. In other words, the genomic signature concept defined before, based on 2-tuples, is interpreted as a particular case of a broader collection or spectrum of signatures. The CGR, given its flexibility and convenience, is used as a description of the spectra. They also explore several types of dissimilarity measures between CGR images and compare the results with phylogeny studies. The δ-distance defined above (Eq.19) is shown to be a special case of the Hamming distance between CGR maps. The image distance was also introduced, a new dissimilarity measure between CGR maps that expands the Hamming definition. The key idea is to average neighborhoods of intervals (representing *L*-tuples) and to consider the differences in those masked values. Another important result, obtained by simulation, was that 3-order Markov chains cannot explain the images obtained for real organisms, contradicting the results of (Goldman, 1993). They defend that the DNA structure is too refined to be captured only by a low-order model, with few parameters. The method is applied to the classification of 26 mitochondrial DNA, showing a good agreement with former phylogenetic studies.

It is important, at this point, to stress the relationship between *L*-tuples and chaos game representation, described in the background section: in fact, CGR is a generalization of *L*-tuples frequencies and Markov chains and can be effectively used as a good representation for machine learning algorithms. Furthermore, it is not necessary to define the resolution *a priori*, a main advantage of CGR maps, which might be included in the optimization procedure that will investigate and automatically chose the best combination of *L*-tuples for specific discriminatory analysis.

In a recent article, the genomic signature concept along with CGR maps was used to distinguish between Eukaryotic coding and non-coding sequences (Nikolaou and Almirantis, 2005). The authors propose a measure $S^{(L)}$ which is the sum of all the *L*-tuple frequencies, filtering the background mononucleotide composition:

$$\begin{aligned} & R_{ijk\ldots q}{}^{(L)} = \frac{f_{ijk\ldots q}{}^{(L)}}{f_i^{(L)} f_j^{(L)} f_k^{(L)} \cdots f_q^{(L)}}\,; \quad i,j,k,\ldots,q \in \{A,T,C,G\} \\ & S^{(L)} = \sum R_{ijk\ldots q}{}^{(L)} \end{aligned} \qquad \textbf{(20)}$$

In particular, 5-tuples are chosen as an optimum value to discriminate between introns and coding sequences (CDS) and a surrogate sequence collection obtained by shuffling. An evolutionary model is also formulated, which was able do recreate the statistical properties of the sequences considered.

The prolific genomic signature concept was further explored to higher-order tuples using machine learning techniques. In particular, the species-specificity was assessed through the use of self-organizing maps (SOM's) applied to di-, tri- and tetranucleotide frequencies (Abe *et al.*, 2003). A SOM is an unsupervised neural network algorithm which converts complex nonlinear relations among high-dimensional data in simple geometric relations. In practice, SOM's implement nonlinear projections from the multidimensional space of input data onto a two-dimensional array, utilizing complex key combination of oligonucleotide frequencies for the sequence separations. The resulting SOM's revealed clear separation/discrimination between species, easily verified with this technique.

The genomic signature concept has also motivated several software programs that efficiently analyze oligonucleotide frequencies, such as TETRA (Teeling *et al.*, 2004), a tool that might be used as a fingerprint technique for genomic fragment correlation. The algorithm normalizes tetranucleotide frequencies by their expected values under a maximal-order Markov model, which is calculated from 2 and 3-tuple composition (Eq.8). The obtained z-scores are subsequently used to calculate Pearson correlation coefficients and ultimately classify the fragments. Other works have also explored the tetranucleotide frequency to classify genomes (Yap *et al.*, 2003).

Another implementation of alignment-free sequence analysis based on the genomic signature paradigm was developed very recently (Fertil *et al.*, 2005). GENSTYLE is a workspace with several toolboxes that allows the characterization and classification of nucleotide sequences (available online at http://genstyle.imed.jussieu.fr/). It depicts CGR images and fully describes their statistical properties and profiles, allowing the detection of atypical areas in the genome. There is also a genomic signature database with information on different species, which enhances the scope and power of the analysis in comparative genomics. These methodologies were already explored for the detection of horizontal transfer in bacteria in a recent paper (Regeard *et al.*, 2005).

As seen it this section, a wealth of theoretical and experimental work has been performed underlying the genomic signatures concept, from 2-tuple composition to more sophisticated methodologies, combining several other data mining techniques, to achieve a wide scope of applications. One conclusion to retain is the constancy of some frequency-based descriptors among the same species, even taken for relatively short sequences, and the evident differences seen for different organisms, which undoubtedly substantiate the genomic signature spectrum concept.

3.4 Proteins

The previous sections have focused on the application of vector maps to DNA sequences, overlooking the wealth of algorithms currently in use for protein analysis and comparison. Although the basic ideas and dissimilarity measures are the same as the ones described in the previous sections, some of the specificity of the protein datasets, namely its alphabet higher dimensionality, justifies the separate analysis of the comparison methodologies.

A recent study proposed the use of weighted aminoacid composition to classify proteins (Vinga *et al.*, 2004). The basic idea was to combine *L*-tuple composition (in this case, 1-tuple, omitted in the equation) with evolutionary information through a quadratic form:

$$d^W(X,Y) = \left(f^X - f^Y\right)^T \cdot \mathbf{W} \cdot \left(f^X - f^Y\right) = \sum_{i\in\mathcal{A}}\sum_{j\in\mathcal{A}}\left(f_i^X - f_i^Y\right)\cdot\left(f_j^X - f_j^Y\right)\cdot w_{ij} \qquad \textbf{(21)}$$

The weight matrices **W** in equation 21 can be rationalized as being scoring or aminoacid substitution matrices, instead of covariance-based weights as in other distances such as standard Euclidean $d_L^{SE}(X,Y)$ and Mahalanobis $d_L^M(X,Y)$. These matrices, such as PAM – Point Accepted Mutation (Dayhoff *et al.*, 1978) and BLOSUM – BLOcks SUbstitution Matrices (Henikoff and Henikoff, 1992), are used in alignment-based methods and estimate the log-likelihood ratios between probabilities of symbols that best describe mutation rates in known homologous proteins.

