Recurrence distributions in computational genomics

Vincenzo Bonnici[a,*], Vincenzo Manca[a,b]

[a]Department of Computer Science, University of Verona, 37134 Verona, Italy
[b]Center for BioMedical Computing (CBMC), University of Verona, 37134 Verona, Italy

Abstract

A genomic distribution, called *recurrence distance* (RDD), is computed for different genomic sequences (E. coli genome, human chromosomes, all exon regions of H. sapiens). A phenomenon associated to RDD, called *3-peak-regularity*, is discovered that is specific of protein-encoding genomic regions. Other second order regularities associated to the peaks of RDD emerge as an evidence. These findings encourage to develop methods of computational genomics, by means of classical concepts of information theory, aimed at identifying genomic dictionaries that underlie to the informative organization of genomes.

Keywords: Computational genomics, Genomic distribution, Recurrence distance, Protein-coding region, Genomic dictionary.

1. Introduction

In recent years, many works were developed that approached the investigation of DNA strings and genomes by means of concepts from information theory and theory of algorithms [1, 2, 3, 4], formal language theory [5, 6, 7], and linguistics ([8, 9, 10, 11, 12]). Information theoretic notions such as: information, codes, entropy, mutual information, entropic divergence, compression, complexity, randomness; and notions from formal language theory and linguistics, such as: grammars, automata, patterns, words, dictionaries, lexical categories, distributional classes, become crucial when genomes are considered as texts and we try to discover deep levels of their informational organization, providing new clues about the logic of "genomic languages" that life developed during evolution. An emerging perspective [13, 14, 2, 15] is based on *alignment free* methods of genome analysis, where global properties of genomes are investigated, rather than local similarities based on classical methods of string alignment.

In this paper we analyse some genomic distributions by showing their relevance in discriminating genomic zones where *k*-mers play biologically meaningful roles. This paper continues previous investigations developed in [16, 17, 4, 18, 19], that were aimed at defining and computing informational genomic indexes, genomic distributions, genomic representations, and genomic dictionaries. In this approach, shortly named *Infogenomics*, around twenty kinds of distributions were investigated on genomes of different organisms, and the computational results were analyzed and compared.

[*] Corresponding author
Email address: `vincenzo.bonnici@univr.it` (Vincenzo Bonnici)

Similar ideas are expressed and applied in recent papers, even with different methods and perspectives [20, 21, 2, 3].

Specific Algorithms and a set of computational tools, called IGTools, were designed for developing the analyses presented in this paper. In particular, here we focus on a specific phenomenon that is a sort of proof of concept of the power of infogenomics, as informational annotation of genomes. In this sense, infogenomics seems to be a perspective somewhat complementary to the ENCODE project (ENCyclopedia Of DNA Elements) focused on the biochemical annotation of human genome [22].

2. Basic Concepts and Notation

Let us recall basic concepts and notation on strings (see also [23, 4]). Strings are sequences of contiguous symbols. Mathematically, a string is a function from as set of *p*ositions, usually the set of positive integers less or equal than a given number n, called length of the string. We denote generic strings with Greek letters (possibly with subscripts) and reserve λ to the empty string (useful for expressing mathematical properties of strings). The length of a string α is denoted by $|\alpha|$, and $\alpha(i)$ or $\alpha[i]$ is the symbol occurring in α at position i, while $\alpha[i, j]$ is the string occurring in G between the positions i and j (both included). Basic operations are assumed on strings. The most important operation over strings is the concatenation. If α and β are two strings, their concatenation, usually denoted by $\alpha\beta$, is the string where the last element of α is followed by the elements of β, in the order they have in β. Let us denote by Γ the genomic alphabet of four symbols (characters, or letters, associated to nucleotides): $\Gamma = \{A, C, G, T\}$. The set Γ^*, as usual, denotes the set of all possible words over Γ. We assume in Γ the order $A < C < G < T$. For them the binary encoding with two bits $A = 00, C = 01, G = 10, T = 11$. On Γ^* the lexicographic order is considered, according to which shorter words are before longer words and between two words α, β of the same length, when j is the first position of them where different letters occur, then α is before β in the lexicographic order, when $\alpha[j]$ is before $\beta[j]$ in the order given over Γ. The binary encoding preserves the relation of ordering and the relation of complementarity, when 0 is complementary to 1, and $0 < 1$.

A genome G is representable by a sequence over Γ. Symbols are written in a linear order, from left to right, according to the standard writing system of western languages, and to the chemical orientation $5' - 3'$ of DNA molecules. Any string α of length k that is a substring of G (a sequence of characters placed in contiguous positions of G) is called a *string, word, factor, k-mer* of G (also k-string, k-word, k-factor, when the length k has to be put in evidence). A dictionary D is a set of words, and a dictionary of G is a set of strings occurring in G. The coverage of D in G is the set of positions of G where a word of D occurs. A dictionary of G is G-complete, when its coverage in G is equal to the set $\{1, 2, \ldots, |G|\}$ of all positions of G. We denote by $D_k(G)$ the dictionary of all k-mers occurring in G. A word of D can occur in G many times. If it occurs m times, then we say that m is its (occurrence) multiplicity in G. A word of G with multiplicity greater than 1 is called a *repeat* of G, while a word with multiplicity equal to 1 is called a *hapax* of G. This term is used in philological investigation of texts, but it is also adopted in document indexing and compression [24]. The values of word multiplicities can be normalized if we divide them by the sum of the multiplicities of all the words occurring in G. This normalization corresponds to replace multiplicity with frequencies, which can be seen as percentages of multiplicities. Two occurrences of α in G are overlapping, when one of them starts at a position of G where some letter of the other occurrence is placed. If this situation does not happen, then the two occurrences are said to be disjoint. A segmentation of a genome G is a sequence of disjoint factors of G such that their concatenation yields G. A recurrence is an ordered pair (p_1, p_2) of two distinct occurrence starting positions of α in G, such that $p_1 < p_2$. If p_2 is the closest occurrence to p_1, than (p_1, p_2) is called a minimal recurrence. If $p_2 - p_1 \geq |\alpha|$ then they identify a non-overlapping recurrence, since the positions covered by the occurrence starting at p_1 do not overlap the occurrence starting at p_2. In the following, recurrence are intended to be minimal and non-overlapping. The notion of recurrence is a particular case of a more general concept of pair occurrence, where p_1 and p_2 are the occurrences of two distinct words α and β. The recurrence distance is defined as $p_2 - p_1$.

