# A Word Recurrence Based Algorithm to Extract Genomic Dictionaries

Vincenzo Bonnici
*Department of Computer Science*
*University of Verona*
Verona, Italy
email: vincenzo.bonnici@univr.it

Giuditta Franco
*Deptartment of Computer Science*
*University of Verona*
Verona, Italy
email: giuditta.franco@univr.it

Vincenzo Manca
*Deptartment of Computer Science*
*University of Verona*
Verona, Italy
email: vincenzo.manca@univr.it

***Abstract*—Genomes may be analyzed from an information viewpoint as very long strings, containing functional elements of variable length, which have been assembled by evolution. In this work, an innovative information theory based algorithm is proposed, to extract significant (relatively small) dictionaries of genomic words. Namely, conceptual analyses are here combined with empirical studies, to open up a methodology for the extraction of variable length dictionaries from genomic sequences, based on the information content of some factors. Its application to human chromosomes highlights an original inter-chromosomal similarity in terms of factor distributions.**

***Keywords*—Genome languages, information content, Kullback-Leibler, word extraction.**

## I. Introduction

Human genome computational analysis is one of the most important and intriguing research challenges we are currently facing. Genomes carry the main information underlying life of organisms and their evolution, including a system of molecular rules which orchestrate all cell functions [1]. Our work here follows and outlines some trends of research which analyze and interpret genomic information, by assuming the genome to be a book encrypted in a language to decipher [2–7], in order to convert the genomic information into a comprehensible mathematical form, such as a dictionary of variable-length factors that collects words of the unknown genomic language.

According to a common approach in computational genomics [8–12], a genome is represented by a string over the nucleotidic alphabet. This representation easily leads to affinities with a text, written in a natural language, which is comprehensible by means of its vocabulary, giving both syntax and semantic of *words*.

Several studies define properties for words which result to be salient features in analysing genomic sequences [13]. Minimal absent words, maximal or palindromic repeated words are some examples [14–16]. These approaches are focused on finding specific words to be used as key features of a string for analysing its property or for comparing it to another sequence [17]. The extracted words are often sparsely located in the analysed sequence [18], thus they do not constitute a real linguistic analysis of genomic strings.

According to recent advancements, the concept of *functional element* is central, defined as a genomic segment that codes for a defined biochemical product or displays a reproducible biochemical signature [6, 19]. An information theory based analysis clearly plays an important role in deciphering such elements as the genomic language [20], and it allows us to confirm the linkage between DNA fragments and their information content [4, 8, 19, 21–23].

In [24, 25], the authors applied a methodology developed for literary text to extract fixed length genomic dictionaries. Examples of fixed length dictionary extraction procedures could be provided by applying notions such as word multiplicity or word length distributions. On the other hand, graphical investigative analyses, based on expected frequency gaps, show the unpredictable behaviour of genomic sequences and help to detect peculiar words [26].

If we think of a book, semantically significant words have a fairly medium number of occurrences and they are clustered according to the topic described in specific part of the book. Several works are focused on finding genomic words exhibiting some special kind of (somehow clustered) repetitiveness, with a global frequency quite different than the expected frequency in purely random sequences having the same length of an investigated genome [8, 21, 22, 27–29]. A very relevant and peculiar word periodicity is revealed by the *Recurrence Distance Distribution* (RDD), which measures the frequency at which a given word occurs at given distances [30]. Its application to coding regions shows the informational evidence of the codon language, and in [31–33] some characterizations of recurrence behaviours were pointed out for very short $k$-mers. However, only fixed length dictionaries were extracted from real genomes by means of such a distribution [25].

In this paper, we start from a modified version of an algorithm introduced in [24], in order to apply it to real genomes. We call it V-algorithm, from the first name of the authors who designed it. Both these original and modified algorithms are aimed at finding words forming local clusters (the approach is explained in Section II-A). Then, we propose a new RDD-based algorithm, we call it W-algorithm, which extracts variable length dictionaries of interests from several real genomic sequences and collects words having a recurrence distribution maximally different than their random distribution. Such a selection is developed by computing the (locally) maximum divergence, from random sequences, of the RDD of each string obtained by elongating an initial *seed word*

Copyright (c) IARIA, 2021. ISBN: 978-1-61208-859-4 

over the genome. The divergence from random sequences is a crucial issue in information analysis of strings [34, 35] and in analyzing mathematical properties of dictionaries. The methodology in [24] to find dictionaries is therefore here improved by the V-algorithm, and a more general approach is proposed (Section II-B) by means of the RDD based W-algorithm, that works with the global word recurrence distance distribution rather than with only a first slice of it.

## II. Material and Methods

This section summarizes the genomic word extraction methodology reported in [24], which was our starting point to develop a variant of it, the V-algorithm, and then introduces a novel RDD-based extraction algorithm, called W-algorithm. We also propose some criteria to evaluate extracted genomic dictionaries. Following the terminology from our previous work [12], a genome is a string over the genomic alphabet $\Gamma = \{A, C, G, T\}$. Given a genome $G$, we call $D_k(G) \subseteq \Gamma^k$ the *$k$-dictionary* of all $k$-mers occurring in the genome $G$. Given a word $\alpha \in D_k(G)$), a recurrence distance distribution (RDD) informs how many times $\alpha$ occurs at a given distance $d$. Thus, a recurrence is a pair of positions $(p_1, p_2)$ (with $p_1, p_2 < |G|$ and $p_1 < p_2$) such that $\alpha$ occurs in $p_1$ and $p_2$ and no other occurrences of $\alpha$ are in the middle. The recurrence distance is given by $p_2 - p_1$.

