An Investigation to Test Spectral Segments as Bacterial Biomarkers

Silvia Astorino[1], Vincenzo Bonnici[2], and Giuditta Franco[1(✉)]

[1] University of Verona, Strada le Grazie 15, 37134 Verona, Italy
silvia.astorino@studenti.univr.it, giuditta.franco@univr.it
[2] University of Parma, Parco Area delle Scienze, 7/A, 43124 Parma, Italy
vincenzo.bonnici@unipr.it
https://www.univr.it/, https://www.unipr.it/

Abstract. A dictionary-based bacterial genome analysis is performed, through specific k-long factors (called *res*) and their maximal right elongation along the genome (called *spectral segment*), in order to find discriminating biomarkers at the genus and species level. The aim is pursued through a k-mer-based approach previously introduced, here applied on genomes of different bacterial taxa. Intervals for values of k are identified to obtain meaningful genomic fragments, whose collection is a suitable representation to compare genomes according to informational indexes and Jaccard's similarity matrices. Corresponding dictionaries of k-mers are identified to discriminate bacterial genomes at genus and species level. This approach appears competitive in terms of performance (e.g., species discrimination) and size with respect to traditional barcoding methods.

Keywords: Barcoding · k-mers · right special factors · spectral segments

1 Introduction

Computational methodologies avoiding alignment of biological sequences constitute a relevant field of bioinformatics, including alignment-free methods [13,26], which show a considerable reduced computational cost with respect to alignment-based approaches. Alignment-free analysis is often based on dictionaries composed by relatively small words of the same length k, called k-mers, which are extracted from biological sequences [6,23,30]. Those methods find applicability in multiple contexts, such as genome assembly [7,8], genetic reconstruction [15,27,29] and DNA barcoding [10,12]. In particular, they allow handling large quantities of sequences in metagenomic studies [24], which characterize unknown taxa present in an environmental sample (or in a microbiome [28]).

In this paper we continue the investigation initiated in [4], where some informational concepts derived by the notion of k-spectrum applied to genomic k-mers were analyzed. Starting from the dictionary of the k-mers having the property

D. Genova and J. Kari (Eds.): UCNC 2023, LNCS 14003, pp. 1–16, 2023.
https://doi.org/10.1007/978-3-031-34034-5_1

to be followed by the same nucleotide in all their occurrences on the genome (we call RES_k these *Right-Extendable Sequences*, or simply RES when the role of k is obvious), spectral segments have been defined as the iterated $(k-1)$-long overlap concatenations of RES along the genome, maximally and uniquely right-elongated.

In this work we have extracted dictionaries of RES_k, called U_k, and corresponding dictionaries of spectral segments, called Sp_k, from several bacterial genomes downloaded from NCBI (details on the data set are reported in Sect. 2) in order to identify, if existing, a range of the values of k for which this dictionary-based genome representation is valid to classify biological sequences, according to their species or genus membership. We found out some ranges for values of k such that the knowledge of U_k and/or Sp_k would allow us to discriminate the presence or absence of one species or one genus in an unknown bacterial population of an environmental sample. Experimental results are reported in the following, together with interesting notes both on the overlapping of genes (or coding regions) with spectral segments and on the efficiency of the algorithm employed to extract such segments.

The state of the art for this work ranges in a wide variety of contexts. In combinatorics on words the notion of right special factors of a fixed length k has been investigated [1,9,16], where a substring u of a word w is special if there exist at least two occurrences of u in w followed on the right by two distinct letters (i.e., there exist at least two distinct letters a and b such that the strings va and vb are both factors of w). Such k-mers are exactly the non-RES words, because by definition RES words are followed on the right by one same character, in all their occurrences. In the literature of computational genomics, there are several examples of methods that extract genomic substrings, as unitigs [15,27] and omnitigs [25] already defined as discriminant for taxa, carriers of biological significance, and reliable fragments in the reconstruction of a genome. In this work, U_k and Sp_k dictionaries are tested and proposed to classify bacterial genomes, as an alternative to segments in the literature, and potentially to markers commonly used in the laboratory.

Our method could be a competitive solution for supervised machine learning methods, where the values of k use to range from 1 to 6. In [19,20] for example, authors implemented a machine learning approach for the recognition of specific classes of genomic sequences (mainly retrotransposons) based on 6-mers multiplicity. Our results indicate good performance in terms of the ability to discriminate between species (by not necessarily identifying them) in comparison to the use of short DNA sequences, for the purpose of species discrimination (previously coined as DNA barcoding).

A possible application of our approach is indeed DNA barcoding, where usually a single marker gene located on RNA (the 16S rRNA, that is, the coding gene for 16S ribosomal RNA) is employed to characterize bacteria, particularly in the human microbiota. However, traditional barcoding studies usually fail to reach the discrimination at the species level [28]. Metabarcoding is then a point of application of any technique for characterizing a species inside a sample, namely the representation of genomes by their U_k or Sp_k dictionaries.

