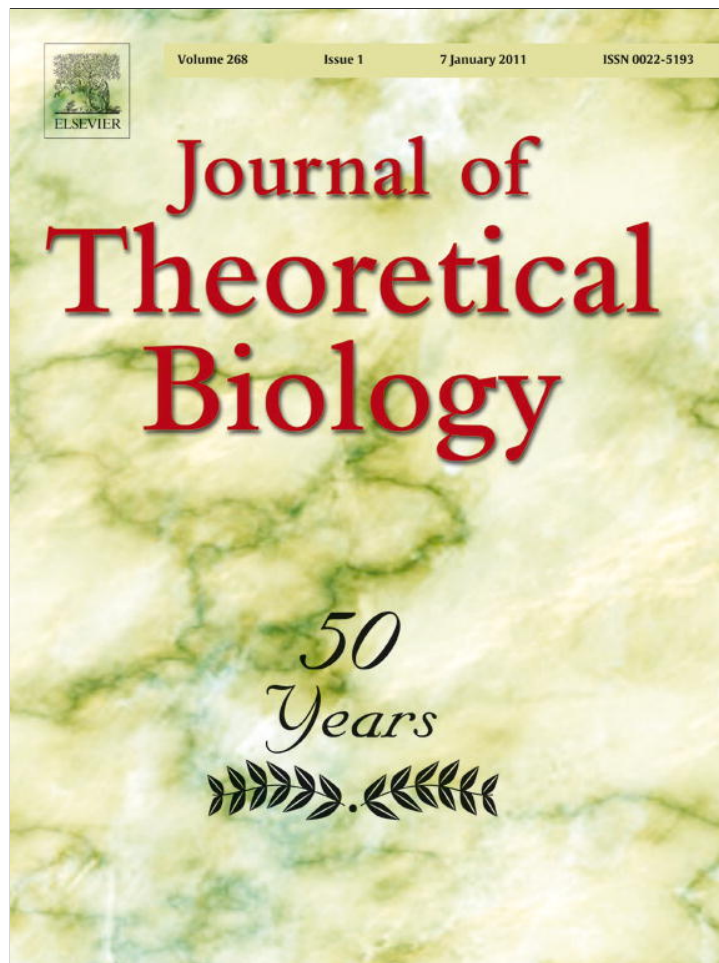


Provided for non-commercial research and education use.
Not for reproduction, distribution or commercial use.



(This is a sample cover image for this issue. The actual cover is not yet available at this time.)

This article appeared in a journal published by Elsevier. The attached copy is furnished to the author for internal non-commercial research and education use, including for instruction at the authors institution and sharing with colleagues.

Other uses, including reproduction and distribution, or selling or licensing copies, or posting to personal, institutional or third party websites are prohibited.

In most cases authors are permitted to post their version of the article (e.g. in Word or Tex form) to their personal website or institutional repository. Authors requiring further information regarding Elsevier's archiving and manuscript policies are encouraged to visit:

<http://www.elsevier.com/copyright>

Contents lists available at [SciVerse ScienceDirect](http://SciVerse.ScienceDirect.com)

Journal of Theoretical Biology

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

Clustering of DNA words and biological function: A proof of principle

Michael Hackenberg^{a,b,*}, Antonio Rueda^{a,b}, Pedro Carpena^c, Pedro Bernaola-Galván^c, Guillermo Barturen^{a,b}, José L. Oliver^{a,b,*}^a Dpto. de Genética, Facultad de Ciencias, Universidad de Granada, Campus de Fuentenueva s/n, 18071-Granada, Spain^b Lab. de Bioinformática, Centro de Investigación Biomédica, Campus de la Salud, Avda. del Conocimiento s/n, 18100-Granada, Spain^c Dpto. de Física Aplicada II, E.T.S.I. de Telecomunicación, Universidad de Málaga, 29071-Málaga, Spain

ARTICLE INFO

Article history:

Received 17 October 2011

Received in revised form

20 December 2011

Accepted 21 December 2011

Available online 30 December 2011

Keywords:

