ISTITUTO DI ANALISI DEI SISTEMI ED INFORMATICA
CONSIGLIO NAZIONALE DELLE RICERCHE

E. Weitschek, D. Polychronopoulos, G. Felici, and Y. Almirantis

CONSERVED NON CODING ELEMENTS CLASSIFICATION

R. 13-15 2013

Emanuel Weitschek - Institute for System Analysis and Computer Science "Antonio Ruberti" (IASI), CNR, Viale Manzoni 30, 00185 Rome, Italy. Email: emanuel.weitschek@iasi.cnr.it.

Dimitris Polychronopoulos - Institute of Biosciences and Applications, National Center for Scientific Research "Demokritos", 15310 Athens, Greece. Email: dpolychr@bio.demokritos.gr.

Giovanni Felici - Institute for System Analysis and Computer Science "Antonio Ruberti" (IASI), CNR, Viale Manzoni 30, 00185 Rome, Italy. Email: giovanni.felici@iasi.cnr.it.

Yannis Almirantirs - Institute of Biosciences and Applications, National Center for Scientific Research "Demokritos", 15310 Athens, Greece. Email: yalmir@bio.demokritos.gr.

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

Conserved non coding elements (CNEs) are non-coding DNA regions that are evolutionarily conserved in different organisms. The functionality of conserved non coding elements (CNEs) is actually unknown and large efforts have to be done for getting insights of their role in several organisms genomes.

In this work we take into consideration a particular type of CNEs analysis: their classification. CNEs classification is difficult and cannot be obtained via common alignment based techniques. Therefore, an alignment free method based on a feature vector representations of the sequences is adopted in this work. The feature vector representation is combined and given as input to rule based supervised machine learning algorithms for CNEs classification. These methodology (composition of the methods) is tested on different public available data sets obtained from up to date CNEs data bases and on CNEs of new sequenced organisms.

The classification results are sound and the reader is provided with classification models (if then rules), that are able to successfully distinguish the different functional classes present in the different datasets. The classification accuracies are compared with a state of the art sequence analysis method, the genomic signature. It is shown, that the proposed methodology has better classification performances: the distinction of CNEs and other sequences is obtained with success and with accurate and compact classification models.

1. Background

1.1. General information about CNEs

High throughput sequencing at a massive scale combined with comparative genome analyses has led to the discovery of a variety of constrained genomic elements, comparable in sequence length and populations to the known protein coding sequences. One of the most interesting discoveries that have arisen from comparative genomics among mammalian genomes are the hundreds of noncoding elements of more than 200bp in length that show absolute conservation among mammalian orders (Bejerano et al. 2004). These only represent the tip of the iceberg of a much larger class of conserved noncoding elements (CNEs), which exhibit conservation of the same order with protein-coding genes and non-coding RNAs (ncRNA).

Conserved non-expressed elements can be found in the literature with various definitions, depending on the percentage of identity between two or more organisms and the mean length (Elgar & Vavouri 2008). Increasing evidence suggests that CNEs are selectively constrained and not mutational cold-spots (Drake et al. 2006) and there is a plethora of studies indicating possible functions of those elements. Among the various reported functional roles of CNE, enhancers appear to be the most plausible (Dermitzakis et al. 2005). Nevertheless, the relative abundance, genomic distribution and variable length of CNEs is indicative of various alternatives. CNEs have been shown to bear resemblance to CTCF insulator sites (Xie et al. 2007), matrix attachment regions (Glazko et al. 2003), while a recent, concise study across 29 mammals revealed constrained elements of smaller sizes that may be directly related to transcriptional regulation as well as to the encoding of functional RNA molecules (Lindblad-Toh et al. 2011). CNE existence may be extended even further back in evolutionary time as suggested by recent works including both bony and cartilagenous fish species (Lee et al. 2011). While the majority of the analyses have been conducted in mammals, there is growing evidence that CNEs are not a vertebrate innovation and can also be found in invertebrates and plants (Vavouri et al. 2007; Lockton & Gaut 2005). Despite the fact that vertebrate and invertebrate CNEs bear no sequence similarity, they share common sequence characteristics, indicating a parallel evolution of those sequences in order to perform the same, possibly essential, functions.