### A. A clustering coefficient based approach

RDD has been used to identify keywords by applying a methodology that associates a clustering coefficient $C$ to $k$-mers [24]. The main idea is based on the fact that keywords are not uniformly distributed among a literary text, instead they are clustered. The approach combines the information provided by the spatial distribution of a word along the text (via the clustering coefficient) and its frequency, since the statistical fluctuation depends on the frequency. This basic approach has been used in [25] to assign a relevance to 6-mers and 8-mers in *Homo sapiens* and *Mus musculus*. The 8-mers were sorted by their normalized clustering coefficient (called $\sigma_{nor}$), and it has been shown that part of the top-200 clustered words (about 70%) appears in known functional biological elements, like coding regions and transcription factor binding sites.

The whole recurrence distribution is synthesised with a single parameter $\sigma$, to quantify the clustering level, previously presented in [9] for studying the energy levels of quantum disorder systems [36], and a clustering degree $\sigma_{nor}$ assigned to words, for the identification of keywords in literary texts, obtained by means of the relation between the $\sigma$ of a real word and the theoretical expected one (coming from a theoretical hypothesized distribution), as in the following.

For a given word, the parameter $\sigma$ is the standard deviation of its normalized set of recurrence distances, $\sigma = s/\bar{d}$, where $s$ is the standard deviation of the recurrence distance distribution, and $\bar{d}$ is the average recurrence distance. When the RDD is a geometric distribution, the parameter is denoted by $\sigma_{geo}$ and it is equal to $\sqrt{1-p}$, since $s = \sqrt{1-p}/p$ and $\bar{d} = 1/p$, where $p$ is the word frequency. Thus, the resultant normalized clustering measuring $\sigma_{nor}$ of the given word is given by $\frac{\sigma}{\sigma_{geo}} = \frac{s/\bar{d}}{\sqrt{1-p}}$. For values of $\sigma_{nor}$ near to 1, the recurrence distribution of the word is close to the geometric one, thus it indicates a randomness of the word. In fact, a random sequence is generated by a Bernoullian process, then different occurrences of a given word are independent events, and the event of having $k$ occurrences of a word (in a segmentation unit) follows a Poisson distribution. Therefore, according to probability theory [37] its waiting time, that is the distance at which a word recurs, is an exponential distribution (having a geometric distribution as a discrete counterpart).

For words with low multiplicity, the statistical fluctuation is much larger, and it is possible to obtain an higher $\sigma_{nor}$ for rare words placed at random, and they would be misidentified as keywords. Thus, the authors applied a correction by a Z-score measure that combines the clustering of a word and its multiplicity $n$. The resultant clustering measure $C$ is given by the following equation: $C(\sigma_{nor}, n) = \frac{\sigma_{nor} - \langle\sigma_{nor}\rangle(n)}{sd(\sigma_{nor})(n)}$, where $\langle\sigma_{nor}\rangle(n) = \frac{2n-1}{2n+2}$ and $sd(\sigma_{nor})(n) = \frac{1}{\sqrt{n}(1+2.8n^{-0.865})}$. Parameter values were obtained via extensive simulations, by taking into account the distribution of $\sigma_{nor}$ in random texts. They represent the mean value and the standard deviation of such empirical distribution. The $C$ coefficient measures the deviation of $\sigma_{nor}$ with respect to the expected value in a random text, in units of the expected standard deviation. In this case, $C = 0$ indicates randomness, $C > 0$ that the word is clustered and $C < 0$ that the word *repels* itself.

In [24] also an approach to explore the lineage of a word (from a short word to one of its possible elongations), without any knowledge about the effective word length, was provided. Given an initial word length $k_0$, some of the words in $D_{k_0}(G)$ are selected, according to their $C$ measure, that must be greater then a $C_0$ measure corresponding to a fixed percentile (usually $0.05$). Successively, for each of these initial words, their lineage is explored by selecting only the *elongations having a $C$ measure greater than $C_0$, and up to a fixed maximal word length*: these are properly the two points we changed in the V-algorithm presented in the next section. The longest visited lineage is selected as a word with semantic meaning, and the process is repeated for different values of $k_0$ (ranging from 2 to 35), until a dictionary is obtained by discarding repeating words.

### B. The RDD-based W-algorithm

We use RDD to calculate the divergence of the real distribution of a word within the genome from its frequency over a random string with the same genome length [29, 38]. Such a divergence is used as a measure of the information content of a word. Low expressive words are elongated by an expansion procedure, until they reach a reasonable level of *significance* according to which they are classified as genomic words of the extracted dictionary.

