Genomics 104 (2014) 79–86

Contents lists available at ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

Classification of selectively constrained DNA elements using feature vectors and rule-based classifiers

Dimitris Polychronopoulos [a,b,1], Emanuel Weitschek [c,d,*,1], Slavica Dimitrieva [e,f], Philipp Bucher [e,f], Giovanni Felici [d], Yannis Almirantis [a]

[a] Institute of Biosciences and Applications, National Center for Scientific Research "Demokritos", 15310 Athens, Greece
[b] Department of Biochemistry and Molecular Biology, Faculty of Biology, National and Kapodistrian University of Athens, 15701 Athens, Greece
[c] Department of Engineering, Roma Tre University, Via della Vasca Navale 79, 00146 Rome, Italy
[d] Institute of Systems Analysis and Computer Science "Antonio Ruberti", National Research Council, Viale Manzoni 30, 00185 Rome, Italy
[e] Swiss Institute for Experimental Cancer Research (ISREC), School of Life Sciences, Swiss Federal Institute of Technology (EPFL), Lausanne, Switzerland
[f] Swiss Institute of Bioinformatics (SIB), CH-1015 Lausanne, Switzerland

ARTICLE INFO

Article history:
Received 9 June 2014
Accepted 15 July 2014
Available online 22 July 2014

Keywords:
Conserved noncoding elements (CNEs)
Ultraconserved elements (UCEs)
Classification
Rule-based classifiers
Feature vector representation
Alignment-free classification

ABSTRACT

Scarce work has been done in the analysis of the composition of conserved non-coding elements (CNEs) that are identified by comparisons of two or more genomes and are found to exist in all metazoan genomes.
Here we present the analysis of CNEs with a methodology that takes into account word occurrence at various lengths scales in the form of feature vector representation and rule based classifiers. We implement our approach on both protein-coding exons and CNEs, originating from human, insect (*Drosophila melanogaster*) and worm (*Caenorhabditis elegans*) genomes, that are either identified in the present study or obtained from the literature. Alignment free feature vector representation of sequences combined with rule-based classification methods leads to successful classification of the different CNEs classes. Biologically meaningful results are derived by comparison with the genomic signatures approach, and classification rates for a variety of functional elements of the genomes along with surrogates are presented.

1. Introduction

1.1. Conserved non-coding elements in metazoan genomes

It is generally believed that sequence conservation across genomes is a key indication for functional relevance. As a consequence, since the early days of comparative genomics several groups have focused on the detection of genomic sequences highly similar between two or more organisms. Bejerano et al. described a set of 481 sequences that display 100% identity between the human, mouse and rat genomes, so-called ultraconserved elements (UCEs) [1]. Surprisingly, the majority of these sequences do not encode for proteins. The UCEs along with several other classes of elements called UltraConserved Non-coding Elements (UCNEs) (>95% identity over > 200 bp between human and chicken) [2] and Long Conserved Non-Coding Sequences (LCNS) (>95% identity over > 500 bp between human and mouse) [3] are among the most conserved sequences described in the biomedical literature with mostly unknown functionality overall. These sequences represent only the tip of the iceberg of a much larger class of conserved non-coding elements (CNEs), whose mean level of conservation frequently exceeds that of protein-coding sequences [4,5].

Conserved non-coding elements can be found in the literature under various definitions, depending on the minimal sequence similarity between species under consideration and the similarity score achieved over a minimal sequence length [6]. Throughout this article, we use the term CNE(s) to describe all such elements despite their specific characterization as UCEs, UCNEs, LCNSs, etc. in the related literature. We use the specific name only when we refer to the corresponding class of elements.

Due to the level of conservation of those elements, their non-random co-localization around developmental genes [7] and their widespread distribution throughout genomes [5,8], many plausible functional roles have been attributed to CNEs (reviewed in [5]). It has been shown, amongst others, that many of the identified CNEs function as regulatory elements that are important in the early stages of vertebrate development and brain formation [9–11].

Although early studies have primarily focused on CNEs in vertebrate genomes, CNEs with similar properties have been also identified in

[*] Corresponding author at: Department of Engineering, Roma Tre University, Via della Vasca Navale 79, 00146 Rome, Italy.
E-mail addresses: dpolychr@bio.demokritos.gr (D. Polychronopoulos), emanuel@dia.uniroma3.it (E. Weitschek), slavica.dimitrieva@epfl.ch (S. Dimitrieva), philipp.bucher@epfl.ch (P. Bucher), giovanni.felici@iasi.cnr.it (G. Felici), yalmir@bio.demokritos.gr (Y. Almirantis).
[1] Joint first authors.

http://dx.doi.org/10.1016/j.ygeno.2014.07.004
0888-7543/© 2014 Elsevier Inc. All rights reserved.

invertebrate genomes (insects, worms) [12,13] and in plants [14]. This has suggested that they are of very ancient origin and has provoked speculations about the emergence of those elements [5].

When compared to their surrounding genomic environment, CNEs are generally AT-rich, often containing runs of identical nucleotides (homopurines or homopyrimidines, unpublished results). Walter et al. analyzed the base composition of human and fugu CNEs at single nucleotide level and found that they are A + T rich, much more so than the region they reside in, in contrast to their flanking region just outside their boundaries, which exhibits a marked drop in A + T content that forms a unique pattern [15]. But surprisingly, very little is known about the sequence intrinsic properties of CNEs that are important for their function and that segregate them from other classes of functional elements. It is therefore of great interest to further investigate the compositional preferences of constrained regions in greater detail.

1.2. Alignment-based and alignment-free methods for sequence analysis

Similarity of sequences is generally used as an indication of a corresponding similarity in their functionality. Also, similarity studies have been widely used in the phylogenetic reconstruction based on molecular grounds. Most of the current sequence analysis methods are based on alignments, i.e. aligning areas of sequences at several length scales, from single genes to whole genomes. Each alignment is evaluated with a score that depends on the number of same and contiguous characters in the sequences. Optimal methods for sequence alignments rely on dynamic programming techniques, the most widely used optimal sequence alignment algorithms being the ones of Needleman and Wunsch [16] and Smith and Waterman [17]. These algorithms are computationally demanding and their complexity grows exponentially with the length of the sequences. Heuristics have been proposed that solve the sequence alignment problem, e.g. BLAST [18] and FASTA [19]. In order to perform the alignment of multiple sequences more efficiently, several algorithms have been proposed: ClustalW [20], Muscle [21], Mafft [22], and Motalign [23].