The concept of *discrete distribution* (over a domain A), in the context of probability and statistics, refers to a function from A to the set $\mathbb{R}$ of real numbers, having the property of finite summability over A, that is,

$$\sum_{x \in A} f(x) \in \mathbb{R}.$$

Analogously, f is a *continuous distribution* over A if it is is finitely integrable over A:

$$\int_{x \in A} f(x)dx \in \mathbb{R}.$$

A continuous distribution is also called a *density distribution*, or simply a density function. Of course for every density f, given an sequence of reals $(x_i)_{i \geq 0, \in I}$, it univocally provides a discrete distributions g such that $g(x_i) = \int_{x_{i-1} \leq y < x_i} f(y)dy$. When the sum or the integral over A is equal to one, then the distribution (discrete or continuous) is a *probability distribution*. We remark that the term *distribution function* is mostly used for the cumulative distributions [25], which in the case f is a discrete distribution is given by the function $F(x) = \sum_{y \leq x} f(y)$ ($F(x) = \int_{y \leq x} f(y)dy$ in the continuous case). However, in the case of genomic analyses, it is useful to extend the term distribution by including also *partition distributions*, that is, a function f from a set A to $\mathbb{P}(S)$, the subsets of a finite set S, such that $\bigcup f(x)_{x \in A} = S$. Here the finite summability is replaced by the finite cardinality of the set "distributed" among the element of the domain of f. This kind of distribution will be called also the *spectral* distribution, and $f(x)$ is called the *spectrum* of x in S.

We use the word distribution, without further specification, when it is clear from the context the specific meaning of its usage. We refer to [25, 26, 27] for basic concepts about probability and information theory. A *Bernoullian genome* is a synthetic genome generated by means casual (blind) extractions (with insertion after extraction) from urns containing four types of balls, completely identical apart their colors, denoted by the genomic letters A, C, G, T. These random genomes are very useful in the analysis of real genomes.

3. Some genomic distributions

Information theory began with Shannon's famous booklet published in 1948 [28]. This epochal work was essentially based on the idea that the information measure of an event is a function of its probability (the logarithm of its inverse value). But probability is given by distributions (on suitable spaces of events), therefore the essence of an *information source* (according to Shannon's terminology) has to be searched in suitable distributions.

In the case that information sources are genomes, biological meaningful information has to be extracted from distributions defined on genomes. In the following, we consider a list of interesting distributions that naturally arise when we analyze a genome G [4].

Among other mentioned distributions, here we will focus on the distribution (10) in the following list, which was also considered in [29, 30, 31, 32, 33], starting from an attempt at adapting concepts from the energy spectra of physical systems, in order to extract key words from literary texts. As it will appear in the next sections, distribution (10) is here considered with different methods and perspectives, by showing its strong relevance in discriminating biological functionalities. Below, we describe a set of genomic distributions, moreover, formal definitions are given in table 3.

1. **Word position**
   This distribution assigns to any word α of D the set of positions of G where it (its first character) occurs, that constitute what is also called the *spectrum* of α in G. When the spectrum is given for every word of a dictionary that is complete for G, then the whole genome G can be easily reconstructed.
2. **n-Word count**
   This distribution assigns to any value of $n < |G|$ the cardinality of $D_n(G)$.
3. **n-Repeat count**
   This distribution assigns to any value of n, from 0 to some maximum value, the cardinality of the set of the repeats of G having length n.
4. **Rank frequency**
   If we order the words, of a complete dictionary for G, according to their frequencies in G (in decreasing order) we say that the most frequent words have rank 1, the most frequent words, after words of rank 1, have rank 2, and so on. Therefore, this distribution assigns to each rank the value corresponding its frequency. This distribution, also called *Zipf distribution*, after the scholar who introduced it, was extensively studied in natural languages.
5. **Word Multiplicity**
   This distribution assigns to each word of D the number of times it occurs in G.
6. **Multiplicity class**
   This distribution assigns to any value of n, from 0 to some maximum value, the set of words that occur in G with multiplicity n.
7. **Word Co-multiplicity**
   This distribution assigns to each value of n, from 0 to some maximum value, the number of words of D that occur in G with multiplicity n, which we call *co-multiplicity* (it coincides with the cardinality of the set provided by the previous distribution).