We assume that the higher the entropic divergence from the above exponential distribution, the more specialized and evolutionary selected is the genomic element. In this sense, low multiplicity words already represent elements owning high

Copyright (c) IARIA, 2021. ISBN: 978-1-61208-859-4 

level of significance. Instead, for what concerns repeats, we associate their *meaning* with their repetitiveness-profile, as it is revealed by their RDD. A word has to occur along the genomic sequence several times and at different distances. See an example in Figure 1, where the exponential distribution represents the random recurrence behaviour of the word. RDD of words along real genomes is often sparse, meaning that several distances (of recurrence) actually do not appear in the genome. This is why we evaluate the sound (i.e., more fitting) exponential distribution after removing peaks, that are absent in exponential functions, and by imposing a normalization ensuring the overall unitary probability.

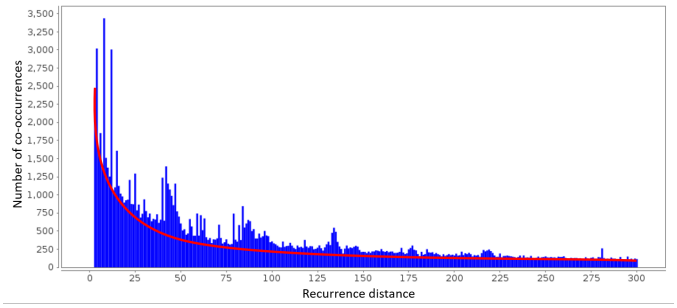


Fig. 1. RDD of word CGC (the jagged curve) in human chromosome 22

The degree of significance of a word is its *random deviation*, measured by the function in 1, based on the the entropic divergence (Kullback-Leibler divergence [27]), between the real RDD of a word (over the analysed genome) and its expected exponential distribution.

More technically, given a word $\alpha$, which occurs in a genome $G$, we calculate its random deviation as the entropic divergence between its RDD and a suitable exponential distribution. To this aim, we first extract the real RDD of $\alpha$ over $G$, which we refer as $R_\alpha$. Then, we estimate a two parameters exponential distribution $E_\alpha$, by making use of the Nelder and Mead Simplex algorithm [39]. A denoised distribution is used as input for the estimation procedure: it is obtained by applying a low-pass filter (over $R_\alpha$) in order to attenuate peaks. Afterwards, we remove from $E_\alpha$ the domain values which are not present in $R_\alpha$, namely the gaps of $R_\alpha$. Successively, both $R_\alpha$ and $E_\alpha$ are normalized in order to become probability distributions. Finally, the random deviation of $\alpha$ is chosen as:

$$r(\alpha) = \max(KL(R_\alpha, E_\alpha)\ ,\ KL(E_\alpha, R_\alpha)), \qquad (1)$$

where $KL$ is the asymmetric Kullback-Leibler entropic divergence.

In our algorithm (reported in Listing 1) estimation of the information content of a word $\alpha$ is computed by the function $r(\alpha)$. Word elongation is realized until the random deviation does not start to decrease. As it may be seen in Figure 2, smaller seeds allow the algorithm to generate words $\alpha$ corresponding to the first peak (local maximum) of $r(\alpha)$. To produce a longer significant word $\alpha$, corresponding to the second peak of $r(\alpha)$, a longer seed has to be taken as a starting string. In all our computational experiments, $r(\alpha)$ showed

```
W:=∅;
ForEach α ∈ D0:
    Elongate(α, W)
W := W \ D0;
Return W
```

Listing 1. Extraction Algorithm

```
if r(αx) ≤ r(α), ∀x ∈ Γ then W := W ∪ {α}
else ForEach x ∈ Γ
        if r(αx) > r(α) then Elongate(αx, W)
```

Listing 2. Elongation procedure: Elongate $(\alpha, W)$

to have only two peaks, whose localization depends on the genome length.

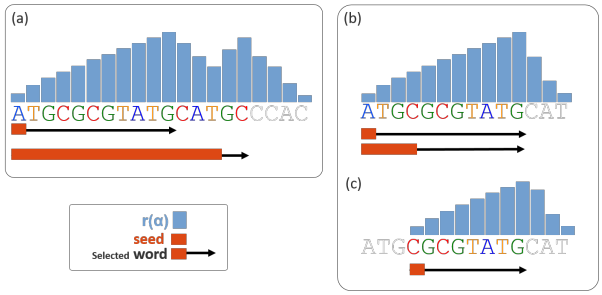


Fig. 2. Expansion procedure

We would like to extract all the words $\alpha$ such that both $\alpha[1, |\alpha|-1]$ and $\alpha x$ (where $\alpha x$ is any elongation of $\alpha$ occurring in $G$ at least once) own a lower level of significance, namely a lower random deviation, with respect to $\alpha$. The goal can be reached by examining all the words within $G$ from monomers up to a word length equal to the maximum repeat length of $G$, and by discarding hapaxes. However, such an approach turns out highly expensive, and it cannot be applied efficiently for long genomes. Thus, we developed an expansion procedure with the aim of elongating seed words, let say monomers, up to more meaningful words. The (variable length dictionary) extraction algorithm, combining word elongation and random deviance test (in the expansion procedure) is given by two recursive functions in Listings 1 and 2, where $D_0$ denotes the set of seeds $D_{k_0}(G)$.