The results were applied to the SCOP database and correctly classify family and superfamily levels. In this paper the quantitative assessment of those accuracies is compared across protein hierarchical levels (family, superfamily, fold and class) and different metrics, using Receiver Operating Characteristic (ROC) curves.

Aminoacid composition was also recently used to discriminate between several protein categories, in particular the beta-barrel transmembrane (bbtm) proteins from other types (Garrow *et al.*, 2005). The authors showed that the combination of aminoacid composition with evolutionary information leads to a better accuracy. (See (Garrow *et al.*, 2005) and references therein for other aminoacid base classification methods for distinct protein categories.). Squared Euclidean distances between the aminoacid frequency vectors and the k-nearest neighbor algorithms were used to classify the proteins in classes. A weighted Euclidean distance (similar to the d2 metric described above) was also tested, where the weights are optimized using a genetic algorithm. Another weighting procedure is done with the frequency vectors themselves, averaging for several homologous proteins. The overall accuracies are high, when compared with other methods.

In another study (Qi *et al.*, 2004), the authors propose the same approach of contrast vocabularies seen above for DNA sequences (Pietrokovski, 1994). Proteins are mapped onto their 6-tuple composition, i.e., considering the frequency of all the oligopeptides constituted by 6 aminoacids. The expected probability of each L-tuple $\alpha_1\alpha_2\cdots\alpha_L$, assuming a maximal order Markov model, is estimated from the observed (L-2) and (L-1)-tuples frequencies:

$$p^0_{\alpha_1\alpha_2\cdots\alpha_L} = \frac{f_{\alpha_1\cdots\alpha_{L-1}} \cdot f_{\alpha_2\cdots\alpha_L}}{f_{\alpha_2\cdots\alpha_{L-1}}}$$

The observed frequencies f are then subject to a normalization step:

$$a_{\alpha_1\alpha_2\cdots\alpha_L} = \begin{cases} \dfrac{f_{\alpha_1\cdots\alpha_L} - p^0_{\alpha_1\cdots\alpha_L}}{p^0_{\alpha_1\cdots\alpha_L}} & \text{, when } p^0_{\alpha_1\cdots\alpha_L} \neq 0 \\ 0 & \text{, when } p^0_{\alpha_1\cdots\alpha_L} = 0 \end{cases} \qquad \textbf{(22)}$$

This expression is similar to Eq.11 but divides the deviations between f and p^0 by their estimated expected values. Following this normalization, the dissimilarity is obtained by considering, as above, the cosine of the angle between two vectors a^X and a^Y obtained from sequence X and Y, subsequently normalized to the range between 0 and 1 (instead of -1,1).

This alignment-free measure is proposed as an evolutionary distance between species and phylogenetic trees are constructed from whole proteome prokaryote data. The statistical significance of the trees was also taken into account in this study and was assessed by bootstrap and jackknife methods. The phylogenetic trees thus obtained show excellent agreement with other evolutionary studies. In fact, the authors suggest a “K-string picture of evolution”, where a small fraction of primordial polypeptides (e.g. from the possible 20^6=64,000,000, considering K=6) evolved in this K-tuple space by growth, fusion and mutation.

Following directly this work, other recent article explored whole genome analysis for evolutionary studies using frequency estimates (Yu *et al.*, 2005). The background frequencies from shorter *L*-tuples can be filtered using the equation 23:

$$q_{\alpha_1\alpha_2\cdots\alpha_k} = \frac{f_{\alpha_1\alpha_2\cdots\alpha_{k-1}} f_{\alpha_k} + f_{\alpha_1} f_{\alpha_2\alpha_3\cdots\alpha_k}}{2}$$

$$X_{\alpha_1\alpha_2\cdots\alpha_k} = \begin{cases} f_{\alpha_1\alpha_2\cdots\alpha_k} \big/ q_{\alpha_1\alpha_2\cdots\alpha_k} - 1 \text{ , if } q_{\alpha_1\alpha_2\cdots\alpha_k} \neq 0 \\ 0 \qquad\qquad\qquad\quad \text{ , if } q_{\alpha_1\alpha_2\cdots\alpha_k} = 0 \end{cases} \qquad \textbf{(23)}$$

The application of angle metrics to the vectors $X_i \equiv X_{\alpha_1\alpha_2\cdots\alpha_k}$, as previously described, allows the classification of all protein sequences and the inference of phylogenetic trees that are in agreement with evolutionary information. Once more, the analysis of the compositional statistics of proteins from complete genomes is proven to be a fast and efficient method for phylogenetic reconstruction. This alignment-free approach avoids the problems inherent to alignment-based algorithms when gene rearrangements are present.

Other type of filtering based on SVD (single value decomposition) was recently applied to classify proteins (Stuart and Berry, 2004). This technique had previously been employed in mitochondrial DNA classification, conjugated with the angle metric referred to above (Stuart *et al.*, 2002a; Stuart *et al.*, 2002b). In this new work the 4-tuple composition, i.e. the tetrapeptides frequencies, are decomposed by the SVD algorithm in families that are characterized by linear combinations of the original peptides. These transformed vectors provide motif models and can be useful in the annotation of unknown proteins and/or their accurate clustering, as demonstrated for nine whole eukaryotic genomes.