8. **Word distance**
   Given a repeat word α, this distribution assigns to any n, going from 2 to some maximum value, the distance between the $(n-1)$-th and n-th occurrences of α in G.
9. **Segment-multiplicity**
   Fixed an "unitary" segment length s, such that G is disjointly factorized in $\lceil |G|/s \rceil$ factors, and a word α, this distribution assigns to any multiplicity n, going from 0 to some maximum value, the number of segments where α occurs exactly n times. In a Bernoullian genome, this distribution, when normalized (with respect to the total number of segments), is a Poisson probability distribution, for some value of Poisson parameter [34] (we verified this fact in several computational experiments). Moreover, the *waiting time* of an α-occurrence process following a Poisson law, that is, the distance between two consecutive occurrences of α is an exponential distribution (with the same parameter as in the Poisson law).
10. **Recurrence distance**
   Given a word α, this distribution assigns to any distance n, going from 1 to some maximum value, the number of times it occurs at distance n from its previous occurrence. As we remarked, in presenting the previous distribution, In a Bernoullian genome, recurrence distance distribution, when normalized, corresponds to the waiting time associated to a Poisson process, that is, to an exponential distribution. Given a set of words D, the **average recurrence distance** assigns to any distance n the value

$$\frac{\sum_{\alpha \in D} |R(G, \alpha, n)|}{|D|}$$

   where $|R(G, \alpha, n)|$ is the number of times α occurs in G at distance n ($|D|$ is the cardinality of D).

Table 3 yields the formal definition of distributions described above, where the following definitions are assumed:

$$\alpha \subset \beta \iff \exists ij(\alpha = \beta[i, j])$$

$$D(G) = \{\alpha \in \Gamma^* \mid \alpha \subset G\}$$

$$\alpha \not\subset \beta \iff \neg\exists ij(\alpha = \beta[i, j])$$

Recurrence is a peculiar phenomenon of words. In fact a word is a unity that may occur many times in a text, with a specific identity that is independent from its instances, and at same time, its instances in the text (and in the specific contexts where they occur) are related to some function that it plays within the text.

In next section 4, by using the distance recurrence distribution, we will show in genome of Escherichia coli a phenomenon, which we call *3-peak-regularity*, missing in the whole genome of Homo sapiens. In section 5 we show that 3-peak-regularity appears again when the same distribution is computed only for exon regions in Homo sapiens. An explanation to this observed specificity of 3-peak-regularity is outlined at end of section 5. The peak-analysis of recurrence distance distribution seems to suggest

WordPosition $\alpha \mapsto \{0 < i \leq |G| \mid G[i] = \alpha\}$

WordCount $n \mapsto |D_n(G)|$

RepeatCount $n \mapsto |\{\alpha \subset_i G \mid |\alpha| = n, i > 1\}|$

RankFrequency $n \mapsto |\{\alpha \in D \mid \alpha \subset_n G\}|/|D|$

WordMultiplicity $\alpha \mapsto i : \alpha \subset_i G$

MultiplicityClass $n \mapsto |\{\alpha \in \Gamma^* \mid \alpha \subset_n G\}$

WordComultiplicity $n \mapsto |\{\alpha \in \Gamma^* \mid \alpha \subset_n G\}$

WordDistance $n \mapsto d : \exists d_1, d_2 : \alpha \subset_n G[1, d_1], \alpha \subset_{n+1} G[1, d_2], G[d_1 + 1, d_2] = \beta\alpha, \alpha \not\subset \beta, |\beta\alpha| = d$

SegmentMultiplicity $n \mapsto |\{i \mid \alpha \subset_n G[is, (i+1)s]\}$

RecurrenceDistance $n \mapsto |\{\alpha\beta\alpha \subset G \mid |\alpha\beta| = n, \alpha \not\subset \beta\}|$

Table 1: Formal definitions of ten genomic distributions.

a new method in discriminating genomic regions with different functions. Moreover, the analysis of recurrence distance distributions, could open new perspectives in computational genome analyses, toward the identification of genomic dictionaries of biological relevance, on the basis of which a first level of comprehension in deciphering genome languages could be obtained.

4. Recurrence Distances in Escherichia coli and Homo sapiens

In this section we present the main phenomenon related to distance recurrence, which drew our interest toward recurrence distance distribution. In fact, when we compute this distribution, for a given 3-mer α in E. coli, a clear “ternary phase” of “extra-peaks” appears (see figure 1). Namely, at interval 3, the number of times that a recurrence distance d of α appears is greater (say more than 20%) than the numbers of times recurrence distances $d-2, d-1, d+1, d+2$, placed on an underlying basic curve. Figure 1 shows the recurrence distance distribution of word ATG in Escherichia coli and in Homo sapiens whole genome.

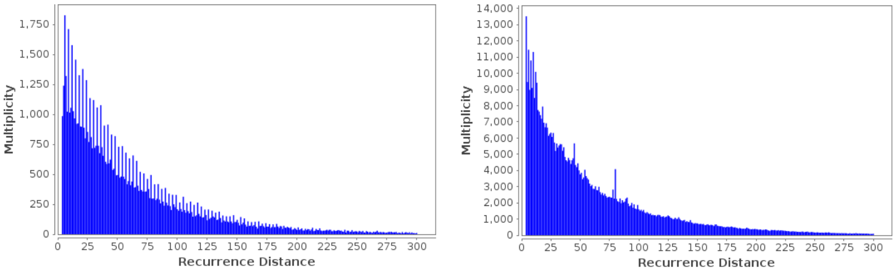

Figure 1: Recurrence distance distribution of the word ATG in Escherichia coli (left side) and Homo sapiens (right side). Distributions are plotted up to distance 300. In Escherichia coli distribution extra-peaks are on recurrence distances that are multiple of 3. In Homo extra-peaks do not have such a regularity and are often sparse.