When compared to non-CNEs and near promoter sequences, CNEs possess an excess of AT-rich motifs, often containing runs of identical nucleotides. In a recent paper, Walter et al have analyzed the base composition of human and Fugi CNE at single nucleotide level. They have 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 (Walter et al. 2005). Such compositional extremes are strong indications of functionality as has been shown for gene regions (Touchon et al. 2009). It is therefore of great interest to further investigate the compositional preference of constrained regions in greater detail. To this end, conventional approaches addressing composition through histograms, bag-of-words will tend to overlook the positional information, while probabilistic sequential methods like Hidden Markov Models are likely to undermine the effect of local sequence boundaries. In the following we describe a novel methodology that is able to address both such aspects.

1.2. Alignment-based and alignment-free sequences classification methods

In general, the similarity of sequences is 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 alignment, i.e. align areas of the 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 Needleman and Wunsch (Needleman and Wunsch, 1970) and Smith and Waterman (Smith and Waterman, 1981). These algorithms are computationally demanding and the complexity is exponential in the length of the sequences. Heuristics have been proposed that solve the sequence alignment problem, e.g. BLAST (Altschul et al., 1997) and FASTA (Pearson, 1990). For performing the alignment of multiple sequences more efficiently several algorithms have been proposed that address this issue: ClustalW (Thompson et al., 2002), Muscle (Edgar, 2004), Mafft (Katoh et al., 2002), and Motalign (Mokaddem and Elloumi, 2013).

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 synteny 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 was the aforementioned discovery of thousands of conserved and highly conserved non-coding genomic elements (CNEs, UCNEs etc) through various strategies. On the other hand, 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 methods 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 abundancies 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 are studied by suitable combinations of both.

Alignment free techniques have been proven to be particularly successful in phylogeny and sequence analysis (Vinga and Almeida, 2003). 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.
A promising alignment free method is based on the feature vector representation of a sequence (Kudenko and Hirsch, 1998), (Vinga and Almeida, 2003), (Kuksa and Pavlovic, 2009) and (Xing et. al 2010), 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 long k characters. 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 a single copy in the genome and have been proposed as markers for phylogenomic studies. We propose, herein, a novel methodology that does not take into account the information content of an alignment or of an overlapping DNA gene region and apply it to the classification of functional sequences of several genomes, ranging from invertebrates to vertebrates. We also proceed to several comparisons that extend the robustness of our methodology by comparing the performance of the algorithm 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 conserved noncoding elements already published in the literature.

2. Methods

2.1 The alignment-free sequences classification method

The main aim of this work is to distinguish between different classes of sequences: Conserved Non coding Elements (CNEs) from other functional types.

For accomplishing this task we propose an alignment-free method, based on:

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

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 (1998), Vinga and Almeida (2003) and Xing *et. al.* (2010), where the authors review sequence representation techniques and combine feature vector representation with supervised machine learning methods, like Support Vector Machines, for classifying biological and generic sequences. Another recent work is the one of Kuksa and Pavlovic (2009), 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.

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 = {k-mer$_1$, k-mer$_2$, ..., 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) $C = \{c_{k\text{-mer}1}, c_{k\text{-mer}2}, \ldots, c_{k\text{-mer}t}\}$.

The frequencies are then computed accordingly and stored in a vector $F = \{f_{k\text{-mer}1}, f_{k\text{-mer}2}, \ldots, f_{kmert}\}$, for a generic k-mer j the frequency is defined as $f_j = c_{kmerj} / (n\text{-}k+1)$.

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 to a given class, e.g. *Vertebrate*, *Invertebrate*; *Amniotic*, *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 approach combines the feature vector representation with rule based supervised machine learning methods.

As a technique for biological sequence analysis, rule based classifier have been proposed in (Bertolazzi et al., 2009) and in (Weitschek et al., 2012), 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 were necessary: an analysis of the characteristics nucleotides present in a determined position for every class was performed, leading to logic rules of the type, e.g. *if in 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 or state), eg. *if pos30=A and pos40=C then the sequence belongs to the squalus edmundsi species; if gene expression x is under a given threshold then the sample is healthy*. 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 following algorithms are taken into account for performing the rule based classification analysis of the CNEs sequences feature vector representations: RIPPER (Cohen, 1995), RIDOR (Gaines & Compton, 1995) and PART (Frank & Witten, 1998).

