ISTITUTO DI ANALISI DEI SISTEMI ED INFORMATICA
“Antonio Ruberti”
CONSIGLIO NAZIONALE DELLE RICERCHE

E. Weitschek, F. Cunial, G. Felici

DISCOVERING GENOME-WIDE K-MER COMPOSITIONAL RULES USING LOGIC FORMULAS

R. 13-19 2013

Emanuel Weitschek – Istituto di Analisi dei Sistemi ed Informatica del CNR, viale Manzoni 30 - 00185 Roma, Italy. Email: `emanuel.weitschek@iasi.cnr.it`.

Fabio Cunial – Department of Computer Science, University of Helsinki, P.O. 68 (Gustaf Hllstrmin katu 2b), FI-00014, Finland. Email: `cunial@cs.helsinki.fi`.

Giovanni Felici – Istituto di Analisi dei Sistemi ed Informatica del CNR, viale Manzoni 30 - 00185 Roma, Italy. Email: `felici@iasi.cnr.it`.

ISSN: 1128–3378

Collana dei Rapporti dell'Istituto di Analisi dei Sistemi ed Informatica "Antonio Ruberti", CNR

viale Manzoni 30, 00185 ROMA, Italy

tel. ++39-06-77161
fax ++39-06-7716461
email: iasi@iasi.cnr.it
URL: http://www.iasi.cnr.it

Abstract

The increasing availability of biological sequences from massive experiments lead to the growth of the field of sequence analysis. In this field the similarity of sequences is used to prove related biological functions or detect common organisms. Analysis algorithms include methods and techniques from statistics and computer science. Most current sequence analysis methods are based on alignment, i.e. align areas of the sequences sharing common properties. These algorithms are computational demanding and the complexity is exponential in the length of the sequences, therefore heuristics have been proposed that solve the sequence alignment problem. Alternative methods for sequences classification rely on string matching, pattern recognition and alignment free techniques, that can be also combined with supervised and unsupervised machine learning algorithms. In alignment free methods the similarity of two sequences is assessed based only on the dictionary of subsequences that appear in the strings, irrespective of their relative position. The subsequences can be represented in a feature vector and then treated in a mathematical space and eventually combined with machine learning algorithms, e.g., logic data mining. In this work a method for classifying biological sequences is proposed. The method is based on an alignment free feature vector representation of biological sequences in combination with logic data mining algorithms. The method classifies biological sequences without the strict requirement of alignments or of overlapping gene regions. The method is tested on bacterial whole genomes with promising results and classification accuracy. Finally, the strengths of the method are highlighted: promising classification results on bacterial sequences, no necessity to align them and identification of common subsequences (kmers) for each class (taxon) present in the data set.