Since the first decades of systematic sequencing of protein coding and non-coding genomic regions, it has been noticed that, while alignment of protein coding genomic segments both between organisms and inside genomes reveals a richness of information, it has limited application on the non-coding. This is because the non-coding, in its greatest part is not evolutionary constrained (i.e. conserved due to its functionality in the course of evolutionary time). Nonetheless, alignment methods may be useful when applied in short non-coding DNA stretches. In larger chromosomal regions their use is limited, because in long regions synthesis is not conserved between organisms, which are evolutionarily relatively distant. Alignment is particularly useful in the study of transposable elements, which are found in multiple copies in most organisms, and are marked by variable degrees of homology between them, depending on their age in the genome. A recent application of alignment of whole genomes between two or more species led to the aforementioned discovery of thousands of conserved and highly conserved non-coding genomic elements (CNEs, UCNEs, etc.) through various strategies.

Alignment-free methods are valuable when we want to extract compositional profiles and preferences, and are applicable both in whole-genome comparisons between organisms and in intra-genomic detection of segments which exhibit particular modalities in their composition, often related to their functionality. One classical alignment-free method is based on studying distances between the genomic signatures of sequences, which is briefly presented in the methods section and is used in the present study for comparison with the novel method applied herein. Based on alignment free techniques, there are also various algorithms for locating CpG islands, which are short genomic sequence stretches with no mutual similarity but with several distinct compositional traits. Protein coding segments could be determined by both alignment and alignment-free techniques, as the use of the genetic code and the modalities of the machinery of protein synthesis (mRNA-ribosome binding, tRNA abundances, etc.) endow the protein-coding part of the genome with characteristic compositional biases. Overall, we could say that alignment and alignment-free methods are complementary and that particular components of the genome could be studied by suitable combinations of both.

Alignment free techniques have been proven to be particularly successful in phylogeny and sequence analysis [24]. In alignment free methods the similarity of two sequences is assessed based only on the dictionary of subsequences that appear within the studied sequences, irrespective of their relative position. A promising alignment free method is based on the feature vector representation of a sequence [24-26], and [27] where a sequence is encoded in a vector containing the occurrences of its substrings (k-mers). A k-mer is defined as a subsequence of the original sequence, k characters long. In k-mer frequency analysis every possible k-mer over the nucleotide alphabet {A, C, G, T} is extracted, its occurrences in the original sequence are counted and its frequency is calculated. A vector containing the relative frequency of every k-mer is computed for each sequence in the analyzed data set. Two sequences are rated similar by analyzing the dictionary of their subsequences, without taking care of their relative position.

CNEs, as their name suggests, are identified through pairwise or multiple sequence alignments between two or more genomes. They tend to appear in single copies in the genome [1,28,29] and have been proposed as markers for phylogenomic studies [30]. Herein, we propose an approach that does not rely upon the information content of an alignment between different inter-genomic DNA regions and we apply it to the classification of functional sequences taken from several genomes, which range from invertebrates to vertebrates. We also proceed to several comparisons that test the robustness of our approach by comparing the performance of our proposed method in classifying genomic elements not only between species, but also within the same organism and of different possible functionalities and evolutionary depth. For that purpose, we use a wide collection of vertebrate conserved non-coding elements published in the literature as well as new datasets of human and invertebrate CNEs identified in this study.

2. Methods

2.1. Published datasets

In this study, we consider several published datasets of constrained sequences extracted from the human (*Homo sapiens*), insect (*Drosophila melanogaster*) and worm (*Caenorhabditis elegans*) genomes. We randomly select 1000 elements from each superset for subsequent analysis, given the heterogeneity of the datasets. Only worm elements are studied in their entirety, since this set includes only 1869 elements. The exonic sequences of human, worm and insect genomes are downloaded from UCSC [31]. Apart from exonic sequences, several classes of constrained non-exonic sequences are used as follows (for various metrics, also see the Supplementary Excel available at dmb.iasi.cnr.it/cnes.php):

(a) UltraConserved Non-coding Elements (UCNEs): These are human non-protein coding sequences (hg19 assembly) that display ≥95% sequence identity when compared to the chicken genome and are longer than 200 bp [32].

(b) EU100 non-exonic CNEs (EU100nx CNEs): These are human sequences (hg18 assembly) that are identical over at least 100 bp in at least 3 out of 5 placental mammals (human, mouse, rat, dog and cow) [29]. The whole set is named EU100+ and since we remove elements overlapping exons, we name it EU100nx.

(c) Amniotic and mammalian CNEs: These are elements identified by Kim and Pritchard [33]. Mammalian CNEs are sequences that are conserved within mammals but not found in chicken or fish, while Amniotic CNEs are conserved in mammals and chicken but not found in fish. LiftOver is used in order to convert the coordinates to the most recent release of the human genome (from hg17 to hg19).

2.2. Identification of sets of CNEs in vertebrates, insects and worms

For the needs of our study, we also identify sets of conserved elements in vertebrates, insects and worms using uniform criteria in terms of the similarity thresholds applied between the compared sequences and of the considered minimum length. We choose to compare whole genomes of *D. melanogaster*/*Drosophila virilis* and *C. elegans*/*Camellia japonica* because the species which form these pairs are distant enough to allow a clear separation between functionally conserved and neutrally evolving genomic sequences, while they are selected so that evolutionary distances within every pair are close [34]. Whole-genome alignments between *D. melanogaster (dm3)*/*D. virilis* (droVir3) and *C. elegans (ce10)*/*C. japonica (caeJap4)* are downloaded from the UCSC Genome Browser [31]. Sequence regions, where the percentage of sequence identity is consistently ≥90% and the length is >60 bp, are considered as conserved elements. The sequence identity is computed in an asymmetric fashion by taking as references *D. melanogaster* and *C. elegans* genomes and counting the number of conserved bases in the target species in a 61 bp sliding window. The number obtained from every window is used to assign a percentage identity value to the base at the center of this window, as described in [34]. In total, 45,345 sequences are identified as conserved between *D. melanogaster* and *D. virilis* and 1869 between *C. elegans* and *C. japonica.* Based on the Ensembl gene annotations for *D. melanogaster* and *C. elegans* genomes, each of the conserved sequences is classified as protein-coding or non-coding (intronic, UTR-associated or intergenic).

We also identify vertebrate CNEs (based on comparisons between human and chicken) that exhibit various degrees of identity (75–80%, 80–85%, 85–90%, 90–95%). To obtain CNEs that have identity between e.g. 75% and 80%, we first extract a set of CNEs that exhibits identity of 75% (using the same methodology as described above) and from this set we remove the CNEs that exhibit identity of 80%. From the remaining sequences, we keep only those that are longer than 199 bp. We repeat the same procedure for the other sets of CNEs, accordingly.