The main idea is to compare the random deviation of a word with those of its elongations. If an elongation results in a word more significant than its root, then the root word is discarded and the elongated word is selected. The process is applied recursively over the word branching of the selected elements (see Listing 2). Seeds are discarded from the output dictionary. Three steps are implemented to compute random deviations. For all factors $\alpha$ of the genome $i$) RDD of the current word $\alpha$ is computed, by also removing distribution noise (peaks) and transforming $R_\alpha$ into a probability distribution; $ii$) an exponential distribution $E_\alpha$ is computed from $R_\alpha$ and normalized to be a probability distribution; $iii$) random deviation $r_\alpha$ is computed by means of the Kullback-Leibler (entropic) divergence.

Copyright (c) IARIA, 2021. ISBN: 978-1-61208-859-4

We employ two elongating functions (along both directions of the genome double string) and the resulting dictionary is the union of the dictionaries obtained with the two elongations. We refer with $W_{L2R}$ and $W_{R2L}$ as the dictionaries extracted by following the $5' - 3'$ and $3' - 5'$ verses, respectively, and with $W = W_{L2R} \cup W_{R2L}$ as the resulting dictionary.

### C. Dictionary evaluation

Extracted dictionaries are evaluated by means of information measurements, such as *the word length distribution* of their elements. Two other parameters are the *sequence coverage*, which is the percentage of positions $i$ in the genome such that $G[j, k]$ is a word of the extracted dictionary $D$ for $j < i < k$, and the *average positional coverage*, which is the average over positions $i$ of number of words $G[j, k]$ for $j < i < k$ of the dictionary $D$. They are denoted by $cov(G, D)$ and $avg(covp(G, D))$, respectively. A *good* dictionary must have a high sequence coverage, but a low overlapping degree among its elements. In fact, if we consider $D_k(G)$ as a language, for a certain value of word length $k$, then it has the maximum sequence coverage (all positions of the genome would be involved by at least one $k$-mer) but also the maximum positional coverage, since each position of the sequence is involved by up to $k$ different words of the dictionary. On an ideally good dictionary, both parameters are close to one, meaning that its words cover almost the entire genome and tend to not overlap.

## III. Results

Both algorithms described in previous section were run over all human chromosomes belonging to the reference assembly hg19.

### A. Dictionaries extracted by the V-algorithm

Table I shows the number of extracted words (that is, dictionary sizes), for each single human chromosome, and their union at the bottom, for both the algorithm in [24] and the V-algorithm, by starting from different seed lengths, and by implementing two filters as redundancy strategies: one discarding duplicates (same words coming from different seed lengths) and the other discarding prefixes (in order to estimate the relative amount of prefixes).

The result is that the V-algorithm is able to select a smaller set of words, with a lower gap between the two redundancy discarding strategies. This is essentially due to the fact that the higher is $k$ the lower are the $C$ measures of $k$-mers. Therefore, comparing the $C$ measure of a word, relatively longer than $k_0$, with the measure of its proper prefix is more restrictive than a comparison with the measure of the initial word of length $k_0$. From this behaviour, we can speculate that the V-algorithm selects words with an higher semantic meaning.

In Table I, it is evident that the V-algorithm extracts a smaller amount of duplicates and prefixes than the algorithm in [24] (even when starting from seeds with different length). Indeed, *smaller variable length dictionaries were extracted* by the V-algorithm, with fewer duplicate discarding steps, and a smaller amount of prefixes (which needed to be discarded in the original algorithm).

TABLE I
NUMBER OF EXTRACTED WORDS BY THE ORIGINAL AND MODIFIED ALGORITHMS

| Chr | Orig. no dup. | Orig. no pref. | ratio | V-algo no dup. | V-algo. no pref. | ratio |
|---|---|---|---|---|---|---|
| 1 | 276,178 | 210,728 | 0.763 | 57,064 | 57,055 | 1.000 |
| 2 | 281,698 | 227,544 | 0.808 | 119,582 | 118,368 | 0.990 |
| 3 | 259,805 | 203,888 | 0.785 | 102,640 | 101,142 | 0.985 |
| 4 | 251,067 | 201,760 | 0.804 | 108,229 | 106,879 | 0.988 |
| 5 | 259,167 | 207,300 | 0.800 | 112,846 | 111,581 | 0.989 |
| 6 | 255,025 | 198,487 | 0.778 | 106,193 | 104,510 | 0.984 |
| 7 | 269,392 | 208,465 | 0.774 | 113,139 | 111,840 | 0.989 |
| 8 | 259,586 | 206,241 | 0.794 | 118,551 | 117,295 | 0.989 |
| 9 | 212,362 | 152,523 | 0.718 | 33,886 | 33,878 | 1.000 |
| 10 | 234,663 | 186,844 | 0.796 | 100,616 | 99,595 | 0.990 |
| 11 | 249,374 | 188,012 | 0.754 | 94,484 | 93,417 | 0.989 |
| 12 | 247,842 | 187,931 | 0.758 | 99,147 | 97,579 | 0.984 |
| 13 | 176,546 | 149,563 | 0.847 | 81,634 | 78,868 | 0.966 |
| 14 | 209,881 | 162,515 | 0.774 | 94,312 | 90,313 | 0.958 |
| 15 | 207,173 | 177,125 | 0.855 | 107,114 | 103,917 | 0.970 |
| 16 | 229,208 | 166,653 | 0.727 | 62,732 | 62,673 | 0.999 |
| 17 | 204,905 | 160,475 | 0.783 | 85,091 | 84,303 | 0.991 |
| 18 | 161,710 | 131,900 | 0.816 | 65,985 | 65,558 | 0.994 |
| 19 | 258,781 | 197,822 | 0.764 | 123,913 | 122,541 | 0.989 |
| 20 | 171,474 | 131,434 | 0.766 | 66,320 | 65,597 | 0.989 |
| 21 | 130,763 | 100,427 | 0.768 | 50,698 | 50,233 | 0.991 |
| 22 | 147,002 | 120,259 | 0.818 | 77,797 | 74,511 | 0.958 |
| X | 279,938 | 213,093 | 0.761 | 124,793 | 123,006 | 0.986 |
| Y | 194,014 | 137,284 | 0.708 | 66,088 | 65,986 | 0.998 |
| union | 4,281,701 | 3,737,766 | 0.873 | 1,813,776 | 1,798,241 | 0.991 |