1. Introduction

Massive sequencing experiments lead to an increasing availability of biological sequences. The field of sequence analysis develops and applies new and efficient methods to accomplish the task of detecting similarities among them in order to distinguish different organisms or discover functional relatedness. Sequence analysis is approached with mathematical, statistical, and computer science techniques. Different data mining and machine learning methods are often adopted to get an insight of the analyzed sequences.
Most sequence analysis methods rely on alignment: they align portions of the sequences – from single genes to whole genomes – that have similar properties, like a subsequence of similar nucleotide assignments. A score is given to the alignment, based on the number of contiguous common nucleotides assignments. Alignment-based sequence analysis methods can be divided into two classes: exact methods and heuristics.
Exact methods, which are commonly based on dynamic programming, have an high computational cost, exponential in the length of the input sequences. Widespread methods are Smith-Waterman [25] and Needleman-Wunsch [23]. The former is a local sequence alignment algorithm, which determines similar regions between two sequences by comparing segments of all possible lengths. The latter instead aligns the entire sequences.
Heuristic methods reduce the complexity of sequence alignment by providing suboptimal solutions. The most popular algorithms are FASTA [24] and BLAST [2]. Specific algorithms that compare multiple sequences at once have been designed and developed, and are called multiple sequences alignment algorithms, e.g., ClustalW [33], Muscle [8], Mafft [16], and Motalign [22].
The main problem of alignment-based methods is the high computational cost, in particular when dealing with a large number of sequences [35]. Another issue of alignment-based methods is the impossibility to model recombinations and shuffling of the sequences. Moreover, in many cases sequences are not naturally alignable, for example when dealing with non-coding sequences, or very hard to align, for example when analyzing entire genomes of distantly-related organisms.
To solve these limitations, alignment-free methods for the analysis of biological sequences have been introduced in the last decade. These algorithms are particularly suited for sequences that are not naturally alignable or for which there is a lot of ambiguity (e.g., non-coding regions or GC-rich regions).
The success of modern alignment-free algorithms rests on extensive information on the substring composition of genomes and on codon-usage biases, cumulated over approximately fifty years, with particular emphasis on prokaryotes: from the first studies of GC content [14], to the first detection of biases in the composition of pairs and quadruples of adjacent nucleotides [3, 14, 15], to the discovery of species-specific frequencies of 4-mers and 8-mers preserved in DNA fragments ranging from 40 kilobases to 400 bases [31, 40], to more recent, unsupervised classifications [29] and more complex protein motifs [28].
Alignment-free methods for sequence comparison can be classified in two main groups: methods based on sequence compression and methods that rely on subsequences (oligomers) frequencies [35].
The methods based on sequence compression (e.g. Kolmogorov complexity [20]) aim to approximate the shortest description of a sequence, and are based on Kolomogorov complexity or Universal Sequence Maps [1].
The methods that rely on subsequences (oligomers) frequencies are based on a vector representation of a sequence. A sequence is represented as a feature vector where the frequencies of its substrings are stored [1, 17, 39, 18]. A substring must be typically of a specific length k and is called k-mer. The sequences are so in a mathematical space that is tractable with several tools and algorithms.
The main advantages of alignment-free methods are their speed and scalability. In alignment-free methods there is no need to explicitly model overlapping genes, recombinations and shuffling are automatically detected, and every kind of sequence (coding, non coding, segments) can be analyzed.
In this work an alignment-free technique method on oligomers frequencies is applied to whole bacterial genomes and the feature vectors are given as input to different supervised machine learning algorithms in order to distinguish the different taxa present in the data set.

2. Methods

In this section the alignment free feature vector representation technique of biological sequences is described in detail, and the combination with logic data mining algorithms – a class of supervised machine learning algorithms based on rules – is proposed as a method for classifying specimens to their taxa. We call this approach *Logic Alignment Free* (LAF). In the following, we describe the alignment-free feature vector representation technique, the logic data mining algorithms, and the combination of these two techniques.

2.1. Alignment-free feature vector representation

Alignment-free methods based on oligomer frequency rely on the computation of the substrings frequencies of a given length k in the original sequences, called k-mers. Most alignment-free algorithms use a feature vector representation [1, 17, 39, 18], which consists in a representation of the k-mer frequencies of a sequence in a frequency vector, where each component of the vector is associated with the frequency of a particular k-mer [35], obtained by considering a sliding window of length k.
More formally [35], let S be a sequence of n characters over an alphabet Σ, e.g. $\Sigma = A, C, G, T$, and let $k \in [1, n]$. If K is a generic substring of S of length k, K is called k-mer. Let the set $V = K_1, K_2, \ldots, K_m$ be all possible k-mers over Σ, V has size $m = |\Sigma|^k$. The k-mers are computed by counting the occurrences of the substrings in S with a sliding window of length k over S, starting at position 1 and ending at position $n - k + 1$. A vector F contains for each k-mer the corresponding counts $F = c_1, c_2, \ldots, c_m$. The frequencies are then computed accordingly and stored in a vector $F' = f_1, f_2, \ldots, f_m$, for a k-mer K_i, the frequency is defined as $f_i = \frac{c_i}{n-k+1}$. If we consider $k = 3$ and the sequence $AACGTAAC$ as example the feature and frequencies vectors are depicted in table 1. A numeric coding of the sequences is obtained, which enables the analysis with statistical, mathematical, and computer science techniques. A widely used metric for comparing the sequences is the computation of distance measures among their frequency vectors. In current approaches the sequences are compared according to their vector representations by computing a distance measure among them, a simple and effective distance measure is the Euclidean distance, another very used distance measure is the $d2$ distance [35]. Alternatively, the vectors can be given as input to supervised (classification) or unsupervised (clustering) machine learning algorithms. In [39] the authors combine feature vector representation with supervised machine learning methods, like Support Vector Machines [34], for classifying biological and generic sequences. In this work the feature vector representation is used as input for supervised machine learning algorithms, in particular rule-based classifiers [19, 37].