RIPER 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 extracts also 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 (Quinlan, 1993) 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 ruled based supervised learning) the 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), following computational steps:

1. The reverse complement of the sequence is computed.
2. The counts of the k-mers (for k=3,4,5) are calculated on the sequence S and on its reverse complement, obtaining the feature vector
   $C = \{c_{k\text{-}mer1}, c_{k\text{-}mer2}, \ldots, c_{k\text{-}mert}\}$.
   The k-mer counts are extracted with the Jellyfish software (Marçais and Kingsford, 2011).
   k has been chosen between 3 and 5, based on the following references (Teeling et al. 2004a), (Pride et al., 2003), (Teeling et al. 2004b) , 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 (A,C,G,T) alphabet.
3. The frequencies of each k-mer are computed: $f_j = c_{kmerj} / (n\text{-}k+1)$.
4. A feature frequency vector of each sequence is obtained:
   $F = \{f_{k\text{-}mer1}, f_{k\text{-}mer2}, \ldots, f_{kmert}\}$.
   The feature vectors are combined in a data matrix, where each column represents a sequence and each row the k-mer frequency, e.g.

	Seq 1	Seq2	….	Seq3	Seq4	…
	Vertebrate	**Vertebrate**	….	**Invertebrate**	**Invertebrate**	…
AAAA	0.46	0.26	…	0.24	0.54	…
AAAC	0.12	0.16	…	0.23	0.24	…
AAAG	0.123	0.23213	…	0.2312	0.232	…
….	…	…	…	….	…	

5. The obtained data matrix is given as input to three ruled based supervised machine learning algorithms: RIPPER, RIDOR, and PART.
6. 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; the discretization procedure improves the performance of the rule based algorithms.
7. 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 k disjoints sets, the data mining procedure is run k times with different sets. At a generic run k the k subset is used as test set and the remaining k-1 sets are merged and used as training set for building the model. Every of the k sets contains a random distribution of the data. The cross validation sampling procedure builds k models and each of this model 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.
8. The classification models are extracted and evaluated in terms of correct classification rate, e.g. *if freq(AAAC)<0.195 then the organism is Vertebrate; if freq(AAAC)>0.195 then the organism is Invertebrate*.

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 (Hall et al., 2009) 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.

2.2 The "genomic signature" method

In addition another sequences classification technique has been considered for comparing the obtained results of our approach. 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 (Burge et al. 1992). 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 (Karlin and Mrazek, 1997). Thus they have assigned to the vector of these ten "first neighbour preferences" the name of genomic signature (Karlin and Burge, 1995). In what follows, we use classification based on genomic signatures for comparison with classification based on our alignment-free method described in the previous section.

2.3 CNEs sequences extraction and collection

In the present analysis we have selected coordinates of conserved noncoding elements that are already available in the literature (Kim & Pritchard 2007; Bejerano et al. 2004; Stephen et al. 2008; Vavouri et al. 2007; Glazov et al. 2005; Mattick 2009). A very useful suite of tools, along with several of its features, called BEDTools was used (Quinlan & Hall 2010) in order to extract FASTA sequences from BED files and for other purposes. In addition, we have applied the EMBOSS suite to calculate fractional GC content of sequences (Rice et al. 2000).

3. Results and discussion

In this section we describe the performed experiments and the obtained results. For every classification task we adopted the technique of feature vector representation and rule based supervised classification exposed in previous sections. Every classification task was performed using feature vector representations with k-mers of length 3-4-5, which are given as input to three different rule based algorithms (RIPPER, RIDOR, PART), adopting a 10-fold cross validation sampling technique. The accuracy results are summarized in Table 1 by reporting a mean accuracy for each different rule based algorithm applied on the feature vectors composed of k-mers length 3-4-5. Finally the results are compared with the established genomic signature method (Table 3).