DNA-words

Word clustering

Enrichment/depletion experiments

Biological function

ABSTRACT

Relevant words in literary texts (key words) are known to be clustered, while common words are randomly distributed. Given the clustered distribution of many functional genome elements, we hypothesize that the biological text per excellence, the DNA sequence, might behave in the same way: k -length words (k -mers) with a clear function may be spatially clustered along the one-dimensional chromosome sequence, while less-important, non-functional words may be randomly distributed. To explore this linguistic analogy, we calculate a clustering coefficient for each k -mer ($k=2-9$ bp) in human and mouse chromosome sequences, then checking if clustered words are enriched in the functional part of the genome. First, we found a positive general trend relating clustering level and word enrichment within exons and Transcription Factor Binding Sites (TFBSs), while a much weaker relation exists for repeats, and no relation at all exists for introns. Second, we found that 38.45% of the 200 top-clustered 8-mers, but only 7.70% of the non-clustered words, are represented in known motif databases. Third, enrichment/depletion experiments show that highly clustered words are significantly enriched in exons and TFBSs, while they are depleted in introns and repetitive DNA. Considering exons and TFBSs together, 1417 (or 72.26%) in human and 1385 (or 72.97%) in mouse of the top-clustered 8-mers showed a statistically significant association to either exons or TFBSs, thus strongly supporting the link between word clustering and biological function. Lastly, we identified a subset of clustered, diagnostic words that are enriched in exons but depleted in introns, and therefore might help to discriminate between these two gene regions. The clustering of DNA words thus appears as a novel principle to detect functionality in genome sequences. As evolutionary conservation is not a prerequisite, the proof of principle described here may open new ways to detect species-specific functional DNA sequences and the improvement of gene and promoter predictions, thus contributing to the quest for function in the genome.

© 2011 Elsevier Ltd. All rights reserved.

1. Introduction

The relevant words (key words) of a human-written, literary text are clustered along the text, while common words are randomly distributed (Ortuño et al., 2002; Zhou and Slater, 2003). Our group (Carpena et al., 2009) has proven that this principle is successful in detecting semantic meaning (keyword extraction) in conventional literary texts, as well as in comma-less texts (i.e. texts from which all spaces and punctuation marks have been removed). Since DNA is the comma-less text par excellence, we hypothesize that the method might be also useful

to uncover meaning (i.e. association to biological function) in DNA sequences. This linguistic analogy is supported by the fact that genes (Durand and Sankoff, 2003; Kendal, 2004; Neel, 1961), CpG dinucleotides (Bird, 1986; Hackenberg et al., 2006), Transcription Factor Binding Sites (TFBSs, Berman et al., 2002; Boeva et al., 2007) and 3D structural motifs in ribosomal RNA (Sargsyan and Lim, 2010) are all known to occur in clusters.

A combination of computational and experimental approaches are currently used for the genome-wide elucidation of the genome regulatory code, which may include promoters, enhancers, and repressor elements, along with structural components like origins of replication, centromeric sequences or boundary elements. The underlying principles in most of these methods are evolutionary conservation, motif over-representation and gene co-regulation, as well as combinations among them. However, each of these criteria to predict biological function presents some shortcoming. First, the filtering power of evolutionary

* Corresponding authors at: Dpto. de Genética, Facultad de Ciencias, Universidad de Granada, Campus de Fuentenueva s/n, 18071-Granada, Spain.
Tel.: +34 958243261; fax: +34 958244073.

E-mail addresses: mlhack@gmail.com (M. Hackenberg), oliver@ugr.es (J.L. Oliver).

conservation to identify putative regulatory regions (Carmack et al., 2007; Lindblad-Toh et al., 2011; Siddharthan and Nimwegen, 2007; Sinha et al., 2004; Wang and Stormo, 2003) is challenged by (i) the finding that about half of the functional elements found in non-coding DNA from the ENCODE regions are not evolutionarily conserved (Birney et al., 2007), (ii) the demonstration that conservation of motifs alone cannot predict transcription factor binding (Wilson et al., 2008), and (iii) the recent finding that half of functional sequences is specific to individual eutherian lineages (Ponting et al., 2011). Second, the use of over-represented motifs (Bussemaker et al., 2000; Chakravarty et al., 2007; Eden et al., 2007; Hughes et al., 2000; Li and Wong, 2005; Sinha et al., 2004; Wang et al., 2005) to identify regulatory candidates assumes that motifs are used independently, thus ignoring that the spatial relationship among motifs (i.e. the architecture of the genetic text) is also biologically relevant (Brown et al., 2007). Lastly, the difficulties in choosing an appropriate similarity measure to group genes according to their expression profiles, the uncertainty on the number of clusters to form (Wang et al., 2005), and the combinatorial nature of transcriptional regulation (Han and Zhao, 2009), implying that the same motif may appear in the promoters of genes that express or function differently, cast also doubts on the general validity of the co-regulation principle. On the contrary, as an *ab initio* method in which training is not necessary, the clustering principle could uncover both known and new functions. Furthermore, it can be applied to individual genomes (i.e. no comparison with other related genomes is required), thus being able to uncover phylogenetically conserved as well as species-specific functional elements. In this way, by adding to current methods, the clustering principle could significantly contribute to the quest for function in genome sequences.