Table 1: Feature vector and frequencies vector

FeV		FrV	
AAC	2	AAC	0.333
ACG	1	ACG	0.166
CGT	1	CGT	0.166
GTA	1	GTA	0.166
TAA	1	TAA	0.166

2.2. Classification and logic data mining

Classification is the action of assigning an unknown object to a predefined class after examining its features [7]. An example from biology is the specimen-to-species assignment problem. Classification is called also supervised learning: unknown objects are assigned to a class using a model that is derived from objects within a known class, that compose the training set.
A classification method is a systematic approach for building a classification model from the training data. In Logic Data Mining, i.e. classification with logic formulas or rule-based classification, the classifier uses logic propositional formulas in disjunctive or conjunctive normal form ("if-then rules") for classifying the

given records. Examples of methods for computing separating formulas are RIPPER [6], LSQUARE [9], DMB [5, 38], RIDOR [11] and PART [10]. The main advantage of classification formulas is the human readable model, which is often very compact.
In the following different rule-based classification methods, that are suitable for sequence analysis, are described in detail. It is worth noting that, although they use very different rule extraction approaches, they all obtain very good performances for classifying sequences when combined with alignment-free techniques.

2.2.1. DMB

DMB (Data Mining Big) [5, 38] is a logic data mining method, that relies on several models and algorithms based on combinatorial optimization. The main characteristic of the DMB system is that it extracts knowledge in form of logic classification formulas. DMB is based on the following main computational steps:

- discretization;
- feature selection;
- formula extraction.

The discretization step amounts to determining a set of cutpoints over the range of values that each variable may assume, and thus define a number of intervals over which the original variable may be considered discrete [36].
The main aim of feature selection is to identify and remove irrelevant and redundant information from the data set [4]. For the formula-extraction step, DMB uses a logic resolution method based on a minsat problem formulation described in [9].

2.2.2. RIPPER

Repeated Incremental Pruning to Produce Error Reduction (RIPPER) [6] is a direct propositional rule learner. A direct rule extraction method computes the classification model directly by looking at the data. The method consists in two steps:

- construction of an initial set of rules;
- optimization of the rule set.

RIPPER orders the classes by increasing number of elements size and generates the classification rules by considering the smallest class versus the others grouped in one. So the problem becomes a two-class classification problem. The two steps are iterated for all classes except for the most abundant one, for which a default (empty) rule is generated: if no element is recognized by a rule, then it is assigned to the abundant class. The method provides as output a collection of logic rules for each class present in the dataset.

2.2.3. RIDOR

RIDOR (RIpple-DOwn Rule) [11] is also a direct rule extraction method. However, the procedure is structured in a different way: RIDOR starts with the generation of a default rule for the most numerous class, e.g., “All sequences are E.Coli”, and then computes exceptions, e.g., “except if $freq(ACGT) > 143.34$ then the sequences are from the organism S.Aureus”, with the smallest classification error rate on the training set. Exceptions are rules that model the other classes. Exceptions are optimized and corrected according to a purity measure.

2.2.4. PART

PART [10] is an indirect rule extraction method. PART uses the C4.5 decision tree classifier [27] for generating a classification model. For a given number of iterations, it generates a pruned decision tree and selects the best one according to a verification set. The paths to the leaves of the best performing tree represent the classification rules.

2.3. Logic Alignment Free (LAF)

Logic data mining, i.e., rule-based classification for sequence analysis, was firstly adopted by Bertolazzi et al. in [5] and in [38]. This method is based on a prior alignment of the sequences and can process only sequences of the same gene regions. In fact, it extracts the characteristic positions and their nucleotide assignments that are able to identify a species and distinguish it from the others that are present in the data set. The output of the method are classification rules in the form of “if pos192=A of gene 16S then the sequence belongs to E.Coli”.

As discussed previously, computing an alignment has some drawbacks and it is not always possible to align biological sequences (e.g., non coding regions). To overcome the limits of an alignment-based approach a technique, that does not take positions into account, is necessary to perform an effective analysis for sequences that are not easily alignable or for which there is a lot of ambiguity.