2.3. Surrogate sequences extraction

FASTA sequences are obtained from BED files using the *fastaFromBed* package of BEDTools [35]. Overlapping elements between different datasets are calculated using *intersectBed* package from the same suite of tools. In addition, we apply the EMBOSS suite to estimate fractional GC content of sequences [36]. Given a CNE or another functional element of each collection under study, an analogue of it is extracted from the corresponding genome. Every such element (surrogate sequence) is ensured to be of the same length and GC content (±1%) with its corresponding element in the collection under study. Statistics of the datasets (mean length of sequences and GC content) are available in the Supplementary Excel available at dmb.iasi.cnr.it/cnes.php. As vertebrate and invertebrate CNEs are of different lengths (the ones belonging to the latter category are considerably smaller, see [34]), we make sure that we take elements of same lengths as follows (in all these cases the compared sets include one thousand sequences each): (i) each time, we compute the length of one element from the 'short sequences collection', (ii) then we take one element from the 'long sequences collection' and we trim it retaining only its central part equal to the length of the corresponding 'short sequence', (iii) we repeat this procedure exhaustively (1000 times in all the considered datasets). Thus, we end up with a new dataset for the 'long sequences' (collection of sequences in the form of a BED file), of which members have lengths equal to the members of the 'short sequences collection'.

2.4. The alignment-free sequences classification method LAF

To show that we can distinguish CNEs from other functional sequences and from surrogate sequences based solely on sequence intrinsic properties, we apply an alignment-free method based on

- a feature vector representation of the sequences and
- classification with rule based supervised machine learning techniques.

And we call it LAF (Logic Alignment Free).

The feature vector is a well-established technique for representing biological sequences and for permitting a classification of them. This methodology is described in Kudenko and Hirsch [25], Vinga and Almeida [24] and Xing et al. [27], where sequence representations with feature vectors are introduced and combined with supervised classification methods, e.g., Support Vector Machines, for classifying specimen to species. Another recent work is the one of Kuksa and Pavlovic [26], who apply feature vector representations for DNA Barcode classification.

The feature vector representation of a sequence S is based on the computation of the substrings occurrences of a given length k in the original sequence by applying a sliding window in S. These substrings are called k-mers. The k-mer counts of a sequence are represented in a feature vector, where each component of the vector is associated with the occurrences of a particular k-mer.

The authors of Ref. [24] provide the following formal definition. Let S be a sequence of n characters over an alphabet A, e.g. A = {A,C,G,T}, and let $k \in I$, $k < n$, $k > 0$. If K is a generic subsequence of S of length k, K is called k-mer. Let the set $V = \{\text{k-mer}_1, \text{k-mer}_2, \ldots, \text{k-mer}_r\}$ be all possible k-mers over A, V has size $t = |A|^k$. 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 feature vector C contains for each k-mer its occurrences (or counts) $C = \{c_{\text{k-mer}1}, c_{\text{k-mer}2}, \ldots, c_{\text{k-mert}}\}$. The frequencies are then computed accordingly and stored in a vector $F = \{f_{\text{k-mer}1}, f_{\text{k-mer}2}, \ldots, f_{\text{kmert}}\}$, for a generic k-mer j the frequency is defined as $f_j = c_{\text{kmerj}}/(\text{n-k} + 1)$. For example considering the 4 letters alphabet {A, C, G, T}, the 2-mers, and the sequence ACGACT, the feature vector C is:

AA	AC	AG	AT	CA	CC	CG	CT	GA	GC	GG	GT	TA	TC	TG	TT
0	2	0	0	0	0	1	1	1	0	0	0	0	0	0	0

and the frequencies vector F is:

AA	AC	AG	AT	CA	CC	CG	CT	GA	GC	GG	GT	TA	TC	TG	TT
0	2/5	0	0	0	0	1/5	1/5	1/5	0	0	0	0	0	0	0

The sequences are so represented in a coordinate space, that is mathematically tractable with linear algebra and statistics, e.g., by considering the vector representation of the sequences it is possible to compute different distance measures between two sequences or to give the vector representation as input to a supervised machine learning algorithm, i.e. a classifier.

The supervised machine learning approach is also called classification: any collection of the analyzed sequences must contain an a priori known class label, i.e. every sequence is associated with a given class, e.g. *Vertebrate*, *Invertebrate*, *Amniotic*, and *Mammalian*. Such a collection is called training set. Based on the training set, the method extracts a classification model, e.g., "if-then rules", for distinguishing the different sequences present in the data set. The model can then be evaluated on a test set, that may be composed of unknown sequences or sequences that belong to a known class, in order to verify the classification performances. Our adopted approach combines the feature vector representation with rule based supervised machine learning methods [37,38].

As a technique for biological sequence analysis, rule based classifiers have been proposed in Bertolazzi et al. [39] and in Weitschek et al. [40], in particular for classifying organisms using DNA Barcode sequences and for viruses identification. In these works, the genomic sequences were analyzed with a positional approach: each nucleotide was analyzed independently by referring only to its position. Therefore an alignment of the sequences or an overlapping gene region was necessary: an analysis of the characteristic nucleotides present in a determined position for every class was performed, leading to logic rules of the type, e.g., *if pos90 = A then the sequence belongs to class X*. The alignment was necessary, because a positional analysis is only possible when the sequences come from the same gene regions and are aligned on a reference position.

The output of a rule based classifier is a collection of if-then rules, assigning a sequence to a particular class (species), e.g., *if pos30 = A and pos40 = C then the sequence belongs to the squalus edmundsi species*. The major advantage of rule-based classification is the additional knowledge that is given by the compact human interpretable model (the if-then rules). In this work, the following algorithms are taken into account for performing the rule based classification analysis of the CNEs sequences feature vector representations: RIPPER [41], RIDOR [42] and PART [43].

RIPPER is a classification algorithm that uses a direct method for rules extraction: it infers the rules by processing directly the data. RIPPER is composed of following computational steps:

1) rule growth
2) rule pruning
3) model optimization
4) model selection.

In the first step, the rules are computed in a greedy way by processing the data attributes. The rules are then pruned (simplified) in step 2 according to statistical measures on the training set. In step 3 the rules are optimized by extending them and adding new pruned rules. In the last step the best performing rules are selected and the remaining rules are discarded from the classification model.

The RIDOR classification algorithm also extracts the rules directly from data. It firstly computes a default rule that covers the most represented class (e.g., "*all sequences are Vertebrates*") and then exceptions rules which cover the other classes (e.g., "*except if freq(ACG) > 250 and freq(AGT) < 200*" *these sequences are Invertebrates*).

On the other hand, PART is an indirect method for rules extraction: it processes the data by using the C4.5 tree based classifier [44] generating a pruned decision tree for every iteration. Finally it selects the best classification tree and uses the leaves as logic rules.