### B. Dictionaries extracted by the W-algorithm

The RDD-based W-algorithm was applied (with values for seed length from the range $1 - 12$) to extract genomic dictionaries from each human chromosome, and some analysis was performed also on the union of such 24 dictionaries. However, here we show data only for some (more explicable) chromosomes, for (more significant) seed lengths up to 8.

TABLE II
WORD LENGTH DISTRIBUTION OF HUMAN CHROMOSOME 1

| | $k_0$ | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| $k$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| 4 | 2 | 13 | 20 | | | | | |
| 5 | 31 | 134 | 202 | 272 | | | | |
| 6 | **63** | **349** | **517** | **995** | **1,261** | | | |
| 7 | 57 | 180 | 232 | 350 | 475 | 1,343 | | |
| 8 | 57 | 193 | 277 | 430 | 679 | 3,001 | 10,668 | |
| 9 | 10 | 144 | 241 | 529 | 1,073 | 7,602 | 29,521 | 53,314 |
| 10 | 5 | **201** | **326** | **794** | **1,391** | **9,126** | 59,951 | 129,872 |
| 11 | 2 | 151 | 233 | 569 | 923 | 4,302 | **63,089** | **184,296** |
| 12 | | 64 | 91 | 198 | 323 | 973 | 24,275 | 97,646 |
| 13 | | 21 | 30 | 51 | 81 | 225 | 4,592 | 20,670 |
| 14 | | 2 | 3 | 10 | 18 | 40 | 875 | 3,525 |
| 15 | | 2 | 2 | 5 | 6 | 11 | 190 | 724 |
| 16 | | 4 | 5 | 5 | 5 | 9 | 54 | 165 |
| 17 | | 1 | 1 | 2 | 2 | 3 | 17 | 54 |
| 18 | | | | | | | 5 | 19 |
| 19 | | | | | | | | 5 |
| 20 | | | | | | | | 6 |
| 21 | | | | | | | | 3 |
| 22 | | | | | | | | 6 |
| 23 | | | | | | | | 1 |

The Word Length Distribution (WLD) related to human chromosomes 1 is shown in Table II by reporting the cardinality of words having a given length and being generated by starting from a given seed length. A common feature is to have two modes in the $k$-dictionary sizes, that is, two local maximum values (indicated in bold) for some lengths $k$. In

Table II, such values are 6 (for seeds long from 1 to 5) and 10-11 (for seeds long from 2 to 8). Although they do not have fixed values (for tests performed on the other human chromosomes and not shown here), they are not very variable.

Another empirical result, confirmed on all the other chromosomes, is that the dictionary generated by starting from seeds $k-1$ long is a proper subset of that generated by starting from seeds $k$ long, apart of the words long $k$. In fact, words with the same length of the seed are eliminated by the algorithm and do not appear in the WLD tables.

Extracted dictionaries are evaluated according to both their sequence and their (average) positional coverage: these data related to chromosome 1 are reported in Table III and Table IV respectively, where it is clear that parameter goodness does not increase with the word or seed length $k_0$.

TABLE III
HUMAN CHROMOSOME 1: SEQUENCE COVERAGE VALUES

| | $k_0$ | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| $k$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| 4 | | 0.0291 | 0.0291 | | | | | |
| 5 | 0.0309 | 0.0790 | 0.1362 | 0.1681 | | | | |
| 6 | 0.0269 | **0.3149** | **0.5504** | **0.7767** | **0.8426** | | | |
| 7 | **0.0742** | 0.2479 | 0.3878 | 0.6430 | 0.7691 | **0.8141** | | |
| 8 | 0.0285 | 0.0616 | 0.0899 | 0.1187 | 0.1384 | 0.1643 | **0.2634** | |
| 9 | 0.0115 | 0.0209 | 0.0303 | 0.0499 | 0.0615 | 0.0714 | 0.1593 | **0.6315** |
| 10 | 0.0008 | 0.0054 | 0.0071 | 0.0128 | 0.0206 | 0.0329 | 0.0974 | 0.5388 |
| 11 | **0.0025** | **0.0077** | **0.0088** | **0.0108** | **0.0127** | 0.0174 | 0.0602 | 0.3509 |
| 12 | | 0.0028 | 0.0031 | 0.0081 | 0.0089 | 0.0101 | 0.0342 | 0.2858 |
| 13 | 0.0000 | 0.0006 | 0.0013 | 0.0054 | 0.0065 | 0.0070 | 0.0155 | 0.1209 |
| 14 | 0.0035 | 0.0048 | 0.0049 | 0.0056 | 0.0065 | 0.0066 | 0.0101 | 0.0451 |
| 15 | 0.0026 | 0.0036 | 0.0036 | 0.0050 | 0.0052 | 0.0052 | 0.0065 | 0.2140 |
| 16 | | 0.0016 | 0.0017 | 0.0017 | 0.0071 | 0.0028 | 0.0032 | 0.0090 |
| 17 | | 0.0011 | 0.0011 | 0.0012 | 0.0013 | 0.0013 | 0.0014 | 0.0031 |
| 18 | | 0.0006 | 0.0006 | 0.0006 | 0.0006 | 0.0012 | 0.0012 | 0.0020 |
| 19 | | | | | | | 0.0000 | 0.0003 |
| 20 | | | | | | | 0.0000 | 0.0002 |
| 21 | | | | | | | | 0.0001 |
| 22 | | | | | | | | 0.0000 |
| 23 | | | | | | | | |
| 24 | | | | | | | | 0.0000 |