This approach, that combines alignment-free methods and logic data mining, is called Logic Alignment Free (LAF) and is based on taking as input the frequency vector representation of the sequences for performing their classification. A new classification approach for biological sequences is introduced: the combination of alignment-free k-mer frequency counts and logic data mining allows the analysis of biological sequences without the strict requirement of an alignment or of an overlapping DNA gene region. This leads to the possibility of performing classification of non coding DNA, which is not alignable, and of whole genomes, which are very hard to align, as the problem of whole genome alignment is computationally hard. The performed experiments show that this technique is very promising in distinguishing and classifying diverse organisms whole genomes at different levels of the phylogenetic tree.

The LAF analysis is based on the supervised machine learning paradigm: every sequence has a class assigned a priori, for example a bacterial taxon, and this collection of sequences composes the training set. Considering each genome g in the data set, the following steps are performed by LAF:

1. The reverse complement g' of the genome g is calculated.
2. The genome g is concatenated with its reverese complement g', obtaining the sequence $G = g + g'$.
3. The k-mer counts are computed on G with $k = 3 \dots 6$ and stored in the feature vector F.
4. The frequency vector F' is obtained from F.
5. The frequency vectors are collected in a matrix, whose rows correspond to k-mers, and whose columns correspond to sequences (see table 2 for an example).
6. A discretization of the frequencies is performed using the MDL procedure [12] or the supervised approach of [36].
7. The data matrix is then processed by a rule-based classifier, like DMB, RIPPER [6], RIDOR [11], and PART [10].

For extracting the k-mer counts from the original and reverse-complemented sequences, the software Jellyfish [21] is used. For classifying, the DMB and Weka [12] software are used. All the scripts for processing the sequences (download, reverse complement and frequency calculation) are available upon request. The flow chart of the LAF method is summarized in Figure 1.

Table 2: Data matrix

	Seq1	**Seq2**	**...**	**SeqM**	**SeqN**
	E.Coli	*E.Coli*	...	*S.Aureus*	*S.Aureus*
AAA	0.46	0.26	...	0.24	0.26
AAC	0.12	0.16	...	0.23	0.24
AAG	0.13	0.23	...	0.23	0.22
...	...	...	...	...	...

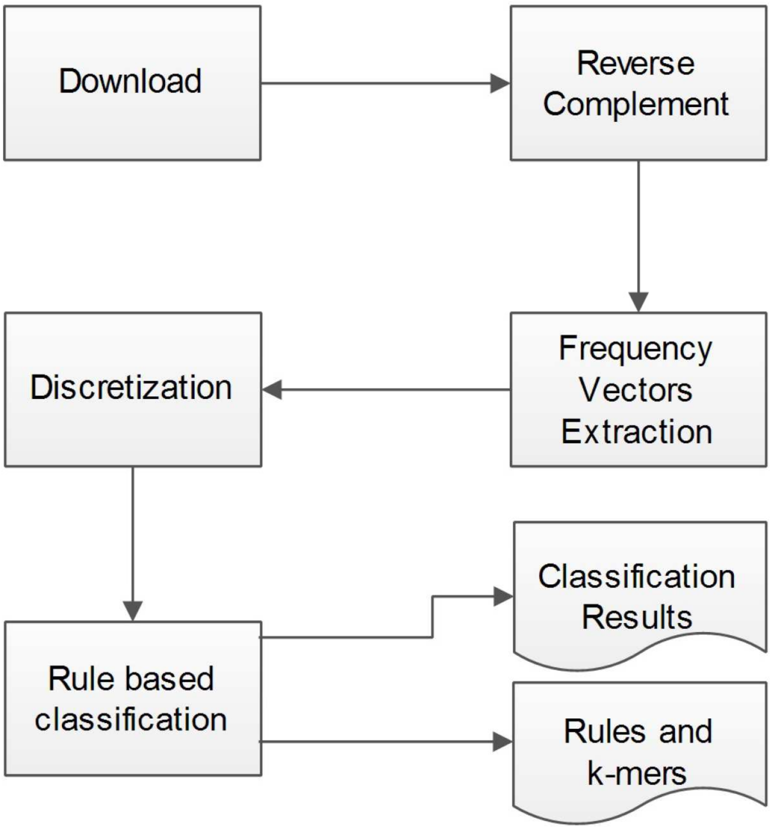

Figure 1: Flow chart of the LAF method.