As Figure 1 shows, in Escherichia coli extra-peaks regularly appears for all distances that are multiples of 3 (up to some maximum value), and they seem to follow a second parallel exponential distribution respect to the basic one. On the contrary, in Homo sapiens, extra-peaks are not localized at a specific recurrence distance, they are smeared and sparse.

The word ATG is not just a special case, in fact all 3-mers show similar distributions. The average recurrence distance confirms the generality of the phenomenon. Figure 2 shows the average recurrence distance multiplicity distributions for several values of k, for both genomes of E. coli and H. sapiens.

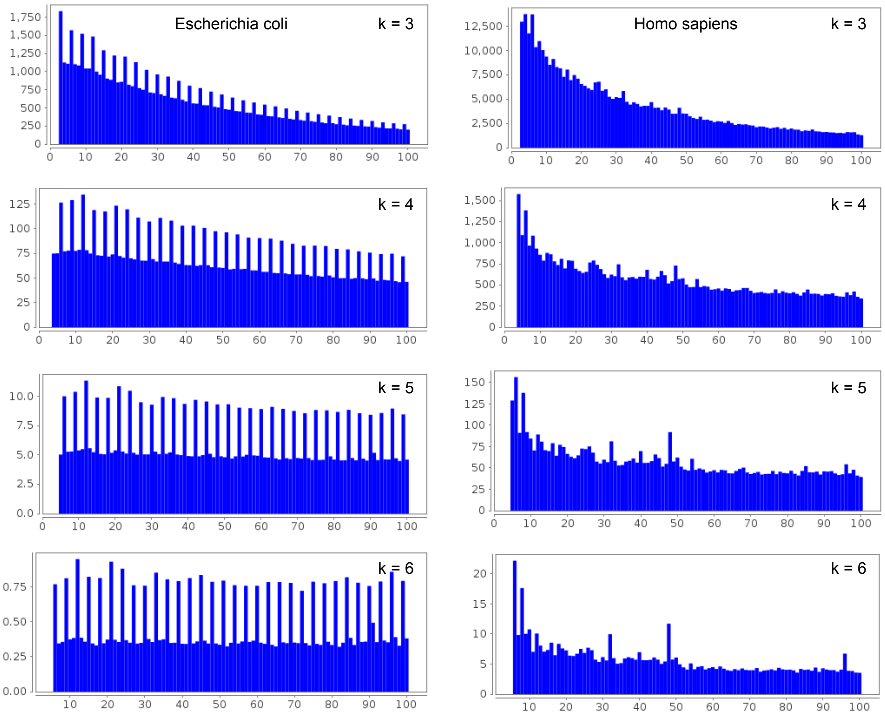

Figure 2: Average recurrence distance distribution in Escherichia coli (left side) and Homo sapiens (right side), for word lengths k equal to 3, 4, 5, and 6. Distributions are plotted up to distance 100. In Escherichia coli regular extra peaks are located at recurrence distances that are multiples of 3.

In conclusion, the 3-peak-regularity of E. coli is evident and systematic, and, at the same time, this regularity signal is almost absent, or weak, in the Human genome. In the next section we apply recurrence distance to other cases, in order to collect results that could help us in answering to the question: "What is the reason of this difference?"

5. Recurrence distance distribution in human exon regions

The Escherichia coli genome is relatively dense of genes. In fact, 40% of its length (in both strands) is covered by exons. We suspected that this great density of encoding parts, with respect to Homo sapiens genome, is the main reason of the appearance of 3-peak-regularity in the former case, and of its disappearance in the latter one. In order to validate this supposition we extracted all exon regions of H. sapiens and computed the recurrence distance distribution for their concatenation. Figure 3 confirms the claim that 3-peak-regularity is related to the encoding function.