3.1 Comparison of human background versus worm background sequences (Interspecies, background genomic compositions)

As a first step, we explored whether our method is able to effectively distinguish between sequences of different genomes that are supposed to be non-functional. For that purpose, we have chosen from the human genome sequences that are not CNEs, not repeats and not exons. We have done the same for the yeast genome. Consideration was taken in order for the sequences of each category to be of the same average length. The sequences were classified correctly at a rate of 94,16% **(see EXP 1, Table 1)**. The algorithm giving the best classification rates is PART

EXP	JRIP	RIDOR	PART	AVG	
[1]	93.37%	92.83%	*94.16%*	93.45%	human bg vs worm bg
[2]	*88.09%*	86.18%	88.04%	87.44%	worm exons vs bg
[3]	*89.50%*	89.26%	87.32%	88.69%	human exons vs bg
[4]	*80.69%*	79.52%	77.46%	79.22%	worm CNEs vs bg
[5a]	*71.95%*	71.41%	76.70%	73.35%	human UCEs vs bg
[5b]	*71.31%*	67.73%	69.90%	69.65%	EU100 nonexonic CNEs vs bg
[5c]	*92.37%*	92.21%	88.31%	90.96%	amniotic vs bg
[5d]	98.98%	*99.05%*	98.56%	98.86%	mammalian vs bg
[6]	*98.35%*	98.32%	97.70%	98.12%	worm exons vs worm CNEs
[7]	91.20%	90.06%	*92.70%*	91.32%	human exons vs EU100 nonexonic CNEs
[8]	93.89%	*93.93%*	91.20%	93.01%	amniotic CNEs vs mammalian CNEs
[9]	83.58%	81.43%	*85.51%*	83.51%	worm exons vs human exons
[10]	*93.88%*	93.48%	94.87%	94.08%	worm CNEs vs human CNEs
AVG	**88.2%**	**87.3%**	**87.9%**	**87.8%**	

Table 1: Classification rates of genomic elements as evidenced by using our approach. Note that in almost every case background (bg) is different.
EXP: experiment. AVG: average.

3.2 Comparison of constrained sequences versus background sequences of different organisms (Intraspecies, functional sequences with the corresponding background)

3.2.1 Case of exons

As a second step, we investigated whether our approach could help towards discriminating functional (or potentially functional as is the case of CNEs which we discuss below) versus background sequences found in the same genome. More specifically we have considered as examples the human and the worm genomes. For the corresponding genomes, we have obtained exons of the same average length and GC content and compared them with background sequences (of the same genome). In this case, the background sequences that are found naturally in the genome, are non-exonic, non-repetive and of the same average length and GC content with the exons. The results are summarized in **EXP 2 & 3, Table 1**. As it is evidenced we get very good results, 87,44% and 88,69% for the worm and human genome respectively

3.2.2 Case of CNEs

We extend our analysis to different classes of conserved noncoding elements of various origin. More specifically, we consider the following comparisons:

(a) worm CNEs versus background sequences **(EXP 4, Table 1)**: These sequences are identified from pairwise alignments of *C.elegans/C.briggsae* (Vavouri et al. 2007) and mapped on the worm genome. The mean length of those CNEs is considerably shorter than those of vertebrates. We have compared them versus sequences specifically chosen in the worm genome such that they are CNE-like (same GC and average length, nonrepetitive and not CNE). For this comparison, we get a classification rate of 79,22%

(b) human UCEs versus background sequences **(EXP 5a, Table 1):** These sequences display 100% identity between human, mouse and rat genomes (Bejerano et al. 2004) and are mapped on the human genome. We have classified those elements comparing them with background sequences of similar GC content and average length found naturally in the human genome. The average classification rate we got was 73.35% with a maximum of 76.7%

(c) EU100 nonexonic CNEs versus background sequences **(EXP 5b, Table 1):** These are CNEs mapped on the human genome (hg18), of various lengths, that are identical over at least 100bp in at least 3 of 5 placental mammals (human, mouse, rat, dog and cow) (Stephen et al. 2008). The whole set is named EU100+ and since we removed elements overlapping exons, we named it EU100 nonexonic. An average of 69.65% classification rate was obtained by comparing those elements with their corresponding background sequences

(d) Amniotic and Mammalian CNEs versus background sequences **(EXPs 5c&5d, Table 1):** Mammalian CNEs (conserved within mammals but not found in chicken or fish) and Amniotic CNEs (conserved in mammals and chicken but not found in fish), mapped on the human genome (hg17) (Kim & Pritchard 2005). In this case we obtained classification rates of the order of 90.96% and 98.86% respectively.