3. Results and discussion

In this section the proposed LAF technique (alignment free k-mer counts frequency analysis combined with logic data mining) is applied for classifying bacteria at multiple levels of the phylogenetic tree (phylum, class, order, genus, species) by analyzing their whole genome. Whole genomes are very difficult to be aligned, because of their length (more than 2 million base pairs on average in bacteria, bilions in eukaryotes). The problem of multiple alignment is computationally demanding and grows exponentially in the size of the input (length and number of sequences) [13]. An alignment-free technique simplifies considerably the analysis process, allowing an increase of the speed and an effective biological classification of the sequences.

1964 bacterial genomes were downloaded from the NCBI database (`ftp.ncbi.nih.gov`). From these sequences we filtered out the under-represented species with less than nine sequenced specimens, to simplify the training phase of supervised machine learning algorithms. The filtering step resulted in 413 sequences belonging to 25 species, 21 genera, 14 orders, 9 classes, and 6 phyla. The proposed LAF method was applied to the filtered bacterial sequences, obtaining promising classification results. The results with

$k = 4$ are reported as example in table 3. We just show the results for $k = 4$ because this setting achieves the best performance inside the interval $k \in [3, 6]$ that we probed in our experiments. Good performance with $k = 4$ has been reported also in a number of previous studies on DNA [30, 26, 32], suggesting that this value achieves a good balance between the length and the frequency of substrings. An example of logic classification rule is the following (the frequencies are multiplied by 10^5 to ease the reading):
if $5558.475 \leq freq(\texttt{ACTA}) < 6248.76$ then the sample is “Campylobacter jejuni”.
Part of the species model computed by the DMB software is reported in figure 2. It is worth noting that, for instance, three and two bacterial species are distinguished by the same k-mer with different frequency values. Additionally, for 25 different species, only 20 k-mers of length 4 are sufficient to distinguish them. The logic classification formulas are all very compact, i.e., composed by only two literals and one `AND` conjunction, except in one case, where an `OR` disjunction is present. The results show that the method is able to correctly classify the bacterial genomes in their taxa. The best performance is obtained at the phylum level, with an average correct classification rate of 98%. Phylum is the highest level in the phylogenetic tree and it is clearly easier to classify precisely an organism at this level, in lower levels the distinctive characteristics of every organism are shuffled together, leading to less precise classification models. Nevertheless, also at lower levels of the phylogenetic tree the classification rates are effective to distinguish the different taxonomic units.

Table 3: Bacteria classification rates (10-fold cross validation)

Level	**JRip**	**Ridor**	**Part**	**DMB**	**Average**
Species	93.21	97.33	96.36	**97.61**	96.14
Genus	93.98	**98.79**	97.1	98.44	97.08
Order	94.45	**99.27**	98.79	98.58	97.77
Class	96.50	97.81	**98.79**	98.06	97.79
Phylum	96.88	**98.78**	98.07	98.53	98.06
Average	95.00	**98.40**	97.82	98.24	97.37

Bifidobacterium animalis: $762.28 \leq TCCA < 819.04$ AND $469.35 \leq TGCA < 515.63$
Corynebacterium diphtheriae: $819.04 \leq TCCA < 875.80$ AND $423.07 \leq TGCA < 469.35$
Corynebacterium pseudotuberculosis: $875.80 \leq TCCA < 932.56$ AND $423.07 \leq TGCA < 469.35$
Escherichia coli: $710.86 \leq GCAC < 860.58$ AND $415.84 \leq GCTA < 525.98$
Listeria monocytogenes: $411.43 \leq GCAC < 561.15$ AND $305.55 \leq GGAC < 393.10$

Figure 2: Part of DMB species level model

4. Conclusions

In this work an approach for classifying biological sequences (LAF), has been proposed. The method is based on an alignment-free feature vector representation of biological sequences in combination with logic data mining algorithms. The method classifies biological sequences without the strict requirement of alignments or of overlapping gene regions.
The tests on bacterial whole genomes are promising, showing accurate classification results. The study sheds light on the power of LAF: promising classification results on bacterial sequences, no need to align them, and identification of common subsequences (k-mers) in each class.
The perspectives of the work are to preprocess the sequences by filtering out high-frequency regions, to apply statistical corrections for improving the reliability of the method, and to study the dependence between classification quality and length of the k-mers. Additionally, we plan to scale the method to include more organisms, like viral sequences, and to perform an accurate analysis of the k-mers selected for each species.