By observing the data in Table III, the best coverage of the chromosome (corresponding value 0.84) is obtained by the examers obtained starting from 5-mers as seeds, while the average positional coverage of such a dictionary is 2.7715 (see Table IV), which is far from one. However, this dictionary was our choice for the chromosome clustering analysis described below, because we gave a priority of importance to sequence coverage. Relatively to only positional coverage values, in Table IV we may notice that words of length 10 (or longer, for instance 15) exhibit good (i.e., less than 2) values for any seed length up to 7, while examers have good positional coverage with shorter seeds (long up to 3).

Finally, we extracted dictionaries of examers on each single human chromosome, and from their pairwise intersections, in absolute and relative terms, we found interesting results, reported in Figure 3, where four groups of chromosomes may be identified at the second level of the dendrogram, having cardinalities of dictionary intersection of the same order of that of the extracted dictionary from each single chromosomes (see leaves of the dendogram). Our dictionary based method was then capable to discriminate by structure similarity the following clusters of human chromosomes.

TABLE IV
HUMAN CHROMOSOME 1: AVERAGE POSITIONAL COVERAGE

| | $k_0$ | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| $k$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| 4 | | 1.0078 | 1.0078 | | | | | |
| 5 | 1.0807 | 1.1690 | 1.2411 | 1.4198 | | | | |
| 6 | **1.1539** | **1.3022** | **1.6590** | **2.3201** | **2.7715** | | | |
| 7 | 1.0934 | 1.2876 | 1.4587 | 1.9817 | 2.5877 | 2.9160 | | |
| 8 | 1.1569 | 1.2590 | 1.3125 | 1.4228 | 1.5184 | 1.5836 | 1.5572 | |
| 9 | **1.4480** | **1.5411** | **1.5211** | **1.7039** | **1.8791** | **1.8661** | 1.5470 | 1.7484 |
| 10 | 1.0006 | 1.1090 | 1.1033 | 1.1697 | 1.1926 | 1.2632 | 1.2580 | 1.5457 |
| 11 | **4.0810** | **2.1729** | **2.0809** | **1.9100** | **1.7829** | **1.6131** | **1.3009** | 1.3658 |
| 12 | | 1.0654 | 1.0624 | 1.1926 | 1.1809 | 1.1716 | 1.1507 | 1.3455 |
| 13 | 1.0000 | 1.0000 | 1.0000 | 1.1355 | 1.3769 | 1.3530 | 1.2340 | 1.3709 |
| 14 | 1.0000 | 1.0000 | 1.0000 | 1.0551 | 1.2244 | 1.2235 | 1.1687 | 1.3807 |
| 15 | 1.000 | 1.1446 | 1.1445 | 1.1065 | 1.1739 | 1.1725 | 1.1444 | 1.2559 |
| 16 | | 1.2684 | 1.2636 | 1.2588 | 1.2539 | 1.1544 | 1.1447 | 1.1148 |
| 17 | | 1.0000 | 1.0000 | **1.3982** | **1.3957** | **1.3948** | **1.3608** | **1.3440** |
| 18 | | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0015 | 1.0187 |
| 19 | | | | | | | 1.0000 | 1.0000 |
| 20 | | | | | | | 1.0000 | 1.0000 |
| 21 | | | | | | | | 1.0000 |
| 22 | | | | | | | | 1.0000 |
| 23 | | | | | | | | |
| 24 | | | | | | | | 1.0000 |