3.3 Comparison of constrained sequences (Intraspecies, functional sequences with other functional sequences)

The aim of this set of comparisons was to validate the sensitivity of our approach towards constrained, functional sequences. For that purpose, we have examined the following cases:

(a) worm exons versus worm CNEs **(EXP 6, Table 1):** Exons and CNEs obtained from the datasets mentioned above were compared. The classification rate was again high, 98,12% on average

(b) human exons vs EU100 nonexonic CNEs **(EXP 7, Table 1):** The average rate of classification here was 91.32%

(c) amniotic CNEs vs mammalian CNEs **(EXP 8, Table 1):** To further examine the potential of our method in distinguishing CNEs and rule out the possibility that the observed results are due to differences in GC content, we apply our methodology to sequences that are mapped on the same genome (hg17), but are characterized by different evolutionary stages, i.e. Amniotic vs Mammalian (Kim & Pritchard 2007). These are non-overlapping elements characterized by the same GC content (38,32% and 40,16% respectively). We have chosen to include equal number of sequences (1000) from the two original datasets that are of the same average length in order to limit the biases. The discriminative power of the best classifier is 94%

3.4 Comparison of constrained sequences (Interspecies, functional sequences with other functional sequences)

(a)worm exons versus human exons **(EXP 9, Table 1):** When we compare worm exons versus human exons of the same average length, we get an average classification rate of 83.51%

(b)worm CNEs versus human CNEs **(EXP 10, Table 1):** The purpose of this analysis was to prove that with the k-mer frequency analysis on CNEs it is possible to distinguish sequences belonging to vertebrate and invertebrate organisms.

An example of logic rules is provided in table 2 for amniotic CNEs vs mammalian CNEs sequences classification, e.g. if a sequence satisfies the logic rule provided in the table than it is assigned to the amniotic class, else to the mammalian class.

Amniotic
191.2055<freq(CAATC)≤302.1175 OR freq(CGCCG)≤151.286 OR freq(ACGCG)>301.2155 AND freq(GCGAC)>302.1175 AND freq(ATATA)>302.1175 OR 191.2055<freq(CGAAT)≤305.344 OR freq(ATATT)>604.7545 AND 302.5725<freq(ATACA)≤382.044 AND 302.1175<freq(CTGGA)≤382.4105 AND 151.0585<freq(CTGGA)≤191.0225

Table 2: RIPPER logic formulas for Amniotic vs Mamalian,
the sequences are classified as Amniotic if recognized by this set of clauses,
else as Mamalian, frequencies are given as base of 10^{-5}

3.5 Comparison of k-mer classification with genomic signatures

In order to assess the performance of our classification, we compared our results with those obtained from a standard method of sequence comparison, namely the genomic signature approach, proposed by (Karlin & Mrázek 1997). Genomic signatures constitute the most widely used approach to assess compositional preferences and it has been shown that it is able to efficiently reconstitute phylogenetic differences in oligonucleotide usage. Genomic signatures were calculated at the level of dinucleotides (programs in Perl are available upon request) and an identical process of classification (using the supervised rule based classification algorithms RIPPER, RIDOR, PART) of the sequences represented as distance matrices was performed. The classification efficiency with the use of the genomic signatures was shown to be much inferior to the one obtained with our proposed approach. Table 3 and provide more information and an overview of the classification results with the genomic signatures.