This paper seeks for an interval of factor lengths that provides us with a distinctive information in terms of corresponding dictionaries (of spectral segments and of *RES*) of bacterial genomes. Moreover, spectral segments which discriminate species and genera turn out to overlap coding regions of bacterial genomes. We briefly describe the bacterial data set used in our experiments in Sect. 2. The methodology is illustrated in Fig. 1 and reported in Sect. 3 together with the software IGtools [3] employed for the analysis. Section 4 is focused on the discussion of achieved outcomes, while Sect. 5 concludes the paper with final remarks.

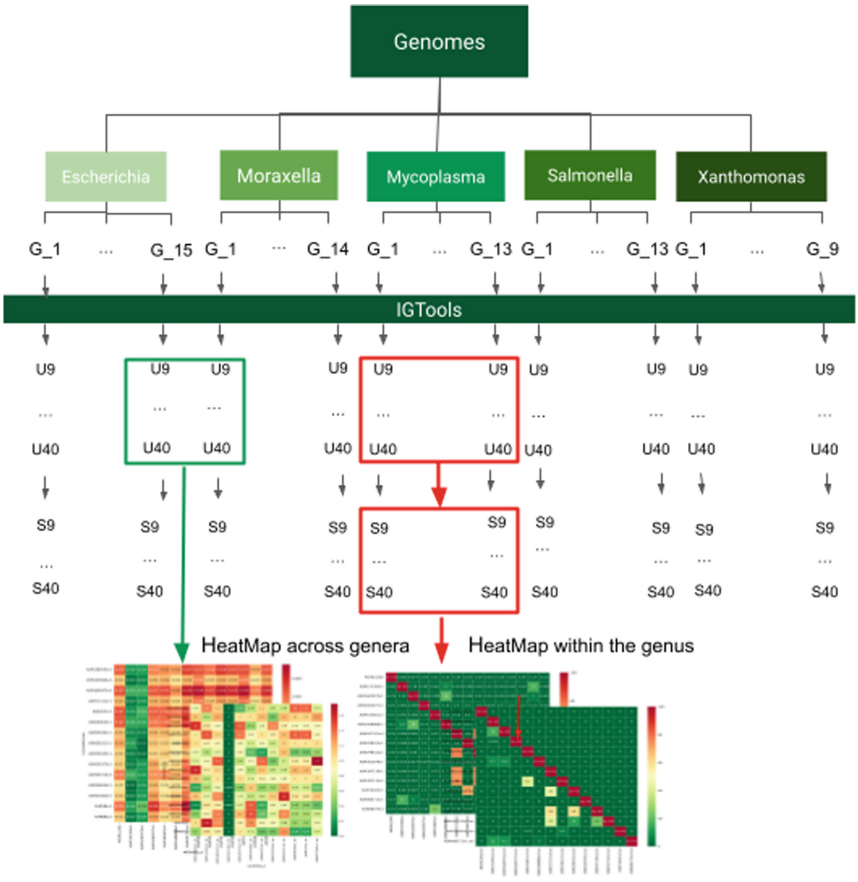


Fig. 1. A sketch of the computational analysis workflow. Dictionaries U_k and Sp_k (in this figure briefly Sk) have been computed by IGtools in the range $9 \leq k \leq 40$ for all the genomes. These are compared by similarity matrices (reported at the bottom), computed on couples of genomes either from two different genera (green square) or representing two different species within the same genus (red square). (Color figure online)

2 Dataset

As we may see in Fig. 1, genera of Escherichia, Moraxella, Mycoplasma, Salmonella and Xanthomonas have been chosen for the work, initially developed in the master thesis of the first author. Each genus collects a number of species, having in turn a few different organisms: in the figure we may distinguish 15 genomes of Escherichia (e.g., with 23 genomes of Escherichia Coli species), 14 genomes of Moraxella, 13 genomes of Mycoplasma and of Salmonella enterica, 9 genomes of Xanthomonas.

In the comparative analysis (by similarity matrices) here reported our dataset has been extended by the additional genus Shigella, with 6 species, in order to work on all genomes employed by the benchmarking *AFproject* [30] and by other alignment-free methods for genetic reconstruction, such as co-phylog [29] and Skmer [23]. Furthermore, in order to work on reference datasets present in the AFproject, to determine a significant k-range for our RES strings and spectral segments, we have extended the dataset with the following genera (having from 1 to 4 species): Citrobacter, Cronobacter, Dickerya, Edwardsiella, Enterobacter, Erwinia, Klebsiella, Pantoea, Pasteurella, Pectobacterium, Photorhabdus, Rahnella, Wigglesworthia, Xenophilus and Yersinia. All downloaded from NCBI.

3 Methods

In order to investigate bacterial genomes, by IGtools software [3,6] we computed statistical indices and specific genomic dictionaries, containing spectral segments and *RES* [4], and we visualize genome similarity by matrices reporting the normalized Jaccard index. These concepts are detailed in the following of this section.

3.1 Theoretical Background

Genomes are formalized by long strings over the alphabet $\Gamma = \{a, c, g, t\}$. In this framework, words, dictionaries and distributions are key instruments to represent genomes. Dictionary D_k collects all distinct k-mers of a string, and it may be split into two disjoint dictionaries: H_k the set of words appearing exactly once (*hapaxes* [6]) and R_k the set of words appearing more than once (*repeats*). Dictionary F_k collects forbidden k-mers, all those k-long words generated from the same alphabet that do not appear in the genome. Of course, by definition, $D_k = H_k \cup R_k$ and $\Gamma^k = D_k \cup F_k$.