A robust method for protein alignment-free classification was recently proposed (Sperisen and Pagni, 2005), named JACOP (Just Another Classification Of Proteins). The protocol consists on randomly selecting 50-length probes from the target protein sequences, which are subsequently filtered, pruning high similarity pairs as to obtain a subset of distinct representative probes. Then the original proteins are compared by alignment with all the probes selected, keeping track, for each protein, of the probes with score higher than a specific threshold, transformed in a binary matrix with ones (above) and zeros (below). The proteins are then compared using the Jaccard distance which accounts for the differences, normalized by the total number of comparisons but excluding the common zeros. This matrix is then used to identify independent groups, to partition the proteins within each group and finally to cluster the complete set. The Jaccard distance guarantees that proteins with no similarity other than noise will not be grouped together. The method performance is tested in two cases. The first one includes prokaryotic lyases, which can be studied using multiple sequence alignment. The second dataset contains all proteins from Swiss-Prot with at least one Src homology domain, a challenging case where it is not possible to arrange them in a meaningful sequence alignment, given the putative reshuffling events. Both cases demonstrate the potential use of this methodology. In fact, the random probes used as an intermediate step allow the uncoupling of the domain architecture, since contiguity – one of the problems with alignment methods – is not required to any further extent. This work, as the one referred to before (Edgar, 2004), highlights the usefulness in combining alignment-based and alignment-free methods for sequence comparison.

The notion of genomic signature described above for DNA (Karlin and Burge, 1995) is shown to have a protein counterpart. In a recent study, an extensive and comprehensive analysis of all the proteome of several species of distinct kingdoms (Eukaryotes, Eubacteria and Archaea) was performed (Pe'er *et al.*, 2004). The analysis was based on the extraction of the aminoacid, di- and tri-peptide composition, i.e. the 1 to 3-tuple frequencies respectively. These vectors were subsequently projected in a two dimensional (2D) space by Principal Component Analysis (PCA), as to extract the main features of each species and the obtained lower dimensional vectors were compared. As noted by other articles (see (Pe'er *et al.*, 2004) and references therein), the species can be differentiated by the *L*-tuple composition of their proteomes, which suggests the use of species-specific compositions as *proteomic signatures*, analogous to the genomic signatures introduced before for DNA sequences. Hierarchical clustering was further applied to the proteins using pairwise standard Euclidean distances between the amino acid frequency vectors, thus creating dendrograms in agreement with phylogenetic relationships. This suggests that closely related proteomes display similar compositions. Furthermore, via whole proteome analysis it is possible to avoid several problems encountered when using single gene methods, such as the “unrecognized horizontal gene transfer, unrecognized paralogy, highly variable rates of gene evolution, or misalignment”.

Other recent endeavor that explicitly uses alignment-free methods was the analysis of correlation between protein aminoacid composition and its corresponding gene expression level ((Raghava and Han, 2005) and references therein for other types of correlation analysis). The analysis of 3468 genes of *Saccharomyces cerevisiae* shows that there are important correlations between the expression of a gene and the 1-tuple composition of its protein, extracted from its primary structure. This promising result might allow the prediction of expression levels from the protein sequences, from microarray data in a given condition. The algorithms, available online, are also based on Support Vector Machines (SVM).

This section presented some applications of vector maps in proteomics, highlighting some of the problems and specificities of this type of data. Several studies dealt with the characteristic of proteins from a composition point of view, leading to fast algorithms to analyze and classify aminoacid sequences.

4 Conclusion

Biological sequence analysis is still a key step in most bioinformatics applications, in tasks that range from searching databases to the inference of regulatory networks and integrative systems biology.

This chapter briefly overviewed the fundamental theoretical basis and the new trends on alignment-free methodologies for sequence comparison. These techniques contrast with alignment-based methods, the dominant paradigm in the field, ubiquitously used in the majority of the bioinformatics applications.

Vector maps represent a vast category for representing biological sequences, providing efficient and elegant algorithms for their study. They are based on representing sequences as *n*-dimensional vectors, exploiting useful linear algebra results and information theory approaches to this image space. In fact, as amply demonstrated in this chapter, alignment-free techniques can deal with several problems that arise in alignment-based algorithms, such as genome recombination and shuffling, thus providing useful alternatives when the later fail.

This chapter was intended to outline some recent literature on this issue, revisiting previous surveys and thus providing a wider view of this subject, which is expected to have high impact in the future with the continuous explosive growth of biological data. This work provides a starting point to explore further alignment-free techniques for sequence analysis and comparison, systematizing the nomenclature and describing possible future applications.

This rich field, proven to be highly productive in the past years, has been the source of a recent wave of new papers and techniques, widening their application to a total novel set of problems, as this chapter briefly illustrated. In fact, the wealth of new methods makes it difficult to perform an exhaustive analysis and to explore all possible applications. The extensions in same cases are straightforward, and it is envisaged that full automated methods to analyze the results would be also studied in the future.

As referred to and highlighted in the literature, alignment-free methods have several advantages over other methods, from their computational efficiency to the possibility of using whole genome and whole proteome for the analysis, thus really using all the information available nowadays. In phylogenetic studies, although not incorporating molecular evolution methodology, they nonetheless perform extremely well, suggesting an underlying important feature about evolutive processes, that somehow maintain sequence signature and features, which should be further investigated.

All these results suggest that vector maps and alignment-free techniques can provide new tools to classify, analyze and integrate biological sequence data.