Acknowledgment

The authors would like to thank Daniele Santoni for precious advice on bacterial genomes. This work is partially supported by the FLAGSHIP "InterOmics" project (PB.P05) funded by the Italian MIUR and CNR institutions, by the PRIN GenData 2020 project, and by Academy of Finland under grant 250345 (Center of Excellence in Cancer Genetics Research).

10.

References

[1] J. S. Almeida and S. Vinga, "Universal sequence map (usm) of arbitrary discrete sequences.," *BMC Bioinformatics*, vol. 3, p. 6, 2002.

[2] S. F. Altschul, T. L. Madden, A. A. Schffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman, "Gapped blast and psi-blast: a new generation of protein database search programs.," *Nucleic Acids Res*, vol. 25, pp. 3389–3402, Sep 1997.

[3] S. D. Bentley and J. Parkhill, "Comparative genomic structure of prokaryotes," *Annu. Rev. Genet.*, vol. 38, pp. 771–791, 2004.

[4] P. Bertolazzi, G. Felici, and G. Lancia, "Application of feature selection and classification to computational molecular biology," in *Biological Data Mining* (S. L. e. J.K. Chen, ed.), pp. 257–294, Chapman & Hall, 2010.

[5] P. Bertolazzi, G. Felici, and E. Weitschek, "Learning to classify species with barcodes," *BMC Bioinformatics*, vol. 10, no. S-14, p. 7, 2009.

[6] W. W. Cohen, "Fast effective rule induction," in *In Proceedings of the Twelfth International Conference on Machine Learning*, pp. 115–123, Morgan Kaufmann, 1995.

[7] S. Dulli, S. Furini, and P. E., *Data Mining*. Springer, 2009.

[8] R. C. Edgar, "Muscle: multiple sequence alignment with high accuracy and high throughput," *Nucleic acids research*, vol. 32, no. 5, pp. 1792–1797, 2004.

[9] G. Felici and K. Truemper, "A minsat approach for learning in logic domains," *INFORMS Journal on Computing*, vol. 13, no. 3, pp. 1–17, 2002.

[10] E. Frank and I. H. Witten, "Generating accurate rule sets without global optimization," in *In: Proc. of the 15th Int. Conference on Machine Learning*, Morgan Kaufmann, 1998.

[11] B. R. Gaines and P. Compton, "Induction of ripple-down rules applied to modeling large databases," *Journal of Intelligent Information Systems*, 1995.

[12] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The weka data mining software: an update," *SIGKDD Explor. Newsl.*, vol. 11, pp. 10–18, Nov. 2009.

[13] S. Hosni, A. Mokaddem, and M. Elloumi, "A new progressive multiple sequence alignment algorithm," in *Database and Expert Systems Applications (DEXA), BIOKDD 2012*, pp. 195 –198, sept. 2012.

[14] S. Karlin and C. Burge, "Dinucleotide relative abundance extremes: a genomic signature," *Trends in genetics*, vol. 11, no. 7, pp. 283–290, 1995.

[15] S. Karlin and J. Mrázek, "Compositional differences within and between eukaryotic genomes," *Proceedings of the National Academy of Sciences*, vol. 94, no. 19, pp. 10227–10232, 1997.

[16] K. Katoh, K. Misawa, K.-i. Kuma, and T. Miyata, "Mafft: a novel method for rapid multiple sequence alignment based on fast fourier transform," *Nucleic acids research*, vol. 30, no. 14, pp. 3059–3066, 2002.

[17] D. Kudenko and H. Hirsh, "Feature generation for sequence categorization," in *AAAI/IAAI*, pp. 733–738, 1998.

[18] P. Kuksa and V. Pavlovic, "Efficient alignment-free dna barcode analytics.," *BMC Bioinformatics*, vol. 10 Suppl 14, p. S9, 2009.

[19] T. Lehr, J. Yuan, D. Zeumer, S. Jayadev, and M. Ritchie, "Rule based classifier for the analysis of gene-gene and gene-environment interactions in genetic association studies," *BioData Mining*, vol. 4, no. 1, p. 4, 2011.

[20] M. Li and P. M. Vitnyi, *An Introduction to Kolmogorov Complexity and Its Applications*. Springer Publishing Company, Incorporated, 3 ed., 2008.