According to the previously described methods (feature vector representation and rule based supervised learning) the LAF approach that is adopted in this work performs for every sequence s in a data set, which is associated to a class (e.g. Vertebrate, Invertebrate), the following computational steps (see Fig. 1):

1. The reverse complement of the sequence s is computed.
2. The sequences S and its reverse complement s_r is combined by concatenation in a sequence S.
3. The counts of the k-mers (for k = 3,4,5) are calculated on the sequence S, obtaining the feature vector $C = \{c_{k\text{-mer}1}, c_{k\text{-mer}2}, \ldots, c_{k\text{-mer}t}\}$.
   The k-mer counts are extracted with the Jellyfish software [45]. k has been chosen between 3 and 5, based on the following references [46,47], which state the optimality of such lengths as they provide an ideal balance between the length of the subsequences and their number, when the sequences are expressed in the 4-letter (A,C,G,T) alphabet.
4. The frequencies of each k-mer are computed: $f_j = c_{kmerj}/(n\text{-}k + 1)$.

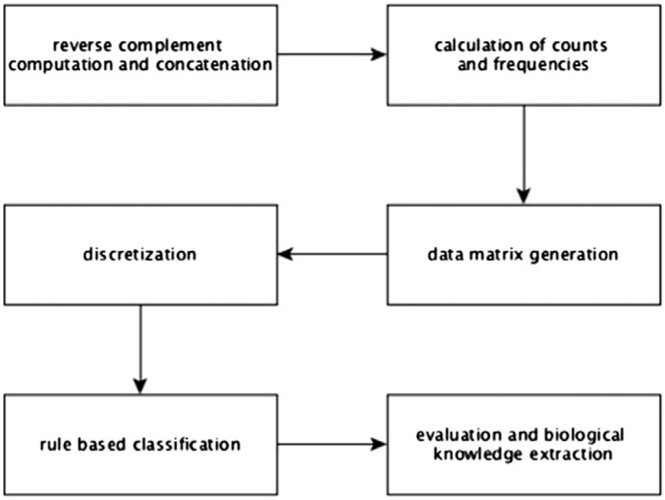
reverse complement computation and concatenation
calculation of counts and frequencies
data matrix generation
discretization
rule based classification
evaluation and biological knowledge extraction

Fig. 1. The LAF method flow chart.

5. A feature frequency vector of each sequence is obtained:

$$F = \{f_{k\text{-mer}1}, f_{k\text{-mer}2}, \ldots, f_{kmert}\}$$

6. The feature vectors are combined in a data matrix, where each row represents a sequence and each column the k-mer frequency. A header with the name of the sequences and their belonging class is present. An example of the data matrix is given in Table 1.
7. The numeric data sets are discretized, i.e. the frequencies are converted from numerical to nominal by the definition of intervals, according to Fayyad & Irani's MDL method [48]; the discretization procedure improves the performance of the rule based algorithms.
8. The data matrix is given as input to three rule based supervised machine learning algorithms: RIPPER, RIDOR, and PART.
9. The classification methods are run in 10 fold cross validation mode to evaluate the performances.
   Cross validation is a standard sampling technique that splits the dataset in a random way in m disjoints sets, and the data mining procedure is run m times with different sets. At a given run n the nth subset is used as test set and the remaining m-1 sets are merged and used as training set for building the model. Each of the m sets contains a random distribution of the data. The cross validation sampling procedure builds m models and each of these models is validated with a different set of data. Classification statistics are computed for every model and the average of these represents an accurate estimation of the data mining procedure performance.
10. The classification models are extracted, e.g.,
    if freq(AAAC) < 0.195 then the organism is Vertebrate;
    if freq(AAAC) > 0.195 then the organism is Invertebrate.
11. The models are evaluated in terms of correct classification rate.

Table 1
Data matrix example of the LAF method.
An example of the data matrix obtained by the application of the LAF method. The data matrix is composed by two heading columns, the first with the identifiers of the sequences and the second with the class labels. The following rows contain the frequency values of the k-mers in the sequences.

	Seq 1	Seq 2	….	Seq N-1	Seq N
	Vertebrate	Vertebrate	….	Invertebrate	Invertebrate
AAA	0.46	0.26	…	0.24	0.54
AAC	0.12	0.16	…	0.23	0.24
AAG	0.12	0.23	…	0.23	0.23
….	…	…	…	….	…

Scripts for filtering, reverse complementing, joining the data sets, calculating the frequencies, discretizing, and classifying have been implemented and are available upon request.

The Weka [48] implementations of the ruled based classification algorithms are adopted for performing the analysis. The experiments have been run under a 64-bit Debian Linux workstation with kernel 2.6.26, 32GB of RAM, and an Intel i7 4-core processor.

2.5. The "genomic signature" method

A different technique for classification of biological sequences is considered for a direct comparison with the results obtained using the LAF approach proposed here. A classical method for quantifying the neighbor preferences in a DNA sequence of an entire genome is the computation of the vector of the odds ratios for dinucleotides [49]. The odds ratio of each dinucleotide is the quantity: $\rho_{ij} = f_{ij}/(f_i\, f_j)$, where f_{ij} and f_i, f_j stand for the frequencies of occurrence in the studied sequence of a dinucleotide and its constituent nucleotides respectively. Thus subscripts *i*, *j* represent any pair of A, G, C and T. This is the ratio of the "observed" dinucleotide frequency over the "expected" one under no neighbor preference, thus it expresses the actual neighbor preferences of the given pair of nucleotides. Before computing the odds ratios for a given sequence, this is concatenated to its reverse complement. Consequently, the relevant ratios are only ten, i.e. four for the self-complementary dinucleotides and six for the mutually complementary couples. Karlin and co-workers found that these quantities differentiate between different genomes, according, approximately, to their evolutionary distance [50]. Thus they have assigned to the vector of these ten "first neighbor preferences" the name of genomic signature [51]. In what follows, we use classification based on genomic signatures for comparison with classification based on the alignment-free method described in the previous section. We also consider RIPPER, PART and RIDOR as classifiers for all the experiments and we also perform the discretization procedure as described previously.

All scripts used throughout this work are available upon request to interested readers. All the used files (in BED format) are available as Supplementary material at dmb.iasi.cnr.it/cnes.php.