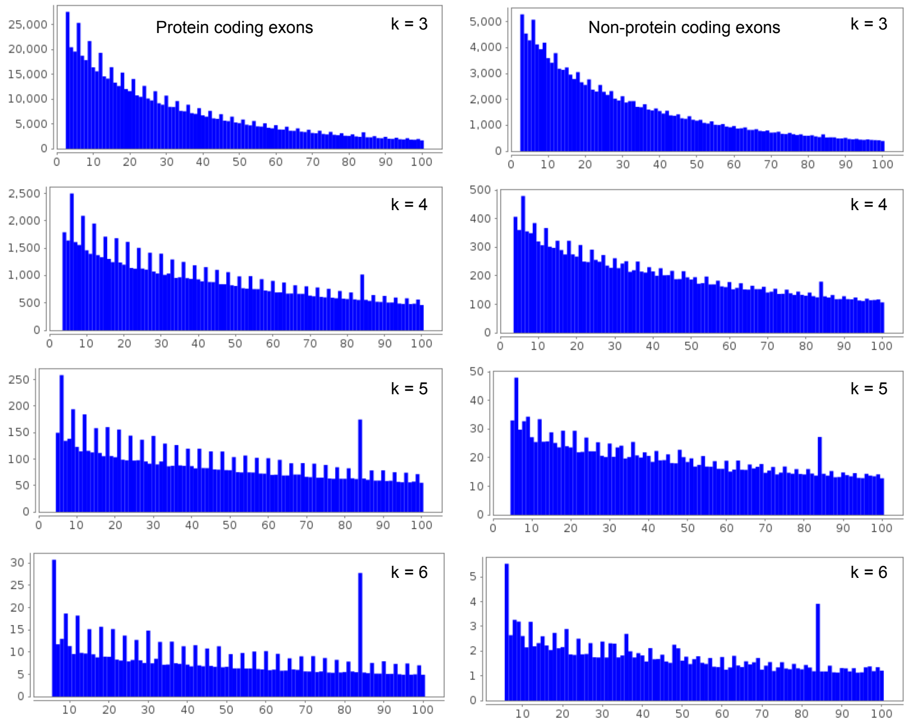

Figure 3: Average recurrence distance distributions inside protein coding exons (left) and non-protein coding exons (right) of Homo sapiens, for word lengths *k* equal to 3, 4, 5, and 6. Distributions are plotted up to distance 100. The distributions corresponding to protein coding regions show a clear 3-peak-regularity signal, which fades for non protein coding regions.

If we consider the recurrence distance distribution of *AGA* in exon and non-exon regions of Escerichia coli, then a difference appears in the two cases, as it is shown in Figure 4.

However, in E. coli this discrimination is less evident when we consider the average distributions in exon and non-exon regions (see supplementary materials). This fact could be explained with two, possibly concurrent, reasons: i) the high density of encoding regions, and ii) the fact that non-encoding regions are potential encoding regions in a silent state, as required by the high dynamic and adaptable nature of its bacterial genome.

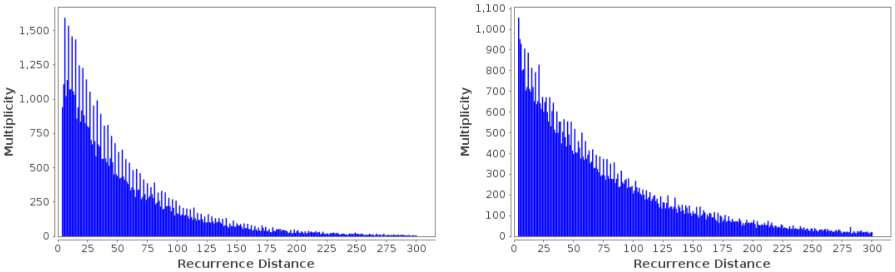

Figure 4: Recurrence distance distribution of the word *AGA* inside (left) and outside (right) the protein coding regions of Escherichia coli. Distributions are potted up to distance 300. The distribution corresponding to protein coding regions shows a more evident and regular extra peaks behavior.

A more accurate analysis on human transcripts reveals that 3-peak-regularity is not present in introns. Besides, we have split exons in two categories, those coming from protein coding genes and those coming from non-protein coding gene. In protein coding exons, the phenomenon is well observable, while it is less evident in non-protein coding exons.

We propose the following explanation to the 3-peak-regularity phenomenon. When we are in a coding region, 3-mers are codons. This means that when a 3-mer α occurs, then a specific sequence of 3-mers follows it, and of course, its recurrence is meaningful after a number of steps that is multiple of three. In conclusion, meaning appears at a 3-step reading intervals. On the contrary, in regions where no specific meaning is related to the occurrence of α, or where a biological semantic is not codon dependent, then the distance recurrence distribution of 3-mers can be approximated to the waiting time of a Poisson process, which is typically an exponential distribution. For this reason, peaks at 3-interval emerge in coding regions where distance recurrences at inter-distance 3 are not driven by change, but from specific reasons due to the encoding function that 3-mers are playing in that context. In other words, peaks emerge where the recurrence of a 3-mer after $3n$ steps is not by chance, but for a specific reason. In this regard, it seems very appropriate to recall Poincare's viewpoint about chance [35]. It is not absence of causes, but the effect of great number of causes that cannot singularly identified, because they act at comparable levels of importance. When one of them emerges among the others, the chance disappear by definition.to par

We want also to stress the fact that 3-peak-regularity is maintained even when we consider words of length greater than 3. However, what changes in passing from 3, to 4, or 5, or 6, and so on, is the magnitude of the peaks, which decrease as the word lengths increase. This is due to the fact that a 4-mer α is a 3-mer β followed by a 1-mer x, therefore the probability of occurrence $P(\alpha)$ is given by ($P(E_2|E_1)$, the conditional probability of event E_2 when event E_1 already occurred):

$$P(\alpha) = P(\beta) \cdot P(x|\beta)$$

this means that the argument developed in the case of 3-mers continues to apply, but with a multiplicative factor such as $P(x|\beta)$ that reduces the occurrence probabilities (in an analogous way we can reason for $k > 4$).

6. Searching for patterns in correspondence of peaks

In the previous sections we argued that peaks are related to the emergence of some cause, among the indiscernible population of causes acting over a random phenomenon. Now we want to distinguish regular peaks, from the irregular ones. The regular peaks, as those observed in recurrence distance of E. coli, and in recurrence distance of exons in H. sapiens, follow a sort of second order distribution with a shape that is parallel to the exponential shape of the basic distribution over which peaks emerge. On the contrary, irregular peaks, as those observed in the whole human genome, are sparse or isolated, therefore their presence is not a part of a collective phenomenon, but something very specific and localized (in the set of distances). However, if regular peaks follow a pattern, peaks that do not behave in this way, apparently without any pattern, include patterns in the words recurring with the recurrence distance corresponding to the peak. This phenomenon is shown in Figure 5, where words are represented by colored rows, and are lexicographically ordered from the top to the bottom. In each row the color encoding maps letters A, C, G, and T to a rectangle of 4×1 pixels of color blue, red, green, and yellow, respectively.