Some locally clustered words have been already related to biological function (FitzGerald et al., 2004; Salisbury et al., 2006). These studies and other related ones are based on the local clustering properties within a given genomic region (promoter, 3' flank) and depend therefore on the existence of a suitable genome annotation. To our knowledge, no method has been applied to characterize the global clustering of DNA words (*k*-mers) in non-annotated genome sequences, then linking the clustering to biological function.

A great deal of work has also been done analyzing the frequencies of DNA words (Arnau et al., 2008; Hampson et al., 2002; Herold et al., 2008; Nussinov, 1981; Subirana and Messegue, 2010), and DNA dictionaries based on word frequencies (Trifonov and Brendel, 1986; Tsonis et al., 1997) have been derived. Indeed, many gene-finding algorithms are based on word frequencies (Burge and Karlin, 1997; Guigo et al., 1992; Staden and McLachlan, 1982). In the same way, methods to identify regulatory regions are based on the over-representation of some DNA words or motifs (Bussemaker et al., 2000; Chakravarty et al., 2007; Eden et al., 2007; Hughes et al., 2000; Li and Wong, 2005; Sinha et al., 2004; Wang et al., 2005). However, all these word-counting approaches assume either independence among different words, or at best uniform dependencies (i.e. hidden Markov models). Such simplifying assumptions ignore more complex spatial relationship among the words (i.e. the architecture of the genetic text) that may be biologically meaningful (Brown et al., 2007). Our approach takes into account not only the word frequencies of the genetic text (as do the word-counting methods), but also the text structure (i.e. the spatial distribution of words along the genome sequence).

By extrapolating from our linguistic approach (Carpena et al., 2009), we develop here a method to compute word-clustering in DNA genome sequences. To explore its biological implications, we first calculate the clustering level for each *k*-mer ($k=2, 9$) in

human and mouse chromosome sequences, then presenting experimental results supporting the linguistic analogy. We then show that a link exists between DNA word-clustering and the enrichment of the corresponding words in functional genome elements.

2. Materials and methods

2.1. DNA sequences

The genome assemblies for human (hg18) and mouse (mm8) were downloaded from the UCSC server (Karolchik et al., 2008).

The following annotations for genome elements were used: exons and introns from the RefSeq database (Pruitt et al., 2007), cisRED TFBSs (Robertson et al., 2006), TFBS conserved sites (Weirauch and Raney, 2011), DNA repeats detected by RepeatMasker (Jurka et al., 2005; Smit et al., 1996–2010), CpG islands (Hackenberg et al., 2006, 2010), phylogenetically conserved elements or PhastCons (Siepel et al., 2005), and insulators (Bao et al., 2008).

2.2. Measuring *k*-mer clustering in DNA sequences

The clustering of a particular *k*-length word (*k*-mer) within the one-dimensional DNA sequence can be measured by considering the distance distribution (*d*) between consecutive copies of the same *k*-mer (Carpena et al., 2009). In particular, when a word appears with probability *p* and is distributed randomly along the sequence, the distance distribution *P*(*d*) follows the geometric distribution:

$$P(d) = (1-p)^{d-1}p \quad (1)$$

However, when the distribution of a word is not random, *P*(*d*) can present different forms. To quantify this distribution with a single parameter, we can use either the standard deviation of *P*(*d*):

$$\sigma = \sqrt{d^2 - \bar{d}^2} \quad (2)$$

or the coefficient of variation:

$$CV = \sigma/\bar{d} \quad (3)$$

A high CV for a particular word indicates clustering, since *P*(*d*) presents large fluctuations as compared to the mean distance. However, since two randomly distributed words with different frequency of appearance (*p*) would produce different CV values (Carpena et al., 2009), we define the normalized clustering measure σ_{nor} as

$$\sigma_{nor} = \frac{\sigma/\bar{d}}{\sqrt{1-p}} \quad (4)$$

Defined in this way, σ_{nor} estimates the clustering of a word without being biased by its particular frequency. In computing this measure, we only take into account non-overlapping copies of each word. A value $\sigma_{nor} = 1$ means that the fluctuations of the distances are similar to the average distance, thus indicating randomness. When $\sigma_{nor} > 1$ the fluctuations of the distances are larger than the mean, i.e. there are more small and large distances than expected, which is the signature of clustering: the higher the σ_{nor} value, the stronger the clustering. In contrast, when $\sigma_{nor} < 1$ the fluctuations are smaller than the mean and this indicates 'repulsion' between the different instances (or copies) of the word.

2.3. Enrichment/depletion analyses

To quantify the association between the *k*-mers and a given genome element, we define the word enrichment as the percentage of words occurring twice as frequent within the genome element as

expected by chance alone. To calculate this percentage, we first define the k -mer density within the genome element as

$$Den_i^{in} = \frac{n_i^{in}}{Len^{in}} \quad (5)$$

being n_i^{in} the number of instances (copies) of the i th word for a given k completely located within any of the different copies of the genome element, and Len^{in} the length sum of all instances of the genome element under analysis. In the same way, we define the density outside the genome element as

$$Den_i^{out} = \frac{n_i^{out}}{Len^{out}} \quad (6)$$

being n_i^{out} the number of instances (copies) of the i th word for a given k completely located outside the genome element, i.e. not overlapping in a single base with the genome element, and Len^{out} the length of the sequence not occupied by the genome element, i.e. the number of bases of the whole genome minus Len^{in} .

Then, the enrichment ratio r_i for each k -mer is defined as

$$r_i = \frac{Den_i^{in}}{Den_i^{out}} \quad (7)$$

This ratio can be easily interpreted:

- $r_i = 1$: The instances of the k -mer are distributed randomly, i.e. occur with the same probability in- and outside the genome element.
- $r_i > 1$: The k -mer is enriched inside the genome element.
- $r_i < 1$: The k -mer is depleted inside the genome element.

The ratio r_i is a property of each individual k -mer for a given chromosome and genome element. In order to depict the general tendency of r_i as a function of σ_{nor} , we ordered all k -mers by its σ_{nor} value dividing them into 20 bins with the same number of k -mers in each bin (equal frequency). In each bin, we calculate the 'enrichment percentage' as 100 times the number of words with $r_i \geq 2$ (twice more frequent inside the element than expected) divided by the number of k -mers in the bin.

2.4. Statistical significance: randomization tests

The enrichment percentage can be calculated rather fast and gives a good overview on the association between k -mers and genome elements as a function of σ_{nor} . However, it does not include a statistical significance. To associate a p -value to the word enrichment percentage, we perform randomization experiments for a given k -mer in the following way:

- We first detect the number of instances of a given k -mer inside the genome element n_{obs} .
- We then randomize 1000 times the genome location of all k -mer instances and detect the number of instances inside the genome elements n_j (j goes from 1 to 1000). Specifically, we generate random distributions of the k -mer positions in order to obtain the expected mean and standard deviation of the degree of overlap between a given randomly distributed k -mer and a functional element. The null hypothesis is therefore that no relation between k -mer and functional element exists, and we can infer enrichment or depletion if the observed overlap differs significantly from the expected overlap. No explicit probability models are needed in this kind of randomization experiments. Note that we take the N-runs (or islands of N's) of chromosome sequences into account, i.e. a random location is discarded when it overlaps an N-run.
- Finally, we calculate the mean and the standard deviation of the random distribution, then assigning a z-score and a p -value

to each k -mer. The p -value is assigned using a two-sided test, i.e. we can check if the k -mer is significantly enriched or depleted in the genome element.