3. Results and discussion

In this section, we present a systematic classification analysis of short genomic segments displaying different functionalities and found in the human and other genomes by using a supervised machine learning approach. For every classification task, we adopt the technique of feature vector representation and rule based classification explained in previous sections. Every classification task is performed using feature vector representation with k-mers of length 3, 4 and 5, which are given as input to three different rule based algorithms (RIPPER, RIDOR, PART), adopting a 10-fold cross validation sampling technique. The full list of accuracy results with all parameters is summarized in Supplementary Excel available at dmb.iasi.cnr.it/cnes.php. We omit the percentages of false positives and false negatives, because the errors are equally distributed between the different classes in the datasets. We report here and discuss based on the best result (highest classification rate) among each different rule based algorithm, applied on the feature vectors composed of k-mers length 3, 4, 5. The reason for choosing the overall best for this purpose is that we consider that it captures more directly the potential of k-mers based classification and consequently it might be correlated more meaningfully with several biological characteristics and reveal functional modalities of the studied sequences. In the Supplementary Excel, interested readers are provided with the average classification rates as calculated based on all 26 experiments performed. Based on our results, we suggest k = 3 and PART as the optimal parameters for an ab initio classification experiment (see Supplementary Excel). For all the considered experiments, genomic signatures are also used as an alternative to the k-mers based classification method. Genomic signatures, as characteristic constitutional traces of different genomes, have a long history and represent a classical standard to which we would like to compare the k-mers based approach. In the case of genomic signatures, the best overall result in terms of algorithm used is also considered.

The following classification analyses are presented in order to assess the potential usefulness of our approach in distinguishing among different genomic elements in the same or in different metazoan genomes. The classification rules extracted with the alignment-free sequences classification method (LAF), described in the "Methods" section, are available for download at dmb.iasi.cnr.it/cnes.php. The interested reader can identify the characteristic k-mers for every pairwise experiment and class of the investigated data sets.

3.1. Inter-species comparison of background sequences

In these comparisons between human, worm and insect background sequences we use as representative samples of the different genomes surrogate sets, composed either for exons or for CNEs. Comparisons involving *H. sapiens* yield always the best classification rates using both LAF and Genomic Signatures (GS), see Table 2. This may be understood on the grounds of the high difference of neighbor preferences, mainly in CpG and TpA between *H. sapiens* and the invertebrates, while these preferences are found to be close between

Table 2
Comparison of background genomic sequences from different organisms.
Inter-genomic comparisons between pairs of collections of 1000 background sequences each (non-CNE, non-exonic) from the three studied genomes. These sequences are produced as surrogates for the CNEs and exons studied later. Each sequence has same length and same GC% with one member of the CNE or exon collections. The highest value of LAF or GS accuracy in each classification experiment is shown in bold. For more details on the used methods see in the text and supplementary material.

Experiment no	Description	Average length	Average GC	LAF, accuracy	GS, accuracy
#1	Surrogate for human exons	167.84	0.40	84.21	86.93
	Surrogates for worm exons	169.32	0.52		
#14	Surrogates for human UCNEs	86.09	0.36	81.45	84.30
	Surrogates for insect UCNEs	86.58	0.39		
#20	Surrogates for human exons	169.82	0.51	84.66	87.89
	Surrogates for insect exons	169.32	0.52		
#22	Surrogates for worm exons	213.37	0.42	78.60	70.95
	Surrogates for insect exons	212.86	0.52		
#23	Surrogates for worm UCNEs	83.18	0.43	82.10	84.48
	Surrogates for hUCNEs	82.93	0.36		
#13	Surrogates for worm UCNEs	83.41	0.43	64.70	63.65
	Surrogates for insect UCNEs	86.58	0.39		
Average				79.29	79.70

invertebrate species: *D. melanogaster* (insect) and *C. elegans* (worm). GS is exclusively a quantification of first neighbor preference, while in LAF several other components of the sequence composition are implicitly included alongside, e.g. mono-nucleotide composition and higher order neighbor preference. In accordance with the above description, in cases where human sequences are included in the comparison, GS perform systematically better than LAF, while in the remaining two cases LAF is the best.

Another conclusion that stems from simple inspection of the Table 2 is that in interspecies comparisons between background genomic sequences, the best results are obtained for exonic surrogates (vs. CNE surrogates). This relies on the simple fact that the mean lengths of exons, in all cases, clearly exceed mean lengths of CNEs (and the same holds true for their surrogate sequences, by construction). Higher mean lengths obviously allow better statistics and thus higher classification rates in both LAF & GS. Note that, concerning interspecies CNE comparisons, the sequences of the vertebrate CNE set are always trimmed to the mean length of the shorter invertebrate CNEs set (for details see in the Methods section).

3.2. Comparison of constrained sequences vs. their corresponding surrogates

The following experiments refer to comparisons of constrained sequences against their surrogates (Table 3). Note that surrogates share with the initial sequences the same GC% and length (see the Methods section). As evidenced, comparisons involving invertebrate constrained sequences are not classified as successfully as their human counterparts using LAF. The latter may be understood on the grounds of several particularities of the warm-blooded animals (and often of all vertebrates) especially in their non-functional, non-constrained background genomic fraction. These include a high enrichment in repetitive sequences. Such genomes also present a typical profile of avoided dinucleotides (especially CpG and TpA) that are less avoided in the constrained elements (exons, CNEs), which having functional roles, do not strictly follow the average genomic compositional trends. Note that invertebrate genomes are much less abundant in repeats and less marked by underrepresentation of specific dinucleotides. In the comparisons by means of GS, the same trend is followed, but differences are minor, due to the simpler structure of this genomic metrics. Furthermore, we know from earlier studies that genomic signatures do not perform well in intra-species comparisons because neighbor preferences remain relatively constant within the same genome. Therefore, in all cases listed, LAF performs better than GS.

Table 3
Comparison of constrained sequences versus their corresponding surrogates.
Intra-genomic comparisons between pairs of collections of CNEs or exons vs. surrogates. The highest value of LAF or GS accuracy in each classification experiment is shown in bold. Here worm CNEs are studied in their entirety against their surrogates (1869 sequences). For more details on the used methods see in the text and supplementary material.

Experiment no	Description	Average length	Average GC	LAF, accuracy	GS, accuracy
#2	Worm exons	213.37	0.42	65.63	63.75
	Surrogates	213.37	0.42		
#3	Human exons	169.82	0.52	73.31	65.81
	Surrogates	169.82	0.52		
#17	Insect exons	388.82	0.54	63.72	61.95
	Surrogates	381.56	0.54		
#4	Worm UCNEs	82.88	0.43	61.00	57.71
	Surrogates	82.88	0.43		
#5a	Human UCNEs	326.92	0.37	81.15	73.55
	Surrogates	326.92	0.37		
#5b	Human EU100nx CNEs	155.50	0.38	75.95	65.15
	Surrogates	155.50	0.38		
#5c	Amniotic CNEs	289.06	0.38	76.25	66.10
	Surrogates	289.06	0.38		
#5d	Mammalian CNEs	246.49	0.40	73.65	60.00
	Surrogates	246.49	0.40		
#12	Insect UCNEs	86.58	0.39	67.95	66.95
	Surrogates	86.58	0.39		
Average				70.96	64.55

The last six rows of Table 3 denote comparisons of several CNE sequences collections against their surrogates. In general we notice that among constrained sequences, human CNE sequences vs. surrogates exhibit relatively higher classification rates, if compared with exonic sequences vs. their corresponding surrogates. Although, we are not able at this stage to clearly interpret this finding, it might be related to the hypothesis that CNEs orchestrate highly sophisticated developmental processes including brain formation [11]. The instructions that direct these processes might be hardwired and reflected in their sequences themselves. It is known from the literature that CNEs do serve as transcription factor binding sites and bear several motifs [52, 53] that k-mer analysis is possibly sensitive enough to detect and thus increase the found CNE vs. background towards exons vs. background classification rates.