A genome G is often represented by the distribution of k-mers within it. Among the others [11,17], here we recall the *k-spectrum distribution*, where each k-mer α of D_k is associated to its multiplicity $multG(\alpha)$ (i.e., the number of times it occurs in the genome). The k-spectrum of a genome G is defined as

$$Spec_k(G) = \{(\alpha, multG(\alpha)) | \alpha \in D_k\}.$$

Two k-mers of a couple (α, β) from $D_k \times D_k$ are k-concatenated if the $(k-1)$-length suffix of α equals the $(k-1)$-length prefix of β. Given $\alpha = x\gamma$ and $\beta = \gamma y$, where x and y belong to Γ, there is a right elongation of α by the symbol y, resulting in αy. If only one k-mer β elongates α along the genome, just one possible symbol y follows α and then the k-mer α is a RES (uniquely right-extendable string).

To assemble spectral segments, RES are iteratively concatenated, until more than one distinct k-mer of the spectrum competes for concatenation. In [4] some procedures were proposed to construct spectral segments, as words whose factors of length k are all RES, *each occurring at most as many times as it does on the genome G*. This constrain naturally reduces the number of different resulting spectral segments. However, it does not guarantee that they occur in the original genome.

A spectral segment is constructed by k-concatenation (that is, along with an overlap long $k-1$) of RES_k (which are collected in the dictionary U_k). It is elongated to the right until there are no more distinct RES_k capable of doing so or the multiplicity of them runs out. Hence, spectral segments are defined as *maximally uniquely elongated strings from* RES_k. All these spectral segments of variable length are collected in Sp_k.

As final remarks, we may point out that RES is a stronger concept than hapax, and that the concept of k-spectrum is behind the construction of spectral segments. Indeed, an hapax is univocally elongated over the genome since it occurs once, while RES is elongated by the same symbol in its multiple occurrences, and spectral segments are constructed consistently to the multiplicity of each k-mer in the spectrum, by means of k-concatenation.

3.2 IGtool Software

The whole procedure of extraction of spectral segments and RES from a genome G has been executed by IGtools software [3]. Bacterial genomic strings are input to the software in the form of FASTA files. It outputs three different sources of information: statistical indices, RES dictionaries (U_k) and spectral segments dictionaries (Sp_k) for a value of k in the interval defined at the beginning. Namely, it calculates for each sequence eight indices: $|D_k|$, $|H_k|$, $|U_k|$, $|U_k|/|D_k|$, $coverage(U_k, G)$, the *number* of *spectral segments*, the *maximum length*, and the *mean length* among spectral segments.

It implements the procedure of *k-segmentation* explained in [4], which computes the U_k and Sp_k dictionaries through an array that represents the positions of each k-mer in the genome. Formally, a k-mer α from $D_k(G)$ is *univocally elongated* in G if $|\{\beta \in D_k(G) : \alpha[2...k] = \beta[1...k-1]\}| = 1$. The algorithm initializes all positions in the array A of the genome size as false. A position is set as true when the k-mer starting at the position is uniquely elongated to the right in G. As last step, the algorithm searches for consecutive true values in A to construct spectral segments.

Moreover, charts are provided on the *coverage*, that is, the percentage of true positions in the array after the k-segmentation, and the *ratio* $|Uk|/|Dk|$.

The cardinality of dictionaries Sp_k with the average and maximum length of their spectral segments are computed as well (see Table 1). IGtools in comparison with other algorithms for extracting substrings performs the analysis in competitive times. In particular, this observation holds by modifying the software to extract unitigs, being the segments on which most procedures are set. IGtools is here compared with the well-established tool Bifrost [15]. IGtools uses *suffix array* SA and *longest common prefix* LCP *data structures*, both constructed in linear time, and for this reason it can be set for unitigs extraction without an increase of computational cost. On the other hand, Bifrost constructs a de Bruijn graph to extract segments and relies on *Bloom filter (BF)* [2]. Figure 2 shows an example of unitigs computation by Bifrost and IGtools. Especially for $k < 20$, IGtools is particularly efficient. Considering the range $9 \leq k \leq 40$ on different bacterial genomes, IGtools takes between one-third and one-tenth of the time of Bifrost. Those times suggest that IGtools provides unitigs, and dictionaries in general, in the timeframe proposed for spectral segment extraction without being affected by the output dictionary size. Therefore, it allows to be used on a large number of sequences and for a wide range of k in reduced time and space.

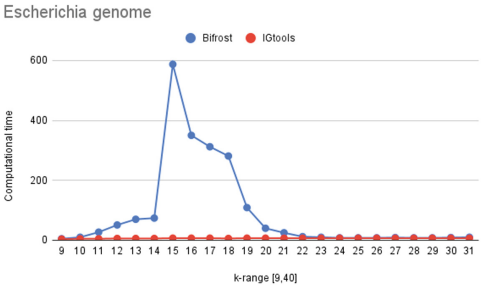


Fig. 2. Computation time on genomes from Escherichia genus. (Color figure online)

The indices computed by IGtools are displayed graphically, while the genome U_k-based and Sp_k-based similarity are retrieved by means of similarity matrices.