The peak-inside analysis can be extended by using a more general notion of recurrence. In fact, if we search for the number of times two words α and β occurs at some distance d, we have a *pair recurrence distance*. Also in this distribution peaks very often individuate words enclosed between the pair (α, β) at a distance d (the value yielding the peak) following very specific patterns.

7. Data sources and software

Here, we give some details about data and procedures used in the paper. All the genomic sequences were obtained from the UCSC (Univeristy of California, Santa Cruz) on-line public database [36]. For Homo sapiens we used the assembly GRCh37 (Genome Reference Consortium Human Build 37), also called hg19, assembled by the Genome Reference Consortium. For Escherichia coli sequence, we used the version numbered as 536, the most recent genome sequence draft released in 2006 by the University of Goettingen. Genomic coordinates of transcripts and their exons, for both the corresponding assemblies of Homo sapiens and Escherichia coli, were retrieved from the UCSC Table Browser [37] based on RefSeq database [38] (to date October 2014).

Genes may reside on both strands of DNA sequences, and they may also produce more than one transcript. Transcripts (and exons) data are provided with absolute coordinates, where 0 corresponds to the first nucleotide of the $5' - 3'$ strand. When the transcript resides on the $3' - 5'$ strand, then the reverse complement has to be extracted. Moreover, we applied a special procedure to remove redundancy due to transcripts that overlap on the same strand. In fact, we marked the nucleotides covered by each transcript, then we extracted only contiguous regions of marked nucleotides. Thus, when

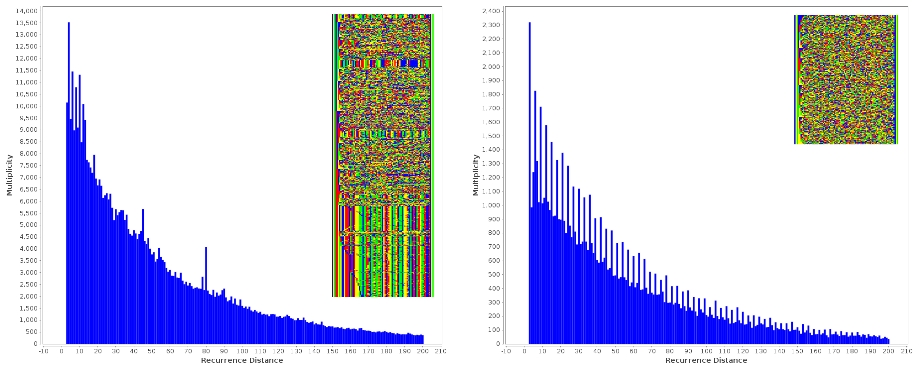

Figure 5: On the left side, recurrence distance distribution (up to distance 200) of the words ATG from the chromosome 22 of Homo sapiens. Beside the distribution curve, there are $4,081$ colored rows, in lexicographic order, of the 84-strings enclosed between the (minimal) recurrences of ATG at distance 81. On the right side, recurrence distance distribution of the word ATG in Escherichia coli. Beside the curve, there are 416 strings $ATG - 78N - ATG$ denoted by colored rows in lexicographic order, that is, the strings enclosed between the recurrences of ATG at distance 81. The enclosed strings of E. coli show low similarity, Contrarily, a great part of the enclosed strings in Homo sapiens are similar. Moreover, approximately, the percentage of strings showing similarity corresponds to the portion of the peak emerging from the basic curve.

two (or more) transcripts overlap, the resultant extracted region starts from the first nucleotide of the earliest (in terms of coordinate order) transcript and ends to the last nucleotide of the latest transcript.

The computation of distributions presented in the paper were performed by an inhouse software called IGTools, which aims at providing a comprehensive platform for informational analysis of genomic and biological sequences. It is based on three main concepts: efficiency, usability, and modularity. IGTools supplies genome representations with their corresponding suffix arrays [39, 40], which are well-established data structures, here adapted to genomic sequences. Graphical user interfaces (GUI) and command-line interfaces (CLI) are available in IGTools. CLIs are suitable for batch and/or extensive computations, while GUIs provide an interactive interface for investigative analyses.

Extraction and visualization of recurrence distance distributions can be performed by means of IGTools with low computational costs. The user can navigate in realtime (after fixing some specific parameters, such as the sequence and words lengths) through the distributions of the words in a given dictionary, and at the same time, can select peaks and extract the regions enclosed between the recurrent word at a given distance.

A collection of API for developers is provided, which is easy to use and to extend for custom analyses. APIs provide access to all the abstraction levels of the framework (see figure 6). The lowest level implements the core data structures and algorithms (extended suffix arrays [41], and succinct nucleotide sequence representation). The middle level implements typical operations for dictionary construction and for the realization of basic elaborations over dictionaries (words enumeration, localization, multiplicity count, elongation, as well as set-theoretic operations among dictionaries). Finally, the highest level implements the informational analyses defined in InfoGenomics, such as the computation of genomic distributions, their analysis and manipulation (approximation and parameter estimation of the closest mathematical distributions). IGTools is an open-source software entirely developed in Java, in order to be platform-independent. It is available upon request by email to the first author and it will be soon released with free license (for non-commercial use).