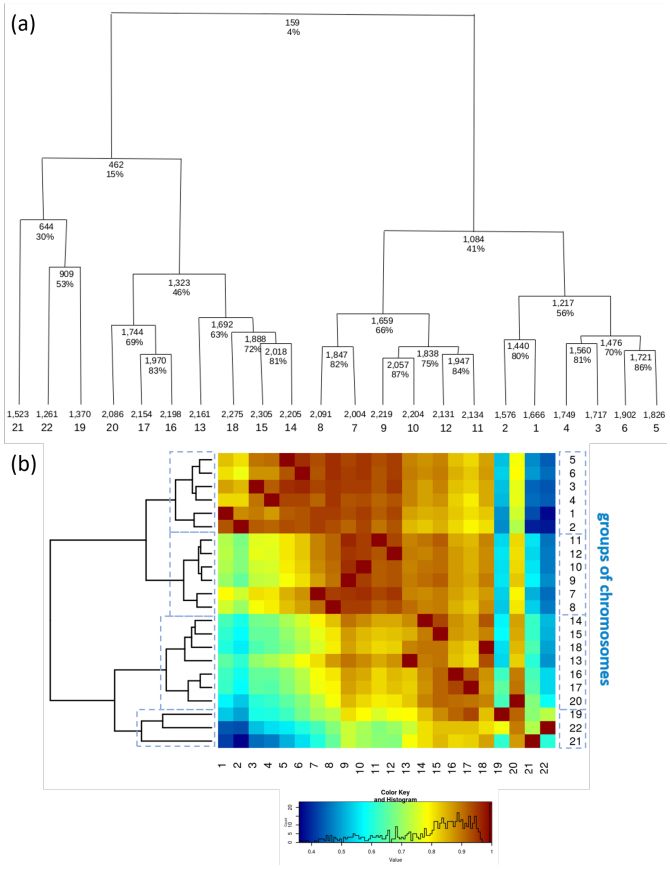


Fig. 3. **(a)** Human chromosome similarities percentages. **(b)** Heatmap of human cromosome similarity.

The dictionary of examers obtained by the algorithm from seeds long 5 was here employed to cluster all human chromosomes (see Figure 3). All chromosomes share very few examers (159 are common to all, over the 1,666 extracted words) which we exhibit as informative conserved sequences, a sort of product by evolution selection, to be further analyzed for their biological characterization.

Copyright (c) IARIA, 2021. ISBN: 978-1-61208-859-4

## IV. Conclusions and Discussion

Given a genome, we extract a specific set of its factors which represent the building blocks, or semantic units, of a dictionary significant for the genome language. In this work, we have described an information theoretical methodology to extract relatively small genomic dictionaries, which have good properties in terms of genome coverage.

Three methods were presented. One from the literature, introduced in [24], which was our starting point in terms of basic ideas, the second method is a variant of this, called V-algorithm, more efficient and appropriate to extract genomic dictionaries, and finally, our RDD based W-algorithm, which originally combines a criterion of anti-randomness with a criterion of elongation of seeds to select variable length factors. The application of the state of the art methodology and the V-algorithm to human chromosomes show that both algorithms often fail in extending seeds, and when they success, they more-likely extract very long words, which sparsely cover the investigated sequences. The point of our approach is to produce relatively small dictionaries with both sequence and average positional coverage as close as possible to one. The goal is reached thanks to the proposed W-algorithm. We have shown that preferred seed lengths emerge, from an observation of sequence and positional genome coverage that provide a better coverage. Moreover, dictionaries of examers were identified to reveal a clear similarity pattern for human chromosomes.

## References


[1] G. S. Ginsburg and H. F. Willard, Eds., *Genomic and Precision Medicine – Foundations, Translation, and Implementation*. (Third Edition): Elsevier, 2017.

[2] R. Mantegna et al., "linguistic features of noncoding dna sequences," *Physical Review Letters*, vol. 73, no. 23, pp. 3169–72, 1994.

[3] D. B. Searls, "The language of genes," *Nature*, vol. 420, pp. 211–217, 2002.

[4] M. Sadovsky, J. Putintseva, and A. S. Shchepanovsky, "Genes, information and sense: Complexity and knowledge retrieval," *Theory in Biosciences*, vol. 127, no. 2, pp. 69–78, 2008.

[5] S. Neph et al., "An expansive human regulatory lexicon encoded in transcription factor footprints," *Nature*, vol. 489, pp. 83–90, 2012.

[6] G. Franco and V. Manca, "Decoding genomic information," in *Computational Matter*, S. Stepney, S. Rasmussen, and M. Amos, Eds. Springer, Cham, 2018, ch. 9, pp. 129–149.

[7] U. Ferraro Petrillo, G. Roscigno, G. Cattaneo, and R. Giancarlo, "Informational and linguistic analysis of large genomic sequence collections via efficient hadoop cluster algorithms," *Bioinformatics*, vol. 34, no. 11, pp. 1826–1833, 2018.

[8] Z. Zhang et al., "Statistical analysis of the genomic distribution and correlation of regulatory elements in the encode regions," *Genome Res.*, vol. 17, no. 6, pp. 787–97, 2007.

[9] M. Ortuno, P. Carpena, P. Bernaola-Galván, E. Munoz, and A. Somoza, "Keyword detection in natural languages and DNA," *EPL (Europhysics Letters)*, vol. 57, no. 5, p. 759, 2007.

[10] T. E. P. Consortium, "An integrated encyclopedia of DNA elements in the human genome," *Nature*, vol. 489, no. 7414, pp. 57–72, 2012.

[11] F. Zambelli, G. Pesole, and G. Pavesi, "Motif discovery and transcription factor binding sites before and after the next-generation sequencing era," *Briefings in bioinformatics*, p. bbs016, 2012.

[12] A. Castellini, G. Franco, and V. Manca, "A dictionary based informational genome analysis," *BMC Genomics*, vol. 13, no. 1, p. 485, 2012.