[21] G. Marais and C. Kingsford, "A fast, lock-free approach for efficient parallel counting of occurrences of k-mers.," *Bioinformatics*, vol. 27, pp. 764–770, Mar 2011.

[22] A. Mokaddem and M. Elloumi, "Motalign: A multiple sequence alignment algorithm based on a new distance and a new score function," in *DEXA Workshops*, pp. 81–84, 2013.

[23] S. B. Needleman and C. D. Wunsch, "A general method applicable to the search for similarities in the amino acid sequence of two proteins.," *Journal of molecular biology*, vol. 48, pp. 443–453, Mar. 1970.

[24] W. R. Pearson, "Rapid and sensitive sequence comparison with fastp and fasta.," *Methods Enzymol*, vol. 183, pp. 63–98, 1990.

[25] W. R. Pearson, "Searching protein sequence libraries: comparison of the sensitivity and selectivity of the smith-waterman and fasta algorithms.," *Genomics*, vol. 11, pp. 635–650, Nov 1991.

[26] D. T. Pride, R. J. Meinersmann, T. M. Wassenaar, and M. J. Blaser, "Evolutionary implications of microbial genome tetranucleotide frequency biases," *Genome Research*, vol. 13, pp. 145–158, 2003.

[27] J. R. Quinlan, *C4.5: Programs for Machine Learning (Morgan Kaufmann Series in Machine Learning)*. Morgan Kaufmann, 1 ed., January 1993.

[28] I. Rigoutsos, A. Floratos, C. Ouzounis, Y. Gao, and L. Parida, "Dictionary building via unsupervised hierarchical motif discovery in the sequence space of natural proteins," *Proteins*, vol. 37, no. 2, pp. 264–277, 1999.

[29] M. Takahashi, K. Kryukov, and N. Saitou, "Estimation of bacterial species phylogeny through oligonucleotide frequency distances," *Genomics*, vol. 93, no. 6, pp. 525–533, 2009.

[30] H. Teeling, A. Meyerdiekers, M. Bauer, and F. O. Glckner, "Application of tetranucleotide frequencies for the assignment of genomic fragments," *Environmental Microbiology*, vol. 6, no. 9, pp. 938–947, 2004.

[31] H. Teeling, A. Meyerdierks, M. Bauer, R. Amann, and F. O. Glöckner, "Application of tetranucleotide frequencies for the assignment of genomic fragments," *Environmental Microbiology*, vol. 6, no. 9, pp. 938–947, 2004.

[32] H. Teeling, J. Waldmann, T. Lombardot, M. Bauer, and F. O. Glckner, "Tetra: a web-service and a stand-alone program for the analysis and comparison of tetranucleotide usage patterns in dna sequences," *BMC Bioinformatics*, vol. 5, p. 163, 2004.

[33] J. D. Thompson, T. Gibson, D. G. Higgins, *et al.*, "Multiple sequence alignment using clustalw and clustalx," *Current protocols in bioinformatics*, pp. 2–3, 2002.

[34] V. N. Vapnik, *Statistical learning theory*. Wiley, 1998.

[35] S. Vinga and J. Almeida, "Alignment-free sequence comparison-a review.," *Bioinformatics*, vol. 19, pp. 513–523, Mar 2003.

[36] E. Weitschek, G. Felici, and P. Bertolazzi, "Mala: A microarray clustering and classification software," in *23rd International Workshop on DEXA, BIOKDD*, 2012.

12.

[37] E. Weitschek, G. Fiscon, and G. Felici, "Supervised dna barcodes species classification: analysis, comparisons and results," *BioData Mining*, vol. 7, p. 4, 2014.

[38] E. Weitschek, A. Lo Presti, G. Drovandi, G. Felici, M. Ciccozzi, M. Ciotti, and P. Bertolazzi, "Human polyomaviruses identification by logic mining techniques," *BMC Virology Journal*, vol. 58, no. 9, 2012.

[39] Z. Xing, J. Pei, and E. Keogh, "A brief survey on sequence classification," *ACM SIGKDD Explorations Newsletter*, vol. 12, no. 1, pp. 40–48, 2010.

[40] F. Zhou, V. Olman, and Y. Xu, "Barcodes for genomes and applications," *BMC bioinformatics*, vol. 9, no. 1, p. 546, 2008.