3.3. Intra-species comparison of constrained sequences (functional sequences vs. functional sequences of a different type)

3.3.1. Case of exons vs. UCNEs

We next proceed to the study of the sequence characteristics of elements that are characterized as functional; exons that are known to be under selective constraints and encode for polypeptide chains and CNEs that are mostly of unknown functionality, which is though implied by their high degree of conservation. At a first glance, differences in classification rates between LAF and GS, tabulated in Table 4, may be related to the gradation of GC content between the studied classes. Indeed, it is known from the literature that CNEs are generally AT rich [15], while protein coding genes are relatively GC rich, see also GC content values included in Table 4. However, this may not be entirely true, as classification success in the case of worm exons vs. worm UCNEs is medium (see experiment #6, 68.65%) while CG content difference is minimal. This might indicate that exons and CNEs functional differences are inscribed in their sequence composition. Known such modalities are codon biases and amino-acid frequencies in coding exons, or inscription of several protein binding and other consensus sequences frequent within CNEs.

3.3.2. Case of different classes of CNEs

Here we compare different classes of CNEs which are identified using the same criteria but with gradually increasing similarity thresholds, using a step of 5 percentage points per dataset (see the Methods section). Note that differentiation on the basis of GC composition between these datasets is quite low. However, we still obtain good

Table 4
Comparison of different types of constrained sequences.
Pair-wise intra-genomic comparisons between collections of CNEs, vs. exons, containing 1000 sequences each. The highest value of LAF or GS accuracy in each classification experiment is shown in bold. For more details on the used methods see in the text and supplementary material.

Experiment no	Description	Average length	Average GC	LAF, accuracy	GS, accuracy
#6	Worm exons	82.64	0.42	68.65	61.88
	Worm UCNEs	83.11	0.43		
#7	Human exons	169.82	0.52	82.79	77.66
	Human UCNEs	169.32	0.37		
#18	Insect exons	86.09	0.52	82.80	81.90
	Insect UCNEs	86.58	0.39		
Average				78.08	73.81

Table 5
Comparison of CNEs identified using the same criteria but with different thresholds. Intra-genomic comparisons between pairs of collections of 1000 human CNE sequences each. These datasets consist of sequences identified between pairwise human/chicken alignments with gradually increasing similarity thresholds. The highest value of LAF or GS accuracy in each classification experiment is shown in bold. For more details on the used methods, see in the text and supplementary material.

Experiment no	Description	Average length	Average GC	LAF, accuracy	GS, accuracy
#24	CNEs 75–80%	248.92	0.39	55.45	53.95
	CNEs 80–85%	248.39	0.38		
#25	CNEs 75–80%	248.92	0.39	57.45	50.00
	CNEs 85–90%	250.63	0.38		
#26	CNEs 75–80%	248.92	0.39	60.95	59.05
	CNEs 90–95%	268.64	0.37		
#27	CNEs 75–80%	248.92	0.39	68.50	63.25
	CNEs 95–100%	319.70	0.37		
#28	CNEs 80–85%	248.39	0.38	50.00	50.00
	CNEs 85–90%	250.63	0.38		
#29	CNEs 80–85%	248.39	0.38	58.90	54.30
	CNEs 90–95%	268.64	0.37		
#30	CNEs 80–85%	248.39	0.38	68.60	61.80
	CNEs 95–100%	319.70	0.37		
#31	CNEs 85–90%	250.63	0.38	58.00	50.00
	CNEs 90–95%	268.64	0.37		
#32	CNEs 85–90%	250.63	0.38	65.65	59.30
	CNEs 95–100%	319.70	0.37		
#33	CNEs 90–95%	268.64	0.37	61.55	50.00
	CNEs 95–100%	319.70	0.37		
#34	CNEs 75–80%	248.92	0.39	76.20	64.10
	Surrogates	248.92	0.39		
#35	CNEs 85–90%	250.63	0.38	78.45	65.25
	Surrogates	250.63	0.38		
#5a	CNEs 95–100%	326.92	0.37	81.15	73.55
	Surrogates	326.92	0.37		
Average				64.68	58.04

classification rates, clearly better than those using genomic signatures. Especially high accuracy values are obtained when we deal with ultraconserved elements (UCNEs, 95–100% similarity between human and chicken, experiments #27, #30, #32, #33, #5a). UCNEs are presumed to form a distinct class, distinguishing themselves from CNEs of the other datasets included in Table 5 comparisons, that are identified from the same organisms but using less stringent criteria of sequence identity. Furthermore, when we compare CNEs 75–80%, CNEs 85–90% and CNEs 95–100% vs. their corresponding surrogate sequences in the human genome, we observe a gradual increase in performance (compare #34 and #35 with #5a). Note here, that we consider for this kind of experiments raw sequences that are not processed in any way (i.e. trimming is not performed). Overall, as we have mentioned earlier, we get the best classification rates when we have UCNEs (Experiment #5a). The information content hardwired in those sequences probably renders them detectable by our method with high efficiency. The genomic signatures approach also performs well in this case, which means that those sequences do probably form a distinct category of ultraconserved sequences from the point of view of first neighbor preferences too.

3.4. Classification based on k-mer analysis and rule based classifiers proves efficient in distinguishing sequences from the same species (intra-species comparisons)

By definition, genomic signatures are widely used for the identification of different species [50]. Using a graphical representation of genomic composition termed Chaos Game Representation (CGR), Deschavanne et al. reported that subsequences of a genome display the main characteristics of the whole genome converging to the genomic signature concept [54]. Interestingly, the authors attributed the variability observed among sequences to base concentrations, runs of complementary bases or purines/pyrimidines and over- or under-expressed words of various lengths with the aim to measure phylogenetic proximity. We categorize the performed experiments as inter-species or intra-species based on whether the sequences under study derive from different or the same organism respectively. Based on the statistics presented in Table 6, it seems that our method performs almost equally well with the GS approach, when dealing with sequences from different species (80.20% vs. 80.19% for the genomic signatures). When it comes to intra-species comparisons, however, the method proposed herein performs fairly better (70.90% vs. 65.83%). This lies in accordance with the fact that k-mer analysis of relatively short sequences combined with logic mining classifiers is a promising method, since it also takes into account the occurrence in the studied sequences of other oligonucleotide stretches such as homopurines, homopyrimidines (unpublished results) or overlapping transcription factor binding sites that are found in the sequences and represent another type of “signature” that distinguishes them from the bulk or from other non-constrained sequences.