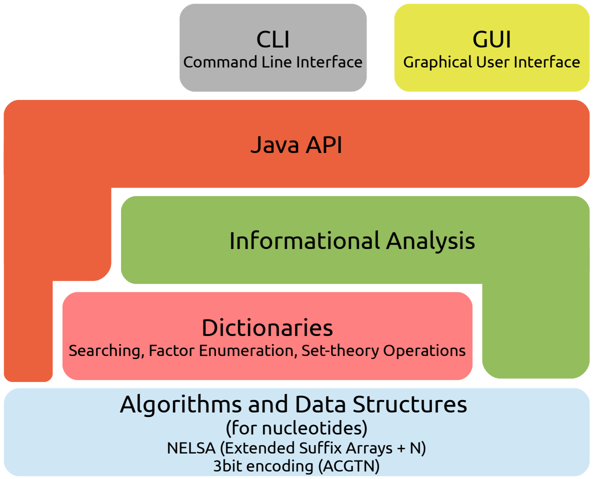

Figure 6: Abstraction layers and functionalities of the IGTools platform.

8. Conclusion and open problems

In this paper, in the context of computational methods based on genomic distributions, we focused on the recurrence distance distribution, which we analysed in different genomes and portions of genomes, by discovering a phenomenon, called 3-peak regularity, which occurs when recurrence distance distribution is applied to regions in most part, or exclusively, devoted to encoding function. This result is a proof of concept of the capability of distributional analyses in revealing biological functions of genome components. A natural development, currently under investigation, is the definition of algorithms able to localize regions where 3-peak regularity pattern occurs, in order to gain genome informational annotations useful to recognize biological functions.

More specific questions concern with some other kinds of peak regularity that we noticed in both types of exon regions, where a second level of regular extra peaks appear at distances that are multiples of 84, as it begins to be apparent in Figure 3, and it is confirmed in Figure 7.

Another issue, directly related to the recurrent distance distribution, is the identification of genome fragments that can be considered as "genomic words", that is, small recurrent units entering the composition of genomic texts with specific biological meanings. It was already discovered, in connection with literary texts, that algorithms can be defined, based on recurrence distance, that are able to recognize key-words of

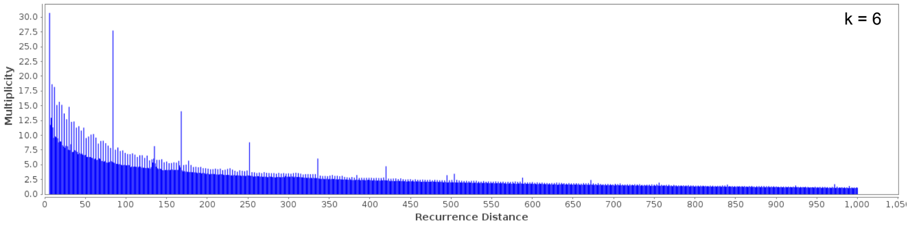

Figure 7: Average recurrence distance distributions for $k = 6$, inside protein coding exons of Homo sapiens. Distributions are plotted up to distance 1000. The distribution shows supplementary extra peaks on distances multiples of 84.

a text [29]. We implemented the same kinds of algorithms for genomes, but several pitfalls of this method, when adapted to genomes, generate results that are not reliable. We combined recurrence distance distributions with other informational notions (random genomes, entropic divergence, parameter evaluation, word elongation). In this way, some genomic words clearly emerge, as fragments with lengths mostly between 6 and 18, which present an informational "density" significantly greater than typical fragments with the same lengths. The search for genomic dictionaries generated by means of these analyses are in progress, and of course, biological validations and interpretations of the obtained results are necessary. However, we are convinced that this kind of investigations could provide powerful insights for a deep understanding of internal organization of genomes.

References

[1] R. C. Deonier, S. Tavaré, M. Waterman, Computational genome analysis: an introduction, Springer, 2005.

[2] S. Vinga, J. Almeida, Alignment-free sequence comparison—a review, Bioinformatics 19 (4) (2003) 513–523.

[3] S. Vinga, Information theory applications for biological sequence analysis, Briefings in bioinformatics 15 (3) (2013) 376–389.

[4] V. Manca, Infobiotics: information in biotic systems, Springer, 2013.

[5] T. Head, Formal language theory and DNA: an analysis of the generative capacity of specific recombinant behaviors, Bulletin of mathematical biology 49 (6) (1987) 737–759.

[6] V. Manca, G. Franco, Computing by polymerase chain reaction, Mathematical biosciences 211 (2) (2008) 282–298.

[7] G. Franco, V. Manca, Algorithmic Applications of XPCR, Natural Computing 10 (2011) 805–819.

[8] V. Brendel, H. Busse, Genome structure described by formal languages, Nucleic acids research 12 (5) (1984) 2561–2568.

[9] D. B. Searls, The language of genes, Nature 420 (6912) (2002) 211–217.

[10] D. B. Searls, Molecules, Languages and Automata, in: Grammatical Inference: Theoretical Results and Applications, Springer, 2010, pp. 5–10.

[11] A. Puglisi, A. Baronchelli, V. Loreto, Cultural route to the emergence of linguistic categories, Proceedings of the National Academy of Sciences 105 (23) (2008) 7936–7940.

[12] M. Gimona, Protein Linguistics—a grammar for modular protein assembly?, Nature Reviews Molecular Cell Biology 7 (1) (2006) 68–73.

[13] B. Hao, J. Qi, Prokaryote phylogeny without sequence alignment: from avoidance signature to composition distance, Journal of Bioinformatics and Computational Biology 2 (01) (2004) 1–19.