3.3 Graphical Tools

A similarity matrix is calculated for each value of k from 9 to 40, thus between specific dictionaries of k-mers. For each couple of genera, 31 similarity matrices exist (one for each value of k), and each matrix represents the U_k-based similarity for a defined k and any genomes pair (see green box in Fig. 1). Each matrix mxn is composed of n rows and m columns, where n and m are the number of species among different genera. If the intersection is computed to compare species inside one genus we have that $m = n$. For example, in Fig 1, matrices on the left side have dimensions 15×14 while matrices on the rigth side are 13×13 squares.

In each cell (i.e., matrix component) the *Jaccard index* is reported, as a measure of the similarity between two sets. It is defined as the intersection size divided by the union size of the dictionaries A and B: $J(A,B) = |A \cap B|/|A \cup B|$. It is a percentage, that is a value between 0 and 1. As it may be deduced from the colour legend, in the matrices red color represents an higher value, while a lower one is identified by the green.

For Sp_k-based similarity within a genus, there are still 31 matrices for each genus (one for each k). Moreover, U_k and Sp_k similarities are calculated for genomes between genera, still through the construction of similarity matrices (see Fig. 1). The similarity matrices have different numbers of rows and columns, as they represent the sequences of two different genera. Since each species has in turn different genomes, for each value of k and any couple of species we have computed 10 matrices, each representing a possible combination of species, either of two genera or within a genus, among the different bacteria.

The purpose has been to demonstrate that RES sequences are significantly present within different genomes in the same species or genus, so that shared segments can identify and characterize sequences of the groups. The analysis starts by searching the similarity between genomes through U_k dictionaries.

Species discriminants identify subtrees of a phylogenetic tree. The phylogenetic trees, constructed by CVtree software [21], employ the distance $D(A,B) = \frac{1-C(A,B)}{2}$, where $C(A,B)$ is the correlation between two species A and B, and often identify genomes from the same family as being the closest. However, these may be not the closest according to the Jaccard coefficient, which is a more demanding string similarity measure.

4 Results

Through a graphical representation of the eight indices calculated by IGtools for k ranging in the interval $9 \leq k \leq 40$, the appropriate k range is defined to extract meaningful spectral segments and specific information on Sp_k and U_k.

4.1 Significant Intervals for Values of k

We studied statistical indices for U_k and Sp_k dictionaries to find a meaningful word length interval, if any, to obtain taxa classification. Indeed, index values and charts have shown likeness over the different bacterial genomes for the k-range equal to $10 \leq k \leq 18$. This information is valid only for bacteria domain. In fact, the study of indices on genomes of eukaryotes, such as Saccharomyces, Ostreococcus and Drosophila, showed that there are no domain-specific k-ranges. Possibly there is a relation of this interval with the genome size, since bacteria in the dataset report common domain genome length (200 000 bp–10 000 000 bp).

In Table 1 we may collect some observed regularities. Even if coverage varies among the organisms in the dataset, for $k = 13$ it reaches its maximum in all cases. The ratio U_k/D_k has been computed to see how different the two dictionaries are.

Only for $k = 17$ the ratio is over 0.90, while for the other values of k the two dictionaries carry different information, according to the (negative) correlation between D_k-based similarity and U_k-based similarity trends. Specifically, Pearson's correlation index has been calculated by a vector containing the U_k-based similarity of a pair of genomes and a vector containing the D_k-based similarity, for the same pair, with respect to the k variation. The two dictionaries lead to negatively correlated similarities for the genomes under investigation, hinting that the sets of RES_k carry more specific information than the sets $D_k(G)$.

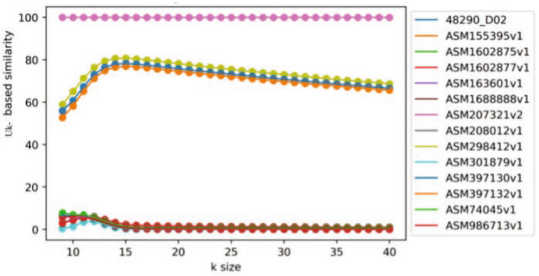
We observed that $|Sp_k|$ has a fast increase for $9 \leq k \leq 13$ and an equally rapid decrease for $14 \leq k \leq 20$. It reaches the minimum values for $k > 20$. Mean and Max represent the k from which the values of mean and maximum length of spectral segments begin to increase. Indeed, maximum and average lengths remain low and constant in the interval $9 \leq k \leq 13/15$ (numerous relatively short segments). After $k = 13$ or $k = 15$, both indices increase until $k = 40$. Correspondingly, the cardinality of the Sp dictionary decreases generally under 3000 items, and the mean length of the segments does not grow fast. As a consequence, for $k > 20$, Sp_k dictionaries contain few and long segments, which are less remarkable for analysis.