4. Conclusions

We present and apply an approach based on k-mer analysis and rule based classification called Logic Alignment Free (LAF) in order to represent and subsequently classify biological sequences of different functionalities and origins. We compare this approach vis-à-vis Karlin's Genomic Signature (GS) method, and present classification rates. To our knowledge, our study is the first one that applies those methodologies on short biological sequences, as up to now GS had only been applied in more extended genomic regions of 50 kb or more [49]. Based on our results, we deduce that LAF performs particularly well when dealing with sequences from the same species surpassing the performance of Genomic Signatures (GS), while in inter-species comparisons, where the potential of GS has already been validated by comparing large genomic regions, the two methods perform equally well. Therefore, LAF analysis of biological sequences could be further used in applications involving sequence analysis such as categorization of different possible functionalities within groups of CNEs or identification of metagenomic components.

Competing interests

The authors declare that they have no competing interests.

Table 6
Overall statistics describing the effectiveness of our method. LAF (Best).
The best combination for each experiment, in terms of k-mer and classifier used, is considered and then averaged, LAF (average of 9): All the different combinations for each experiment of k-mers and rule-based classifiers are averaged, LAF (k = 3, PART): Based on all the experiments performed, the most suitable combination of k and algorithm is used for the analysis in each experiment, GS (without discretization): genomic signatures as calculated without the extra discretization step (described in the Methods section) that we apply when we perform our analysis.

	LAF, accuracy, Best	LAF, accuracy, average of 9	LAF, accuracy, k = 3, PART	GS, accuracy, without discretization	GS, accuracy, Best, with discretization	GS, accuracy, mean of 3, with discretization
Overall	**75.19**	71.33	74.39	70.86	72.45	71.78
Inter-species	**80.20**	75.56	79.81	79.62	80.19	79.53
Intra-Species	**70.90**	67.69	69.74	63.35	65.83	65.14

Authors' contributions

D.P, E.W and Y.A ideated the work, wrote the paper, designed and performed the classification experiments. E.W implemented the scripts and software for feature vector representation and ruled based supervised learning. D.P implemented scripts for genomic analysis and construction of genomic surrogates. S.D generated new CNE datasets. P.B, G.F and Y.A directed research, discussed results and revised the manuscript.

Acknowledgments

E.W and D.P would like to express their gratitude to the organizers of the 2012 European Molecular Biology Organization (EMBO) Practical Course: Bioinformatics and Comparative Genome Analyses in Naples, which enabled this collaboration. This work is partially supported by the Italian PRIN GenData 2020 (2010RTFWBH) and the Flagship Project InterOmics (E.W., PB.05). D.P would also like to thank EMBO for giving him the opportunity to visit the lab of P.B. at EPFL, Switzerland under a short-term fellowship (EMBO ASTF 453-2012).