EXP	JRIP	RIDOR	PART	AVERAGE	DESCRIPTION
[1]	82.099 %	80.1901 %	80.6403 %	**80.97 %**	human bg vs worm bg
[2]	63.4336 %	60.5761 %	64.0144 %	**62.67 %**	worm exons vs bg
[3]	55.8779 %	54.2771 %	55.3277 %	**55.16 %**	human exons vs bg
[4]	57.6894 %	58.5896 %	53.7134 %	**56.66 %**	worm CNEs vs bg
[5a]	69.719 %	63.9958 %	68.1582 %	**67.29 %**	human UCEs vs bg
[5b]	65.7329 %	61.2306 %	65.5328 %	**64.16 %**	EU100 nonexonic CNEs vs bg
[5c]	61.5831 %	58.628 %	58.9974 %	**59.73 %**	amniotic vs bg
[5d]	53.0343 %	51.5567 %	49.657 %	**51.41 %**	mammalian vs bg
[6]	58.0046 %	56.4037 %	56.0892 %	**56.83 %**	worm exons vs worm CNEs
[7]	65.4827 %	61.3807 %	63.6318 %	**63.50 %**	human exons vs EU100 nonexonic CNEs
[8]	53.5092 %	53.7731 %	52.8232 %	**53.37 %**	amniotic CNEs vs mammalian CNEs
[9]	72.2827 %	70.126 %	72.7952 %	**71.73 %**	worm exons vs human exons
[10]	74.7874 %	71.936 %	73.3367 %	**73.35 %**	worm CNEs vs human CNEs

Table 3: Classification rates of genomic elements as evidenced by using the genomic signature approach, EXP: experiment. Note that in almost every case background (bg) is different

4. Conclusions

All the cases where GS perform better consist intra-species comparisons as expected . All the cases where LM perform better are the ones containing human sequences

The above conclusions seem to influence the numerical values obtained for the classification rates as well. GS method yields lower values in intra-species comparisons and high in inter-species comparisons.

The only cases where GS method performs above 70% in intra-species comparisons are met when different functionalities are involved (Experiments [7]).

Experiments containing CNEs versus their corresponding surrogates in the human genome display the highest rates of sequence classification using our method, LM (Experiments 5a – 5d). This is compatible with the highest information content hypothesis for human CNEs based on their conservation.

The same applies in the case of exons versus their corresponding surrogates for different species (compare Experiments [2] with [3]).

Acknowledgements

This work is partially supported by the Italian PRIN GenData 2020 and the Flagship Project InterOmics. EW and DP would like to express their gratitude to Fredj Tekaia, Giuseppe D'Onofrio and EMBO for the organization of the 2013 EMBO Practical Course:

Bioinformatics and Comparative Genome Analyses in Naples, which enabled this collaboration

References

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic acids research, 25(17), 3389-3402.

Bejerano G et al. 2004. Ultraconserved elements in the human genome. Science. 304:1321–1325. doi: 10.1126/science.1098119.

Bertolazzi P, Felici G, Weitschek E (2009) Learning to classify species with barcodes. *BMC Bioinformatics*, 10:1-12.

Burge C, Campbell AM, Karlin S (1992). Over- and under-representation of short oligonucleotides in DNA sequences. PNAS 89, 1358-62.

Boros E, Ibaraki T, and Makino K (1999) Logical analysis of binary data with missing bits. Artificial Intelligence, 107:219–263.

Cohen W (1995) Fast effective rule induction. Proceedings of the Twelfth International Conference on Machine Learning, 115–123, Morgan Kaufmann.

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic acids research, 32(5), 1792-1797.

Frank E and Witten I H (1998) Generating accurate rule sets without global optimization. In Proceedings of the 15th. International Conference on Machine Learning, Morgan Kaufmann.

Felici G and Truemper K (2002) A minsat approach for learning in logic domains. INFORMS Journal on Computing, 13(3):1–17

Gaines B and Compton P. (1995) Induction of ripple-down rules applied to modeling large databases. Journal of Intelligent Information Systems.

Glazov EA, Pheasant M, McGraw EA, Bejerano G, Mattick JS. 2005. Ultraconserved elements in insect genomes: a highly conserved intronic sequence implicated in the control of homothorax mRNA splicing. Genome Res. 15:800–808. doi: 10.1101/gr.3545105.

Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P and Witten I H (2009) The weka data mining software:an update.; SIGKDD Explor. Newsl., 11(1):10–18.

Karlin S, Burge C (1995). Dinucleotide relative abundance extremes: a genomic signature, TIG, 11, 283-290.