Acknowledgements

The author thanks the Biomathematics Group at Instituto de Tecnologia Química e Biológica, Universidade Nova de Lisboa (ITQB/UNL) for using the computer resources during the preparation of this chapter. SV also thanks Prof. Jonas Almeida from the MD Anderson Cancer Center (University of Texas) for important discussions on alignment-free methodologies, Prof. Arlindo Oliveira from INESC-ID for careful reading the manuscript and providing valuable corrections and suggestions and Simone Preziati for useful comments that greatly improved this chapter. This work was partially supported by projects DynaMo (PTDC/EEA-ACR/69530/2006 – S.Vinga, PI) from the Portuguese Science Foundation (FCT) and MaGiC (IE02ID01004 – A.T.Freitas, PI) from INESC-ID.

5 References

Abe, T., Kanaya, S., Kinouchi, M., Ichiba, Y., Kozuki, T. & Ikemura, T. (2003). Informatics for unveiling hidden genome signatures. *Genome Res, 13*, 693-702.

Almeida, J. S., Carriço, J. A., Maretzek, A., Noble, P. A. & Fletcher, M. (2001). Analysis of Genomic Sequences by Chaos Game Representation. *Bioinformatics, 17*, 429-437.

Almeida, J. S. & Vinga, S. (2002). Universal sequence map (USM) of arbitrary discrete sequences. *BMC Bioinformatics, 3,* 6.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990). Basic Local Alignment Search Tool. *Journal of Molecular Biology, 215*, 403-410.

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res, 25,* 3389-3402.

Bastola, D. R., Otu, H. H., Doukas, S. E., Sayood, K., Hinrichs, S. H. & Iwen, P. C. (2004). Utilization of the relative complexity measure to construct a phylogenetic tree for fungi. *Mycol Res, 108*, 117-125.

Blaisdell, B. E. (1986). A measure of the similarity of sets of sequences not requiring sequence alignment. *Proc Natl Acad Sci U S A, 83*, 5155-5159.

Bolshoy, A. (2003). DNA sequence analysis linguistic tools: contrast vocabularies, compositional spectra and linguistic complexity. *Appl Bioinformatics, 2*, 103-112.

Bronson, E. C. & Anderson, J. N. (1994). Nucleotide composition as a driving force in the evolution of retroviruses. *J Mol Evol, 38*, 506-532.

Burge, C., Campbell, A. M. & Karlin, S. (1992). Over- and under-representation of short oligonucleotides in DNA sequences. *Proc Natl Acad Sci U S A, 89*, 1358-1362.

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

Chen, L. Y., Lu, S. H., Shih, E. S. & Hwang, M. J. (2002). Single nucleotide polymorphism mapping using genome-wide unique sequences. *Genome Res, 12*, 1106-1111.

Chen, X., Kwong, S. & Li, M. (1999). A Compression Algorithm for DNA Sequences and Its Applications in Genome Comparison. *Genome Inform Ser Workshop Genome Inform, 10*, 51-61.

Chen, Y. H., Nyeo, S. L. & Yeh, C. Y. (2005). Model for the distributions of k-mers in DNA sequences. *Phys Rev E Stat Nonlin Soft Matter Phys, 72,* 011908.

Cover, T. M. & Thomas, J. A. (1991). *Elements of Information Theory*. New York: Wiley.

Crochemore, M. & Verin, R. (1999). Zones of low entropy in genomic sequences. *Comput Chem, 23,* 275-282.

Dayhoff, M. O., Schwartz, R. & Orcutt, B. (1978). A model of evolutionary change in proteins. In M. O. Dayhoff (Ed.), *Atlas of protein sequence and structure* (pp. 345-352). Washington, D.C.: National Biomedical Research Foundation.

Dehnert, M., Helm, W. E. & Hutt, M. T. (2005). Information theory reveals large-scale synchronisation of statistical correlations in eukaryote genomes. *Gene, 345,* 81-90.

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

Dufraigne, C., Fertil, B., Lespinats, S., Giron, A. & Deschavanne, P. (2005). Detection and characterization of horizontal transfers in prokaryotes using genomic signature. *Nucleic Acids Res, 33*, e6.

Durbin, R., Eddy, S. R., Krogh, A. & Mitchison, G. (1998). *Biological Sequence Analysis*. Cambridge University Press.

Edgar, G. A. (1990). *Measure, topology, and fractal geometry*. New York: Springer-Verlag.

Edgar, R. C. (2004). MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics, 5*, 113.

Everitt, B., Landau, S. & Leese, M. (2001). *Cluster Analysis*. London: Arnold.

Fertil, B., Massin, M., Lespinats, S., Devic, C., Dumee, P. & Giron, A. (2005). GENSTYLE: exploration and analysis of DNA sequences with genomic signature. *Nucleic Acids Res, 33*, W512-515.

Fofanov, Y., Luo, Y., Katili, C., Wang, J., Belosludtsev, Y., Powdrill, T., Belapurkar, C., Fofanov, V., Li, T. B., Chumakov, S. & Pettitt, B. M. (2004). How independent are the appearances of n-mers in different genomes? *Bioinformatics, 20*, 2421-2428.

Gangal, R. & Sharma, P. (2005). Human pol II promoter prediction: time series descriptors and machine learning. *Nucleic Acids Res, 33*, 1332-1336.

García, J. A. & José, M. V. (2005). Mathematical properties of DNA sequences from coding and noncoding regions. *Revista Mexicana de Física, 51*, 122-130.

Garrow, A. G., Agnew, A. & Westhead, D. R. (2005). TMB-Hunt: an amino acid composition based method to screen proteomes for beta-barrel transmembrane proteins. *BMC Bioinformatics, 6*, 56.