Table 1. Mean log is the average (on sequences inside the genus) logarithmical genome size. For $k = 13$ a coverage close to 95% reaches its maximum. At $k = 17$ the *ratio* $|U_k|/|D_k|$ is over 0.90. Peak is the value of k at which the number of spectral segments is maximum, while Mean/Max are the values of k at which the mean/maximum length of spectral segments begin to increase.

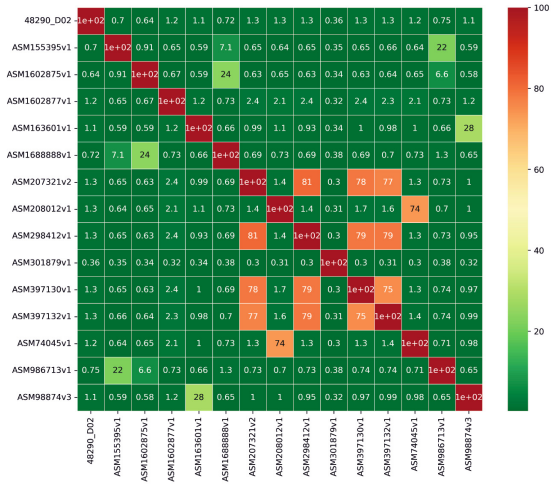
Relevant values of k length for the informational indexes					
Genera	Mean log	Coverage (> 95%)	Ratio (>0.90)	Peak	Mean/Max
Escherichia	11	13	17	13	13/15
Moraxella	10.50	13	17	12	13/15
Mycoplasma	10	13	17	13	13/15
Salmonella	11	13	17	13	13/15
Shigella	11.10	13	17	13	13/15
Xanthomonas	11	14	17/18	13	13/15
Other bact. genera	10–12	13	17/18	12/13	13/15

4.2 U_k-Based and Sp-Based Analysis

The main quest is to determine whether spectral segments and RES are biomarkers at species and genus levels and for what range of k. We search for biomarkers by means of computing U_k similarity and Sp_k similarity between genomes of the same genus and between genomes of distinct genera. These values were displayed through similarity matrices, within each genus or by pairing two different genera, along with different dictionaries, and k-values.



(a)



(b)

Fig. 3. The U_{20}-based similarity within the genus Moraxella is shown. Section (a): example of comparison between one species from the genus Moraxella with each of all the others. The legend on the right represents the species inside the genus, and each line shows the values of similarity between one species and another one along with the value of k. No pair of genomes has zero similarity and there are clusters according to Jaccard coefficients. Section (b): of the figure highlights the clustering of genomes according to similarity. Cells that are identified by pairs of genomes of the same species are those that are not dark green and have similarity values greater than 10%, often 30%, upwards to 99%. (Color figure online)

***U_k*-Based Similarity.** Computational experiments are reported where the similarity is calculated as the Jaccard index (a value between 0 and 1) on the sets U_k, for $9 \leq k \leq 40$, taken from a couple of genomes either within one single genus or from different genera. By calculating similarity within one genus, matrices have different configurations depending on the value of k.

For $k = 9$, the matrices are predominantly green, with no significant peaks in the values and no cells with values close to zero. Notice that green cells contain a value ranging from zero (when the color is darker) to about 30%. As k increases, the pattern of the matrices changes. At $k = 15$, peaks of values, in colours ranging from orange to red, emerge and the green cells assume values close to zero. Here, good similarity values reach a maximum and then slowly decrease (while k increasing). Likewise, as in the case of Escherichia Coli, similarity occurs at the strand level within the same species. Specifically, there are both orange/red and green cells, with no one color predominating over the other.

From the above observations we may hypothesize that RES dictionaries potentially function as sets of identifiers at the genus level for $k < 15$. On the other hand, for $k > 15$, some similarities have values close to zero and only similarities within specific clusters are evident. Consequently, RES dictionaries are possible identifiers at the species level.

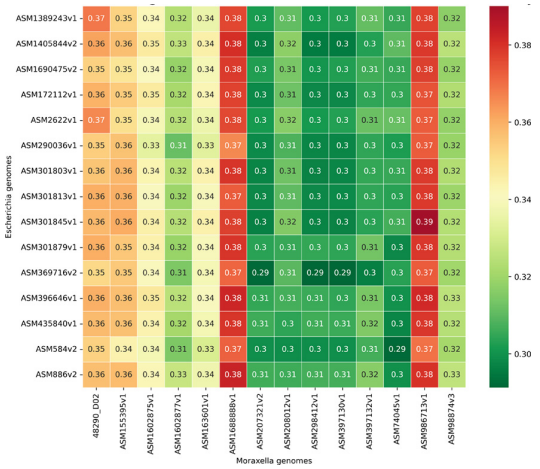


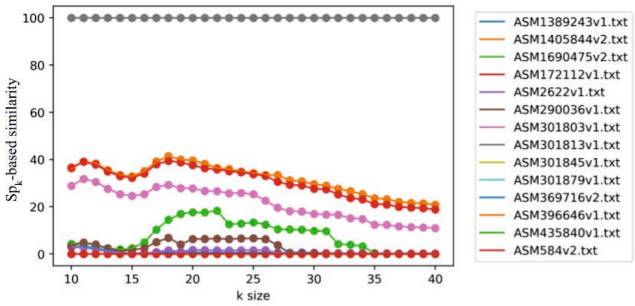
Fig. 4. U_{15}-based similarity between the genera Moraxella and Escherichia is shown. (Color figure online)