[14] G. Hampikian, T. Andersen, Absent sequences: nullomers and primes., in: Pacific Symposium on Biocomputing, Vol. 12, 2007, pp. 355–366.

[15] C. Yin, Y. Chen, S. S.-T. Yau, A measure of DNA sequence similarity by fourier transform with applications on hierarchical clustering, Journal of Theoretical Biology (2014).

[16] A. Castellini, G. Franco, V. Manca, A dictionary based informational genome analysis, BMC Genomics 13 (1) (2012) 485.

[17] G. Franco, Discrete and topological models in molecular biology, in: N. Jonoska, M. Salto (Eds.), Perspectives in computational genome analysis, Springer-Verlag Berlin Heidelberg, 2014, Ch. 1, pp. 3–22.

[18] G. Franco, A. Milanese, An investigation on genomic repeats, in: The Nature of Computation. Logic, Algorithms, Applications, Springer, 2013, pp. 149–160.

[19] A. Castellini, G. Franco, A. Milanese, A genome analysis based on repeat sharing gene networks, Natural Computing (2014) 1–18.

[20] B. Chor, D. Horn, N. Goldman, Y. Levy, T. Massingham, Genomic DNA k-mer spectra: models and modalities, Genome Biology 10 (10) (2009) R108.

[21] Y. Fofanov, Y. Luo, C. Katili, J. Wang, Y. Belosludtsev, T. Powdrill, C. Belapurkar, V. Fofanov, T.-B. Li, S. Chumakov, et al., How independent are the appearances of n-mers in different genomes?, Bioinformatics 20 (15) (2004) 2421–2428.

[22] The ENCODE Project Consortium, An integrated encyclopedia of DNA elements in the human genome, Nature 489 (7414) (2012) 57–72.

[23] G. Rozenberg, A. Salomaa, Handbook of Formal Languages: Beyonds words, Vol. 3, Springer, 1997.

[24] I. H. Witten, A. Moffat, T. C. Bell, Managing gigabytes: compressing and indexing documents and images, Morgan Kaufmann, 1999.

[25] W. Feller, An introduction to probability theory and its applications, Vol. 1, John Wiley & Sons, 1968.

[26] S. Kullback, R. A. Leibler, On information and sufficiency, The Annals of Mathematical Statistics (1951) 79–86.

[27] T. M. Cover, J. A. Thomas, Elements of Information Theory, John Wiley & Sons, 1991.

[28] C. E. Shannon, A Mathematical Theory of Communication, The Bell System Technical Journal 27 (1948) 379–423,623–656.

[29] P. Carpena, P. Bernaola-Galván, M. Hackenberg, A. Coronado, J. Oliver, Level statistics of words: Finding keywords in literary texts and symbolic sequences, Physical Review E 79 (3) (2009) 035102.

[30] V. Afreixo, C. A. Bastos, A. J. Pinho, S. P. Garcia, P. J. Ferreira, Genome analysis with inter-nucleotide distances, Bioinformatics 25 (23) (2009) 3064–3070.

[31] M. Hackenberg, A. Rueda, P. Carpena, P. Bernaola-Galván, G. Barturen, J. L. Oliver, Clustering of DNA words and biological function: A proof of principle, Journal of theoretical biology 297 (2012) 127–136.

[32] C. A. Bastos, V. Afreixo, A. J. Pinho, S. P. Garcia, J. Rodrigues, P. J. Ferreira, Inter-dinucleotide distances in the human genome: an analysis of the whole-genome and protein-coding distributions, Journal of Integrative Bioinformatics 8 (3) (2011) 172.

[33] C. Carretero-Campos, P. Bernaola-Galván, A. Coronado, P. Carpena, Improving statistical keyword detection in short texts: Entropic and clustering approaches, Physica A: Statistical Mechanics and its Applications 392 (6) (2013) 1481–1492.

[34] J. K. Percus, Mathematics of Genome Analysis, Vol. 17, Cambridge University Press, 2002.

[35] J. Poincaré, La science et l'hypothèse, Flammarion, Paris, 1968.

[36] D. Karolchik, G. P. Barber, J. Casper, H. Clawson, M. S. Cline, M. Diekhans, T. R. Dreszer, P. A. Fujita, L. Guruvadoo, M. Haeussler, et al., The UCSC genome browser database: 2014 update, Nucleic acids research 42 (D1) (2014) D764–D770.

[37] D. Karolchik, A. S. Hinrichs, T. S. Furey, K. M. Roskin, C. W. Sugnet, D. Haussler, W. J. Kent, The UCSC Table Browser data retrieval tool, Nucleic acids research 32 (suppl 1) (2004) D493–D496.

[38] K. D. Pruitt, T. Tatusova, D. R. Maglott, NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins, Nucleic acids research 35 (suppl 1) (2007) D61–D65.

[39] U. Manber, G. Myers, Suffix arrays: a new method for on-line string searches, SIAM Journal on Computing 22 (5) (1993) 935–948.

[40] S. Kurtz, A. Narechania, J. C. Stein, D. Ware, A new method to compute k-mer frequencies and its application to annotate large repetitive plant genomes, BMC genomics 9 (1) (2008) 517.

[41] M. I. Abouelhoda, S. Kurtz, E. Ohlebusch, Replacing suffix trees with enhanced suffix arrays, Journal of Discrete Algorithms 2 (1) (2004) 53–86.