Gibbs, M. J., Armstrong, J. S. & Gibbs, A. J. (2005). Individual sequences in large sets of gene sequences may be distinguished efficiently by combinations of shared sub-sequences. *BMC Bioinformatics, 6*, 90.

Goldberger, A. L. & Peng, C. K. (2005). Genomic Classification Using an Information-Based Similarity Index: Application to the SARS Coronavirus. *J Comput Biol, 12*, 1103-1116.

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

Gordon, A. D. (1999). *Classification*. Boca Raton: Chapman & Hall/CRC.

Gusfield, D. (1997). *Algorithms on strings, trees, and sequences: computer science and computational biology*. Cambridge [England]; New York: Cambridge University Press.

Haubold, B., Pierstorff, N., Moller, F. & Wiehe, T. (2005). Genome comparison without alignment using shortest unique substrings. *BMC Bioinformatics, 6*, 123.

Henikoff, S. & Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci U S A, 89*, 10915-10919.

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.

Hooper, S. D. & Berg, O. G. (2002). Detection of genes with atypical nucleotide sequence in microbial genomes. *J Mol Evol, 54*, 365-375.

Jeffrey, H. J. (1990). Chaos Game Representation of Gene Structure. *Nucleic Acids Res, 18,* 2163-2170.

Jernigan, R. W. & Baran, R. H. (2002). Pervasive properties of the genomic signature. *BMC Genomics, 3,* 23.

Kanehisa, M. & Bork, P. (2003). Bioinformatics in the post-sequence era. *Nat Genet, 33 Suppl*, 305-310.

Karlin, S. (2001). Detecting anomalous gene clusters and pathogenicity islands in diverse bacterial genomes. *Trends Microbiol, 9*, 335-343.

Karlin, S. & Burge, C. (1995). Dinucleotide relative abundance extremes: a genomic signature. *Trends Genet, 11*, 283-290.

Karlin, S. & Ladunga, I. (1994). Comparisons of eukaryotic genomic sequences. *Proc Natl Acad Sci U S A, 91,* 12832-12836.

Kirzhner, V., Bolshoy, A., Volkovich, Z., Korol, A. & Nevo, E. (2005). Large-scale genome clustering across life based on a linguistic approach. *Biosystems, 81,* 208-222.

Kirzhner, V., Nevo, E., Korol, A. & Bolshoy, A. (2003). A large-scale comparison of genomic sequences: one promising approach. *Acta Biotheor, 51*, 73-89.

Kirzhner, V. M., Korol, A. B., Bolshoy, A. & Nevo, E. (2002). Compositional spectrum - revealing patterns for genomic sequence characterization and comparison. *Physica A, 312,* 447-457.

Kulkarni, V. G. (1995). *Modeling and Analysis of Stochastic Systems*. New York: Chapman & Hall.

Li, M., Badger, J. H., Chen, X., Kwong, S., Kearney, P. & Zhang, H. (2001). An information-based sequence distance and its application to whole mitochondrial genome phylogeny. *Bioinformatics, 17,* 149-154.

Li, M. & Vitanyi, P. (1997). *An Introduction to Kolmogorov Complexity and Its Applications*. Springer Verlag.

Li, W. (1997). The study of correlation structures of DNA sequences: a critical review. *Comput Chem, 21,* 257-271.

Liao, B. Y., Chang, Y. J., Ho, J. M. & Hwang, M. J. (2004). The UniMarker (UM) method for synteny mapping of large genomes. *Bioinformatics, 20*, 3156-3165.

Lippert, R. A., Huang, H. & Waterman, M. S. (2002). Distributional regimes for the number of k-word matches between two random sequences. *Proc Natl Acad Sci U S A, 99*, 13980-13989.

Lodish, H. F., Berk, A., Matsudaira, P., Kaiser, C. A., Krieger, M., Scott, M. P., Zipursky, L. & Darnell, J. (2004). *Molecular cell biology*. New York: W.H. Freeman and Company.

Mantaci, S., Restivo, A., Rosone, G. & Sciortino, M. (2005). An extension of the Burrows Wheeler Transform and applications to sequence comparison and data compression, *Combinatorial Pattern Matching - Proc. Lecture notes in Computer Science* (pp. 178-189).

Needleman, S. B. & Wunsch, C. D. (1970). A General Method Applicable to the Search for Similarities in the Amino Acid Sequence of Two Proteins. *Journal Molecular Biology, 48*, 443-453.

Nieto, J. J., Torres, A. & Vázquez-Trasande, M. M. (2003). A metric space to study differences between polynucleotides. *Applied Mathematics Letters, 16,* 1289-1294.

Nikolaou, C. & Almirantis, Y. (2005). "Word" preference in the genomic text and genome evolution: different modes of n-tuplet usage in coding and noncoding sequences. *J Mol Evol, 61*, 23-35.

Otu, H. H. & Sayood, K. (2003). A new sequence distance measure for phylogenetic tree construction. *Bioinformatics, 19*, 2122-2130.

Pearson, W. R. (1990). Rapid and sensitive sequence comparison with FASTP and FASTA. *Methods Enzymol, 183*, 63-98.

Pearson, W. R. & Lipman, D. J. (1988). Improved tools for biological sequence comparison. *Proc Natl Acad Sci U S A, 85*, 2444-2448.

Pe'er, I., Felder, C. E., Man, O., Silman, I., Sussman, J. L. & Beckmann, J. S. (2004). Proteomic signatures: amino acid and oligopeptide compositions differentiate among phyla. *Proteins, 54*, 20-40.