We tested if RES can act as genomic markers as well, by computing the U_k-based similarity between the genomes of 10 combinations of pairs of different genera. The coefficient never exceeds 6% (for no value of k), for no pair of genomes and for no combination of genera, as namely seen in Fig. 4. Notice the reference scale: colours vary in a range of 0%–6%. The maximum values are reached for $k \geq 15$. After $k = 15$, all matrix values tend to 0%, without distinction. Genomes of different genera do not have U_k similarities and the heatmaps are basically all homogeneous matrices of zeros.

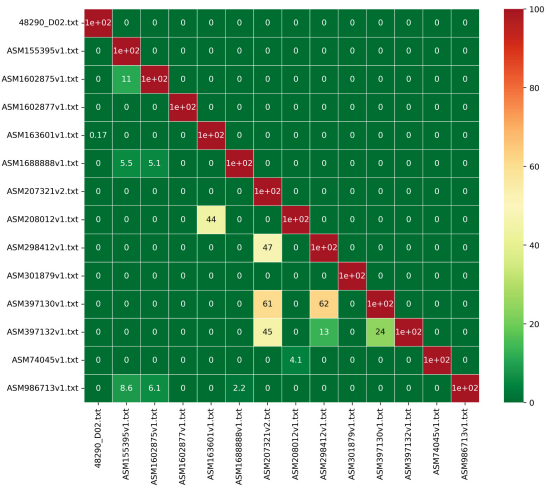
***Sp*$_k$-Based Similarity.** Computational experiments are reported where the similarity is calculated as the Jaccard index (a value between 0 and 1) on the

sets Sp_k, for $9 \leq k \leq 40$, taken from a couple of genomes either within one single genus or from different genera. The optimal k-range for spectral segments is $15 \leq k \leq 25$, since the values have a significant decrease beyond that threshold.

Concerning Sp_k-based similarity within a genus (see Fig. 5 (a) for all values of k), although its values are lower, it shows a division of genomes into clusters corresponding to the same species.



(a)



(b)

Fig. 5. The Sp_{20}-based similarity within the genus Moraxella is shown. Description of details is analogous to the text in caption of Fig. 3. (Color figure online)

As far as Sp_k similarity is concerned, the values are generally lower than those observed with U_k dictionaries, because segments are longer and dictionaries smaller. The similarity matrices have generally low values (lower than those seen

for the U_k based similarity). Orange peaks are rare and the green cells are zero from $k = 9$. The heatmaps are homogeneous and clean, and just show a difference between pairs of genomes of the same species from small k. Genomes belonging to the same species have variable coefficients which depend on the species. The range is 10–90%, and values reach a maximum before $k = 20$.

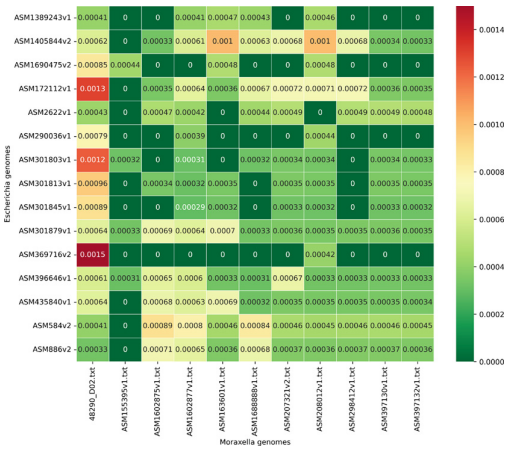


Fig. 6. The Sp_{15}-based similarity between couples of genomes taken from the genera Moraxella and Escherichia is shown. (Color figure online)

The similarity matrices with respect to spectral segments and between genera show values never exceeding 0.5%. In all bacteria organisms, for $k = 9$, there are values less than 1%, decreasing to exactly zero after a few k. For any k, or any combination of genera, the intersection of genomes has size almost zero, as shown in Fig. 6.

The observations above indicate that spectral segments are identifiers within a genus for one species. However, although between genera there is no sharing (of them), all coefficients values within a single genus matrix are not high enough to consider them identifiers of one strain (inside a species).

In the S_p similarity matrices, the cells rarely approach orange, i.e. values above 70%, but they are also surrounded by particularly low values and border on zero. The analysis through Sp_k may remove ambiguity from the intersection study, while emphasising that there is a connection with U_k. The reduced size of the Sp_k, the variability in the length of the segments, and having to deal with segments of increasing size, less prone to repetition, makes the intersection values of greater importance, and means that these results carry new information.