[13] V. Mäkinen, D. Belazzougui, F. Cunial, and A. Tomescu, *Genome-Scale Algorithm Design: Biological Sequence Analysis in the Era of High-Throughput Sequencing*. Cambridge: Cambridge University Press, 2015.

[14] S. P. Garcia, A. J. Pinho, J. M. Rodrigues, C. A. Bastos, and P. J. Ferreira, "Minimal absent words in prokaryotic and eukaryotic genomes," *PLoS One*, vol. 6, no. 1, 2011.

[15] A. L. Price, N. C. Jones, and P. A. Pevzner, "De novo identification of repeat families in large genomes," *Bioinformatics*, vol. 21, no. suppl_1, pp. i351–i358, 2005.

[16] I. Grissa, G. Vergnaud, and C. Pourcel, "Crisprfinder: a web tool to identify clustered regularly interspaced short palindromic repeats," *Nucleic acids research*, vol. 35, no. suppl_2, pp. W52–W57, 2007.

[17] J. Qian and M. Comin, "Metacon: unsupervised clustering of metagenomic contigs with probabilistic k-mers statistics and coverage," *BMC Bioinformatics*, vol. 20, no. 367, 2019.

[18] M. Aharon, M. Elad, and A. Bruckstein, "K-svd: An algorithm for designing overcomplete dictionaries for sparse representation," *Signal Processing, IEEE Transactions*, vol. 54, p. 4311–4322, 2006.

[19] F. Zhou, V. Olman, and Y. Xu, "Barcodes for genomes and applications," *BMC Bioinformatics*, vol. 9, p. 546, 2008.

[20] S. Vinga, "Information theory applications for biological sequence analysis," *Briefings in bioinformatics*, vol. 15, no. 3, pp. 376–389, 2013.

[21] G. E. Sims, S. Jun, G. A. Wu, and S. Kim, "Alignment-free genome comparison with feature frequency profiles (ffp) andoptimal resolutions," *PNAS*, vol. 106, no. 8, pp. 2677–82, 2009.

[22] B. Chor et al., "Genomic dna *k*-mer spectra: models and modalities," *Genome Biology*, vol. 10, p. R108, 2009.

[23] Y. Zheng et al., "Evolutionary mechanism and biological functions of 8-mers containing cg dinucleotide in yeast," *Chromosome Research*, vol. E-pub ahead of print], pp. 1–17, 2017.

[24] P. Carpena, P. Bernaola-Galván, M. Hackenberg, A. Coronado, and J. Oliver, "Level statistics of words: Finding keywords in literary texts and symbolic sequences," *Physical Review E*, vol. 79, no. 3, p. 035102, 2009.

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

[26] G. Franco and A. Milanese, "An investigation on genomic repeats," in *Conference on Computability in Europe – CiE*, ser. Lecture Notes in Computer Science, vol. 7921. Springer, 2013, pp. 149–160.

[27] A. Thomas and T. M. Cover, *Elements of Information Theory*. John Wiley, 1991.

[28] J. H. Holland, *Emergence: from chaos to order*. Perseus books: Cambridge, Massachusetts, 1998.

[29] S. G. Kong et al., "Quantitative measure of randomness and order for complete genomes," *Phys Rev E*, vol. 79, no. 6, p. 061911, 2009.

[30] P. Kolekar, M. Kale, and U. Kulkarni-Kale, "Alignment-free distance measure based on return time distribution for sequence analysis: Applications to clustering, molecular phylogeny and subtyping," *Molecular phylogenetics and evolution*, vol. 65, pp. 510–22, 2012.

[31] A. S. Nair and T. Mahalakshmi, "Visualization of genomic data using inter-nucleotide distance signals," *Proceedings of IEEE Genomic Signal Processing*, vol. 408, 2005.

[32] V. Afreixo, C. A. Bastos, A. J. Pinho, S. P. Garcia, and P. J. Ferreira, "Genome analysis with inter-nucleotide distances," *Bioinformatics*, vol. 25, no. 23, pp. 3064–3070, 2009.

[33] C. A. Bastos, V. Afreixo, A. J. Pinho, S. P. Garcia, J. Rodrigues, and P. J. Ferreira, "Inter-dinucleotide distances in the human genome: an analysis of the whole-genome and protein-coding distributions," *Journal of Integrative Bioinformatics*, vol. 8, no. 3, p. 172, 2011.

[34] L. Gatlin et al., *Information theory and the living system*. Columbia University Press, 1972.

[35] S. P. Harter, "A probabilistic approach to automatic keyword indexing," Ph.D. dissertation, University of Chicago, 1974.

[36] P. Carpena, P. Bernaola-Galván, and P. C. Ivanov, "New class of level statistics in correlated disordered chains," *Physical review letters*, vol. 93, no. 17, p. 176804, 2004.

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

[38] A. Kolmogorov, "On tables of random numbers," *Theoretical Computer Science*, vol. 207, no. 2, pp. 387–395, 1998.

[39] J. A. Nelder and R. Mead, "A simplex method for function minimization," *The computer journal*, vol. 7, no. 4, pp. 308–313, 1965.


Copyright (c) IARIA, 2021. ISBN: 978-1-61208-859-4 