Pham, T. D. & Zuegg, J. (2004). A probabilistic measure for alignment-free sequence comparison. *Bioinformatics, 20*, 3455-3461.

Pietrokovski, S. (1994). Comparing nucleotide and protein sequences by linguistic methods. *J Biotechnol, 35*, 257-272.

Pietrokovski, S., Hirshon, J. & Trifonov, E. N. (1990). Linguistic measure of taxonomic and functional relatedness of nucleotide sequences. *J Biomol Struct Dyn, 7*, 1251-1268.

Qi, J., Wang, B. & Hao, B. I. (2004). Whole proteome prokaryote phylogeny without sequence alignment: a K-string composition approach. *J Mol Evol, 58*, 1-11.

Raghava, G. P. & Han, J. H. (2005). Correlation and prediction of gene expression level from amino acid and dipeptide composition of its protein. *BMC Bioinformatics, 6*, 59.

Regeard, C., Maillard, J., Dufraigne, C., Deschavanne, P. & Holliger, C. (2005). Indications for acquisition of reductive dehalogenase genes through horizontal gene transfer by Dehalococcoides ethenogenes strain 195. *Appl Environ Microbiol, 71*, 2955-2961.

Reinert, G., Schbath, S. & Waterman, M. S. (2000). Probabilistic and statistical properties of words: an overview. *J Comput Biol, 7*, 1-46.

Robin, S., Rodolphe, F. & Schbath, S. (2005). *DNA, words, and models*. New York, NY: Cambridge University Press.

Rocha, E. P., Viari, A. & Danchin, A. (1998). Oligonucleotide bias in Bacillus subtilis: general trends and taxonomic comparisons. *Nucleic Acids Res, 26*, 2971-2980.

Sandberg, R., Branden, C. I., Ernberg, I. & Coster, J. (2003). Quantifying the species-specificity in genomic signatures, synonymous codon choice, amino acid usage and G+C content. *Gene, 311*, 35-42.

Sandberg, R., Winberg, G., Branden, C.-I., Kaske, A., Ernberg, I. & Coster, J. (2001). Capturing Whole-Genome Characteristics in Short Sequences Using a Naive Bayesian Classifier. *Genome Res., 11*, 1404-1409.

Smith, T. F. & Waterman, M. S. (1981). Identification of Common Molecular Subsequences. *Journal Molecular Biology, 147*, 195-197.

Snel, B., Huynen, M. A. & Dutilh, B. E. (2005). Genome trees and the nature of genome evolution. *Annu Rev Microbiol, 59*, 191-209.

Sperisen, P. & Pagni, M. (2005). JACOP: a simple and robust method for the automated classification of protein sequences with modular architecture. *BMC Bioinformatics, 6*, 216.

Stuart, G. W. & Berry, M. W. (2004). An SVD-based comparison of nine whole eukaryotic genomes supports a coelomate rather than ecdysozoan lineage. *BMC Bioinformatics, 5*, 204.

Stuart, G. W., Moffett, K. & Baker, S. (2002a). Integrated gene and species phylogenies from unaligned whole genome protein sequences. *Bioinformatics, 18*, 100-108.

Stuart, G. W., Moffett, K. & Leader, J. J. (2002b). A comprehensive vertebrate phylogeny using vector representations of protein sequences from whole genomes. *Mol Biol Evol, 19*, 554-562.

Teeling, H., Waldmann, J., Lombardot, T., Bauer, M. & Glockner, F. O. (2004). TETRA: a web-service and a stand-alone program for the analysis and comparison of tetranucleotide usage patterns in DNA sequences. *BMC Bioinformatics, 5*, 163.

Thompson, J. D., Higgins, D. G. & Gibson, T. J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research, 22*, 4673-4680.

Tino, P. & Dorffner, G. (2001). Predicting the future of discrete sequences from fractal representations of the past. *Machine Learning, 45*, 187-217.

Torney, D. C., Burks, C., Davison, D. & Sirotkin, K. M. (1990). Computation of d2: a measure of sequence dissimilarity. In G. I. Bell & T. G. Marr (Eds.), *Computers and DNA : the proceedings of the Interface between Computation Science and Nucleic Acid Sequencing Workshop, held December 12 to 16, 1988 in Santa Fe, New Mexico* (pp. 109-125). Redwood City, Calif.: Addison-Wesley.

Torres, A. & Nieto, J. J. (2003). The fuzzy polynucleotide space: basic properties. *Bioinformatics, 19*, 587-592.

Troyanskaya, O. G., Arbell, O., Koren, Y., Landau, G. M. & Bolshoy, A. (2002). Sequence complexity profiles of prokaryotic genomic sequences: a fast algorithm for calculating linguistic complexity. *Bioinformatics, 18*, 679-688.

Vinga, S. (2005). *Biological sequence analysis by vector maps: alignment-free comparison of DNA and proteins*. PhD Thesis. Instituto de Tecnologia Química e Biológica - Universidade Nova de Lisboa (ITQB/UNL), Oeiras - Portugal.

Vinga, S. & Almeida, J. (2003). Alignment-free sequence comparison--a review. *Bioinformatics, 19*, 513-523.

Vinga, S. & Almeida, J. S. (2004). Rényi continuous entropy of DNA sequences. *J Theor Biol, 231*, 377-388.

Vinga, S., Gouveia-Oliveira, R. & Almeida, J. S. (2004). Comparative evaluation of word composition distances for the recognition of SCOP relationships. *Bioinformatics, 20*, 206-215.