Comparison with the Literature of Barcodes. Barcode of life data system (BOLD) [22] is proposed as a reference for potential barcoding sequences. This database provides identifying sequences for species of the bacterial genera of our dataset. In Table 2, the sequences offered by BOLD are 666 bp long and are single strings. For each species, we average between one and four identification sequences. The only highly represented species is Escherichia Coli. Otherwise, the sequence used for barcoding is 16s RNA, which has range 300–470 bp. On the other hand, dictionaries allow a classification into taxa not related to a single sequence, but to a set of words, having length which range from an average of 2000 bp up to a maximum equal to 200000 bp.

Table 2. Characteristics of possible sets of barcode sequences.

Sequence barcodes comparison		
Source	**Segment-length**	**Set-cardinality**
BOLD	666 bp (1000 bp for Escherichia)	1–4
16S RNA	300–470 bp	1
IGTools	Max = 200000 bp (Avg 2000)	3000–100000

4.3 Sp_k-Based Coverage of Genes

It may be relevant to consider the relation between the values of k and the overlapping (or covering) of the spectral segments with the coding portions of genomes.

We say that a spectral segment covers a gene if the two genome portions coincide by at least 95%. We have checked (by means of the Boolean array used by IGtools) the overlap of Sp_k with the genes of each genome, for the interval $10 \leq k \leq 15$. We set $k \leq 15$, to avoid that U_k and D_k dictionaries overlap significantly. Figure 7 shows that in this k-range the spectral segments pass from not covering genes, for $k = 10$ and $k = 11$, to covering them all, for $k = 15$.

Therefore, gene coverage by spectral segments has a very fast growth in the k-range $11 \leq k \leq 13$. Spectral segments cover all coding regions of most genomes already for $k = 13$ or $k = 14$. Either way, for every genome, at $k = 15$ genes are all covered (by keeping in mind that for $k = 13$ the maximum coverage is usually reached, and the dictionaries U_k and D_k are not equal (see Table 1).

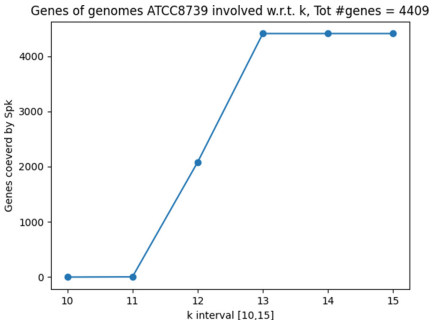


Fig. 7. The picture shows how many genes in a bacterial genome are covered by Sp_k per k-range $10 \leq k \leq 15$. The pattern observed for this specific genome is the same observed for the others.

5 Conclusion

In this paper a k-mer-based method shows to be helpful for determining bacterial species membership, and an accurate set of biomarkers was provided as an alternative to traditional singletons (sets composed by one gene). Main results of this paper may be reported as the identification of two k-parametrized dictionaries, U_k and Sp_k for $15 \leq k \leq 25$, as identifiers of bacterial species. Namely, dictionary U_k for $k < 15$ contains biomarkers at the genus level, while dictionary Sp_k for $15 \leq k \leq 25$, whose spectral segments overlap all the coding regions, discriminates one species within a genus.

The dictionary of a genome traces it back to its taxonomy and characteristics without the sequence itself being known. In fact, comparing dictionaries while following the order of the genetic tree, from leaves to parent, yields a percentage of RES and spectral segments common to the root that is almost zero. Relevantly, there is no set of RES or spectral segments common to all the genomes of a genus. This finding is particularly intriguing and may warrant by itself further investigation thorough a study on other data sets.

The above relationship between a dictionary based similarity and the membership to a phylogenetic tree suggests that spectral segments may be exploited in the phylogenetic domain [14,27,29]. In our experiments, a main difference in the two approaches is emerged with genomes of different species that are located in the same phylogenetic subtree. This observation suggests that dictionaries U_k and Sp_k are more subtle than phylogenetic trees to determine species membership, and that spectral segments distinguish even leaves of a specific subtree.

Future research could focus on a dictionary based method for phylogenetic reconstruction, as a valid alternative to unitigs employed in genome assembly [5, 18]. Indeed, spectral segments are similar but longer than unitigs, so they could be safe and complete solutions for genome assembly. Also potential barcodes could be useful for future applications, such as in the study of the metabiome, overcoming the limitation of distinguishing species in such a large sample.

Biomarker dictionaries are extracted from large amounts of genomes. The method fits with metagenomics, which was developed to handle large quantities of organisms in a less costly and less resource-intensive manner. To generate initial partial tests, we have applied IGtools to the concatenation of 9, 10 and up to 15 sequences. The concatenated sequences representative of a genus show values and peaks that are similar to the individual genomes of that genus, for $10 \leq k \leq 18$. Such a k-range is then pointed out for genomes of the bacterial kingdom, that provides specific information on spectral segments (passing from being many and short to be few and long, and covering all the genes) and RES, which discriminate at the species and genus level, respectively.