References

[1] G. Bejerano, M. Pheasant, I. Makunin, S. Stephen, W.J. Kent, J.S. Mattick, et al., Ultraconserved elements in the human genome, Science 304 (2004) 1321–1325.
[2] S. Dimitrieva, P. Bucher, UCNEbase—a database of ultraconserved non-coding elements and genomic regulatory blocks, Nucleic Acids Res. 41 (2013) D101–D109.
[3] Y. Sakuraba, T. Kimura, H. Masuya, H. Noguchi, H. Sezutsu, K.R. Takahasi, et al., Identification and characterization of new long conserved noncoding sequences in vertebrates, Mamm. Genome 19 (2008) 703–712.
[4] E.T. Dermitzakis, A. Reymond, N. Scamuffa, C. Ucla, E. Kirkness, C. Rossier, et al., Evolutionary discrimination of mammalian conserved non-genic sequences (CNGs), Science 302 (2003) 1033–1035.
[5] N. Harmston, A. Baresic, B. Lenhard, The mystery of extreme non-coding conservation, Philos. Trans. R. Soc. Lond. B Biol. Sci. 368 (2013) 20130021.
[6] G. Elgar, T. Vavouri, Tuning in to the signals: noncoding sequence conservation in vertebrate genomes, Trends Genet. 24 (2008) 344–352.
[7] A. Sandelin, P. Bailey, S. Bruce, P.G. Engstrom, J.M. Klos, W.W. Wasserman, et al., Arrays of ultraconserved non-coding regions span the loci of key developmental genes in vertebrate genomes, BMC Genomics 5 (2004) 99.
[8] D. Polychronopoulos, D. Sellis, Y. Almirantis, Conserved noncoding elements follow power-law-like distributions in several genomes as a result of genome dynamics, PLoS One 9 (2014) e95437.
[9] L.A. Pennacchio, N. Ahituv, A.M. Moses, S. Prabhakar, M.A. Nobrega, M. Shoukry, et al., In vivo enhancer analysis of human conserved non-coding sequences, Nature 444 (2006) 499–502.
[10] A. Visel, S. Prabhakar, J.A. Akiyama, M. Shoukry, K.D. Lewis, A. Holt, et al., Ultraconservation identifies a small subset of extremely constrained developmental enhancers, Nat. Genet. 40 (2008) 158–160.
[11] M. Matsunami, N. Saitou, Vertebrate paralogous conserved noncoding sequences may be related to gene expressions in brain, Genome Biol. Evol. 5 (2013) 140–150.
[12] E.A. Glazov, M. Pheasant, E.A. McGraw, G. Bejerano, J.S. Mattick, Ultraconserved elements in insect genomes: a highly conserved intronic sequence implicated in the control of homothorax mRNA splicing, Genome Res. 15 (2005) 800–808.
[13] T. Vavouri, K. Walter, W.R. Gilks, B. Lehner, G. Elgar, Parallel evolution of conserved non-coding elements that target a common set of developmental regulatory genes from worms to humans, Genome Biol. 8 (2007) R15.
[14] S. Lockton, B.S. Gaut, Plant conserved non-coding sequences and paralogue evolution, Trends Genet. 21 (2005) 60–65.
[15] K. Walter, I. Abnizova, G. Elgar, W.R. Gilks, Striking nucleotide frequency pattern at the borders of highly conserved vertebrate non-coding sequences, Trends Genet. 21 (2005) 436–440.
[16] S.B. Needleman, C.D. Wunsch, A general method applicable to the search for similarities in the amino acid sequence of two proteins, J. Mol. Biol. 48 (1970) 443–453.
[17] T.F. Smith, M.S. Waterman, Identification of common molecular subsequences, J. Mol. Biol. 147 (1981) 195–197.
[18] S. Altschul, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, Nucleic Acids Res. 25 (1997) 3389–3402.
[19] W.R. Pearson, Rapid and sensitive sequence comparison with FASTP and FASTA, Methods Enzymol. 183 (1990) 63–98.
[20] J.D. Thompson, T.J. Gibson, D.G. Higgins, Multiple sequence alignment using ClustalW and ClustalX, Curr. Protoc. Bioinforma. 2 (2002) (Unit 2.3).
[21] R.C. Edgar, MUSCLE: multiple sequence alignment with high accuracy and high throughput, Nucleic Acids Res. 32 (2004) 1792–1797.
[22] K. Katoh, MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform, Nucleic Acids Res. 30 (2002) 3059–3066.
[23] A. Mokaddeml, M. Elloumi, Motalign: a multiple sequence alignment algorithm based on a new distance and a new score function, 2013 24th Int. Work. Database Expert Syst. Appl., IEEE, 2013, pp. 81–84.
[24] S. Vinga, J. Almeida, Alignment-free sequence comparison—a review, Bioinformatics 19 (2003) 513–523.
[25] D. Kudenko, H. Hirsh, Feature generation for sequence categorization, 1998. 733–738.
[26] P. Kuksa, V. Pavlovic, Efficient alignment-free DNA barcode analytics, BMC Bioinforma. 10 (Suppl. 1) (2009) S9.
[27] Z. Xing, J. Pei, E. Keogh, A brief survey on sequence classification, ACM SIGKDD Explor. Newsl. 12 (2010) 40.
[28] C. McLean, G. Bejerano, Dispensability of mammalian DNA, Genome Res. 18 (2008) 1743–1751.
[29] S. Stephen, M. Pheasant, I.V. Makunin, J.S. Mattick, Large-scale appearance of ultraconserved elements in tetrapod genomes and slowdown of the molecular clock, Mol. Biol. Evol. 25 (2008) 402–408.
[30] J.E. McCormack, B.C. Faircloth, N.G. Crawford, P.A. Gowaty, R.T. Brumfield, T.C. Glenn, Ultraconserved elements are novel phylogenomic markers that resolve placental mammal phylogeny when combined with species-tree analysis, Genome Res. 22 (2012) 746–754.
[31] W.J. Kent, C.W. Sugnet, T.S. Furey, K.M. Roskin, T.H. Pringle, A.M. Zahler, et al., The human genome browser at UCSC, Genome Res. 12 (2002) 996–1006.
[32] S. Dimitrieva, P. Bucher, Genomic context analysis reveals dense interaction network between vertebrate ultra-conserved non-coding elements, Bioinformatics 28 (2012) i395–i401.
[33] S.Y. Kim, J.K. Pritchard, Adaptive evolution of conserved noncoding elements in mammals, PLoS Genet. 3 (2007) 1572–1586.
[34] D. Retelska, E. Beaudoing, C. Notredame, C.V. Jongeneel, P. Bucher, Vertebrate conserved non coding DNA regions have a high persistence length and a short persistence time, BMC Genomics 8 (2007) 398.
[35] A.R. Quinlan, I.M. Hall, BEDTools: a flexible suite of utilities for comparing genomic features, Bioinformatics 26 (2010) 841–842.
[36] P. Rice, I. Longden, A. Bleasby, EMBOSS: the European Molecular Biology Open Software Suite, Trends Genet. 16 (2000) 276–277.
[37] T. Lehr, J. Yuan, D. Zeumer, S. Jayadev, M.D. Ritchie, Rule based classifier for the analysis of gene-gene and gene-environment interactions in genetic association studies, BioData Min. 4 (2011) 4.
[38] E. Weitschek, G. Fiscon, G. Felici, Supervised DNA Barcodes species classification: analysis, comparisons and results, BioData Min. 7 (2014) 4.
[39] P. Bertolazzi, G. Felici, E. Weitschek, Learning to classify species with barcodes, BMC Bioinforma. 10 (Suppl. 1) (2009) S7.
[40] E. Weitschek, A. Lo Presti, G. Drovandi, G. Felici, M. Ciccozzi, M. Ciotti, et al., Human polyomaviruses identification by logic mining techniques, Virol. J. 9 (2012) 58.
[41] W.W. Cohen, Y. Singer, A simple, fast, and effective rule learner, 1999. 335–342.
[42] B.R. Gaines, P. Compton, Induction of ripple-down rules applied to modeling large databases, J. Intell. Inf. Syst. 5 (1995) 211–228.
[43] E. Frank, I.H. Witten, Generating accurate rule sets without global optimization, Proceedings of the Fifteenth International Conference on Machine Learning, 1998, pp. 144–151.
[44] J.R. Quinlan, C4.5: Programs for Machine Learning, 1993.
[45] G. Marçais, C. Kingsford, A fast, lock-free approach for efficient parallel counting of occurrences of k-mers, Bioinformatics 27 (2011) 764–770.
[46] H. Teeling, J. Waldmann, T. Lombardot, M. Bauer, F.O. Glöckner, TETRA: a web-service and a stand-alone program for the analysis and comparison of tetranucleotide usage patterns in DNA sequences, BMC Bioinforma. 5 (2004) 163.
[47] V. Brendel, J.S. Beckmann, E.N. Trifonov, Linguistics of nucleotide sequences: morphology and comparison of vocabularies, J. Biomol. Struct. Dyn. 4 (1986) 11–21.
[48] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I.H. Witten, The WEKA data mining software, ACM SIGKDD Explor. Newsl. 11 (2009) 10.
[49] S. Karlin, Global dinucleotide signatures and analysis of genomic heterogeneity, Curr. Opin. Microbiol. 1 (1998) 598–610.
[50] S. Karlin, J. Mrázek, Compositional differences within and between eukaryotic genomes, Proc. Natl. Acad. Sci. U. S. A. 94 (1997) 10227–10232.
[51] S. Karlin, C. Burge, Dinucleotide relative abundance extremes: a genomic signature, Trends Genet. 11 (1995) 283–290.
[52] T. Viturawong, F. Meissner, F. Butter, M. Mann, A DNA-centric protein interaction map of ultraconserved elements reveals contribution of transcription factor binding hubs to conservation, Cell Rep. 5 (2013) 531–545.
[53] X. Xie, T.S. Mikkelsen, A. Gnirke, K. Lindblad-Toh, M. Kellis, E.S. Lander, Systematic discovery of regulatory motifs in conserved regions of the human genome, including thousands of CTCF insulator sites, Proc. Natl. Acad. Sci. U. S. A. 104 (2007) 7145–7150.
[54] P.J. Deschavanne, A. Giron, J. Vilain, G. Fagot, B. Fertil, Genomic signature: characterization and classification of species assessed by chaos game representation of sequences, Mol. Biol. Evol. 16 (1999) 1391–1399.