Wang, Y., Hill, K., Singh, S. & Kari, L. (2005). The spectrum of genomic signatures: from dinucleotides to chaos game representation. *Gene, 346*, 173-185.

Wu, T. J., Hsieh, Y. C. & Li, L. A. (2001). Statistical measures of DNA sequence dissimilarity under Markov chain models of base composition. *Biometrics, 57*, 441-448.

Wu, T. J., Huang, Y. H. & Li, L. A. (2005). Optimal word sizes for dissimilarity measures and estimation of the degree of dissimilarity between DNA sequences. *Bioinformatics, 21*, 4125-4132.

Wu, Z.-B. (2004). Recurrence plot analysis of DNA sequences. *Physics Letters A, 332*, 250-255.

Yang, A. C., Hseu, S. S., Yien, H. W., Goldberger, A. L. & Peng, C. K. (2003a). Linguistic analysis of the human heartbeat using frequency and rank order statistics. *Phys Rev Lett, 90*, 108103.

Yang, A. C., Peng, C. K., Yien, H. W. & Goldberger, A. L. (2003b). Information categorization approach to literary authorship disputes. *Physica A, 329*, 473-483.

Yap, Y. L., Zhang, X. W. & Danchin, A. (2003). Relationship of SARS-CoV to other pathogenic RNA viruses explored by tetranucleotide usage profiling. *BMC Bioinformatics, 4*, 43.

Yu, Z. G., Zhou, L. Q., Anh, V. V., Chu, K. H., Long, S. C. & Deng, J. Q. (2005). Phylogeny of prokaryotes and chloroplasts revealed by a simple composition approach on all protein sequences from complete genomes without sequence alignment. *J Mol Evol, 60*, 538-545.

Zhou, L. Q., Yu, Z. G., Deng, J. Q., Anh, V. & Long, S. C. (2005). A fractal method to distinguish coding and non-coding sequences in a complete genome based on a number sequence representation. *J Theor Biol, 232*, 559-567.