References

1. Berstel, J., Karhumäki, J.: Combinatorics on words-a tutorial. current trends in theoretical computer science. Challenge New Century **2**, 415–475 (2004)
2. Bloom, B.H.: Space/time trade-offs in hash coding with allowable errors. Commun. ACM **13**(7), 422–426 (1970)
3. Bonnici, V., Manca, V.: Infogenomics tools: A computational suite for informational analysis of genomes. J. Bioinforma Proteomics Rev. **1**, 8–14 (2015)
4. Bonnici, V., Franco, G., Manca, V.: Spectral concepts in genome informational analysis. Theoret. Comput. Sci. **894**, 23–30 (2021)
5. Cairo, M., Rizzi, R., Tomescu, A.I., Zirondelli, E.C.: Genome assembly, from practice to theory: safe, complete and linear-time. arXiv preprint arXiv:2002.10498 (2020)
6. Castellini, A., Franco, G., Manca, V.: A dictionary based informational genome analysis. BMC Genomics **13**(1), 1–14 (2012)
7. Compeau, P.E.C., Pevzner, P.A., Tesler, G.: How to apply de bruijn graphs to genome assembly. Nat. Biotechnol. **29**(11), 987–991 (2011)
8. Compeau, P.E.C., Pevzner, P.A., Tesler, G.: Why are de bruijn graphs useful for genome assembly? Nat. Biotechnol. **29**(11), 987 (2011)
9. De Luca, A.: On the combinatorics of finite words. Theoret. Comput. Sci. **218**(1), 13–39 (1999)
10. DeSalle, R., Goldstein, P.: Review and interpretation of trends in DNA barcoding. Front. Ecol. Evol. **7**, 302 (2019)
11. Franco, G.: Perspectives in computational genome analysis. In: Jonoska, N., Saito, M. (eds.) Discrete and Topological Models in Molecular Biology. NCS, pp. 3–22. Springer, Heidelberg (2014). https://doi.org/10.1007/978-3-642-40193-0_1
12. Goldstein, P.Z., DeSalle, R.: Integrating DNA barcode data and taxonomic practice: determination, discovery, and description. Bioessays **33**(2), 135–147 (2011)
13. Hao, B., Qi, J.: Prokaryote phylogeny without sequence alignment: from avoidance signature to composition distance. J. Bioinform. Comput. Biol. **2**(01), 1–19 (2004)
14. Haubold, B., Klötzl, F., Pfaffelhuber, P.: andi: fast and accurate estimation of evolutionary distances between closely related genomes. Bioinformatics **31**(8), 1169–1175 (2015)
15. Holley, G., Melsted, P.: Bifrost: highly parallel construction and indexing of colored and compacted de bruijn graphs. Genome Biol. **21**(1), 1–20 (2020)
16. Lothaire, M.: Combinatorics on Words, vol. 17. Cambridge University Press, Cambridge (1997)

17. Manca, V.: The principles of informational genomics. Theoret. Comput. Sci. **701**, 190–202 (2017)
18. Acosta, N.O., Mäkinen, V., Tomescu, A.I.: A safe and complete algorithm for metagenomic assembly. Algorithms Mol. Biol. **13**(1), 1–12 (2018)
19. Orozco-Arias, S., et al.: K-mer-based machine learning method to classify ltr-retrotransposons in plant genomes. PeerJ, **9**, e11456 (2021)
20. Orozco-Arias, S., S Piña, J., Tabares-Soto, R., Castillo-Ossa, L.F., Guyot, R., Isaza, G.: Measuring performance metrics of machine learning algorithms for detecting and classifying transposable elements. Processes **8**(6), 638 (2020)
21. Qi, J., Luo, H., Hao, B.: Cvtree: a phylogenetic tree reconstruction tool based on whole genomes. Nucleic Acids Res. **32**(suppl-2), W45–W47 (2004)
22. Ratnasingham, S., Hebert, P.D.N.: Bold: the barcode of life data system (http://www.barcodinglife.org). Mol. Ecol. Notes **7**(3), 355–364 (2007)
23. Sarmashghi, S., Bohmann, K., Gilbert, M.T.P., Bafna, V., Mirarab, S.: SKMER: assembly-free and alignment-free sample identification using genome skims. Genome Biol. **20**(1), 1–20 (2019)
24. Thomas, T., Gilbert, J., Meyer, F.: Metagenomics-a guide from sampling to data analysis. Microb. Inf. Exp. **2**(1), 1–12 (2012)
25. Tomescu, A.I., Medvedev, P.: Safe and complete contig assembly through OMNIT-IGS. J. Comput. Biol. **24**(6), 590–602 (2017)
26. Vinga, S., Almeida, J.: Alignment-free sequence comparison-a review. Bioinformatics **19**(4), 513–523 (2003)
27. Wittler, R.: Alignment and reference-free phylogenomics with colored de bruijn graphs. Algorithms Mol. Biol. **15**(1), 1–12 (2020)
28. Yen, S., Johnson, J.S.: Metagenomics: a path to understanding the gut microbiome. Mamm. Genome **32**(4), 282–296 (2021). https://doi.org/10.1007/s00335-021-09889-x
29. Yi, H., Jin, L.: Co-phylog: an assembly-free phylogenomic approach for closely related organisms. Nucleic Acids Res. **41**(7), e75–e75 (2013)
30. Zielezinski, A., et al.: Benchmarking of alignment-free sequence comparison methods. Genome Biol. **20**(1), 1–18 (2019)