Karlin S, Mrazek J (1997). Compositional differences within and between eukaryotic genomes. PNAS, 94, 10227-32.

Katoh, K., Misawa, K., Kuma, K. I., & Miyata, T. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic acids research, 30(14), 3059-3066.

Katzman S et al. 2007. Human genome ultraconserved elements are ultraselected. Science. 317:915. doi: 10.1126/science.1142430.

Kim SY, Pritchard JK. 2007. Adaptive evolution of conserved noncoding elements in mammals. PLoS genetics. 3:1572–86. doi: 10.1371/journal.pgen.0030147.

Kudenko, D. and Hirsh, H. (1998). Feature generation for sequence categorization. In *AAAI/IAAI* (pp. 733-738).

Kuksa P and Pavlovic V (2009) Efficient alignment-free DNA barcode analytics; BMC Bioinformatics, 10(Suppl 14):S9.

Marçais G and Kingsford C (2011) A fast, lock-free approach for efficient parallel counting of occurrences of k-mers; Bioinformatics, 27(6): 764–770.

Mattick JS. 2009. Deconstructing the dogma: a new view of the evolution and genetic programming of complex organisms. Ann N Y Acad Sci. 1178:29–46. doi: 10.1111/j.1749-6632.2009.04991.x.

Mokaddem A and Elloumi M (2013) Motalign: A Multiple Sequence Alignment Algorithm Based on a New Distance and a New Score Function; DEXA - Database and Expert Systems Applications Workshops, 81-84

Needleman, S. B., & Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. Journal of molecular biology, 48(3), 443-453.

Pearson, W. R. (1990). [5] Rapid and sensitive sequence comparison with FASTP and FASTA. Methods in enzymology, 183, 63-98.

Pride, D. T., Meinersmann, R. J., Wassenaar, T. M., & Blaser, M. J. (2003). Evolutionary implications of microbial genome tetranucleotide frequency biases. Genome research, 13(2), 145-158.

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 26:841–842. doi: 10.1093/bioinformatics/btq033.

Quinlan R (1993). *C4.5: Programs for Machine Learning*. San Mateo, CA: Morgan Kaufmann Publishers.

Smith, T. F., & Waterman, M. S. (1981). Identification of common molecular subsequences. Journal of molecular biology, 147(1), 195-197.

Stephen S, Pheasant M, Makunin IV, Mattick JS. 2008. Large-scale appearance of ultraconserved elements in tetrapod genomes and slowdown of the molecular clock. Mol Biol Evol. 25:402–408. doi: 10.1093/molbev/msm268.

Teeling, H., Meyerdierks, A., Bauer, M., Amann, R., & Glöckner, F. O. (2004). Application of tetranucleotide frequencies for the assignment of genomic fragments. *Environmental microbiology*, *6*(9), 938-947.

Teeling, H., Waldmann, J., Lombardot, T., Bauer, M., & Glöckner, F. O. (2004). TETRA: a web-service and a stand-alone program for the analysis and comparison of tetranucleotide usage patterns in DNA sequences. *BMC bioinformatics*, *5*(1), 163.

Thompson, J. D., Gibson, T., & Higgins, D. G. (2002). Multiple sequence alignment using ClustalW and ClustalX. Current protocols in bioinformatics, 2-3.

Vavouri T, Walter K, Gilks WR, Lehner B, Elgar G. 2007. Parallel evolution of conserved non-coding elements that target a common set of developmental regulatory genes from worms to humans. Genome Biol. 8:R15. doi: 10.1186/gb-2007-8-2-r15.

Vinga S., Almeida J. (2003): Alignment-free sequence comparison - a review, Bioinformatics 19 (4), 513-523

Weitschek, E., Presti, A. L., Drovandi, G., Felici, G., Ciccozzi, M., Ciotti, M., & Bertolazzi, P. (2012). Human polyomaviruses identification by logic mining techniques. Virology journal, 9(1), 1-6.

Xing, Z., Pei, J., & Keogh, E. (2010). A brief survey on sequence classification. *ACM SIGKDD Explorations Newsletter*, *12*(1), 40-48.