Figure 1

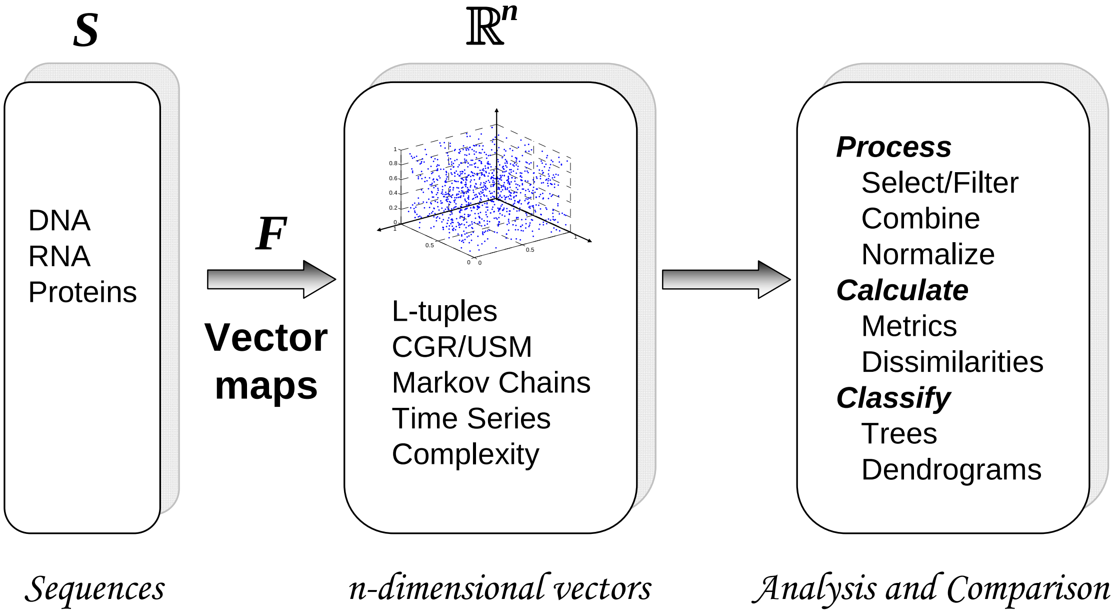


Figure 2a

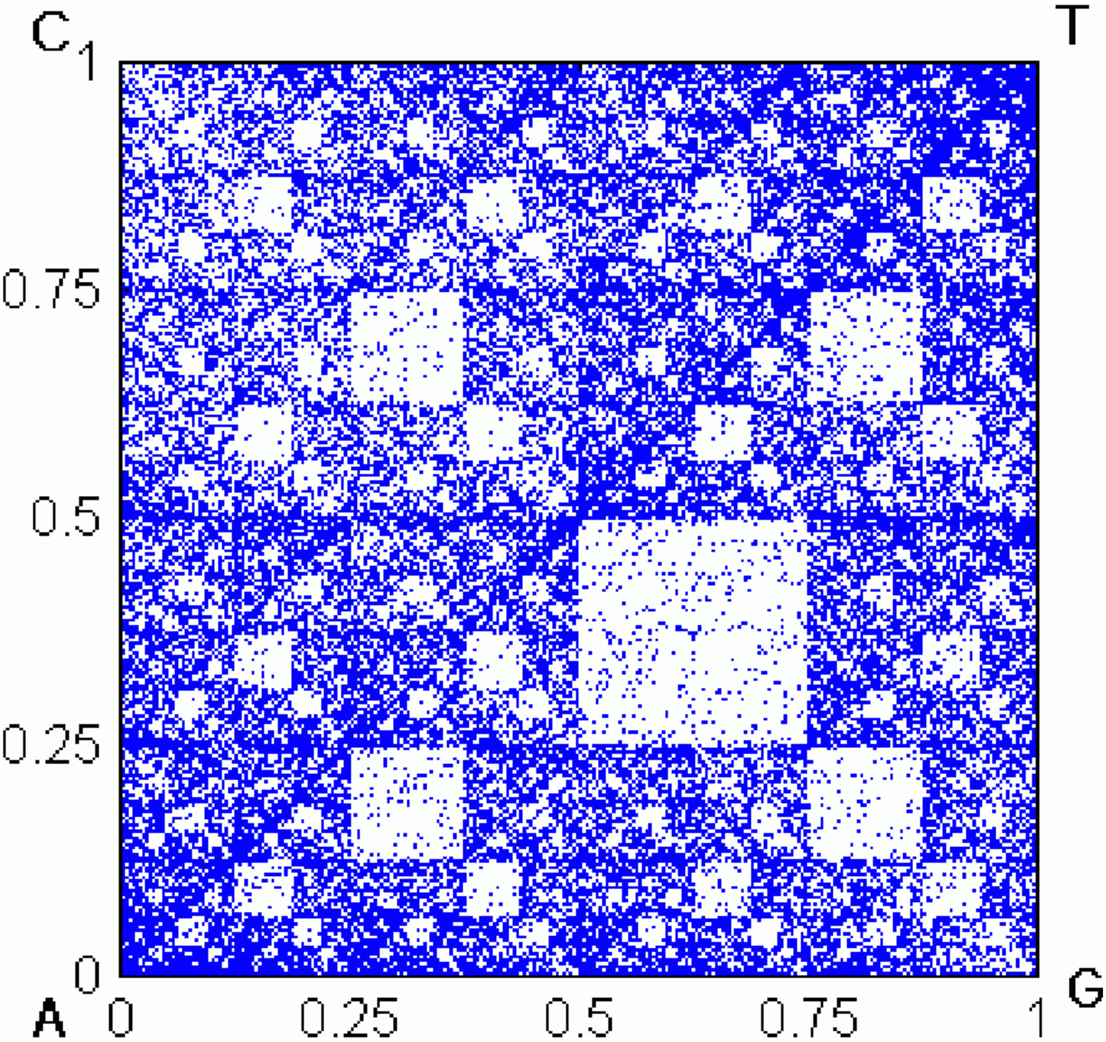
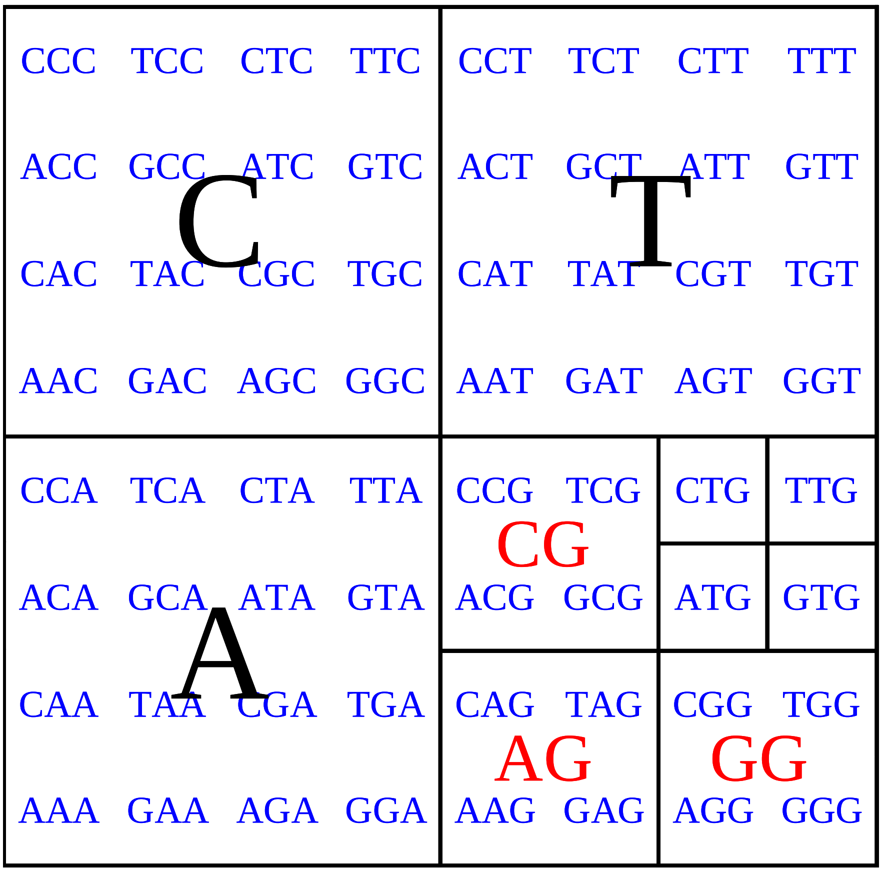


Figure 2b



CCC	TCC	CTC	TTC	CCT	TCT	CTT	TTT
ACC	GCC	ATC	GTC	ACT	GCT	ATT	GTT
CAC	TAC	CGC	TGC	CAT	TAT	CGT	TGT
AAC	GAC	AGC	GGC	AAT	GAT	AGT	GGT
CCA	TCA	CTA	TTA	CCG	TCG	CTG	TTG
ACA	GCA	ATA	GTA	ACG	GCG	ATG	GTG
CAA	TAA	CGA	TGA	CAG	TAG	CGG	TGG
AAA	GAA	AGA	GGA	AAG	GAG	AGG	GGG