Given that these experiments are computationally demanding, we calculate these p -values for only two word sets: (1) the top-200 words with the highest σ_{nor} in each chromosome (case) and (2) the 200 most randomly distributed (non-clustered) words for each chromosome (control). Bonferroni correction for multiple testing was applied to obtain the final p -values.

The enrichment/depletion analyses described in this work were implemented in Java connecting to a MySQL backend in order to store the vast amount of data. Statistical significance was confirmed by means of HyperBrowser (Sandve et al., 2010), a stand-alone application tightly connected to the Galaxy platform (Giardine et al., 2005).

3. Results

3.1. Word clustering in human and mouse genomes

Using Eq. (4) (see Section 2), we first computed the clustering level σ_{nor} for all k -mers ($k=2-9$) that have at least 3 copies in the chromosome. The clustering levels of the k -mers are calculated for each chromosome separately. Table 1 show the basic statistic for σ_{nor} in human and mouse genomes, as a function of word-length (k). A more detailed word clustering statistics by chromosome is shown on the web supplement: <http://bioinfo2.ugr.es/DNAkeywords/>. The distribution of σ_{nor} values in chromosome sequences shows an increasing positive skewness as word-length (k) increases (Fig. 1). Note that the long right-tails differentiating DNA from a random distribution are especially evident from $k=6$ onwards.

3.2. The clustering of a word is related to its enrichment within functional elements

To investigate the link between DNA word clustering and biological function, we first analyzed the fraction of word copies within a given genome element as a function of clustering level. Using the measure of 'word enrichment percentage' (see Section 2), we analyzed how the fraction of word copies within a given genome element varies with word-clustering level (σ_{nor}). To this end, we

Table 1
Word clustering (σ_{nor}) as a function of word-length (k) in human (hg18) and mouse (mm8) genomes.

Genome	Word length (k)	N	Mean	Min.	Max.
Human	2	384	1.258	1.078	1.675
	3	1536	1.259	1.034	1.778
	4	6144	1.239	1.001	2.258
	5	24,576	1.210	0.999	2.517
	6	98,304	1.181	0.905	2.923
	7	393,214	1.156	0.712	4.947
	8	1,571,592	1.131	0.001	10.320
	9	6,175,916	1.097	0.000	24.873
	Mouse	2	336	1.276	1.154
3		1344	1.244	1.050	1.655
4		5376	1.212	1.034	2.217
5		21,504	1.173	0.806	2.485
6		86,010	1.147	0.559	3.131
7		343,442	1.132	0.085	3.202
8		1,359,015	1.119	0.000	3.217
9		5,285,463	1.090	0.000	7.436

plotted the mid-value of the word clustering level from 20 bins, against the percentage of words which are enriched within the particular genome element in the given bin (Fig. 2). When the

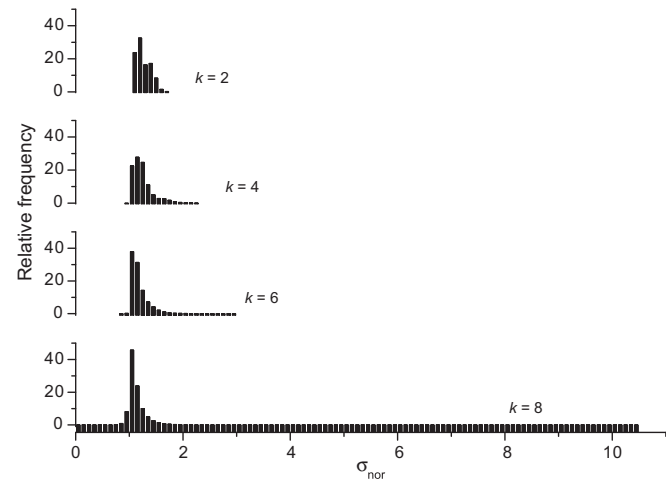


Fig. 1. Observed distributions of σ_{nor} for k -mers ($k=2-8$) in the human genome.

clustering level was not binned on the X axes, the scatterplots (not shown) allow also confirming this trend. Enriched, depleted and random words show different behaviors in these plots, making them more difficult to interpret; therefore, we opted for the enrichment percentage. The plots in Fig. 2 show a positive general trend relating clustering level and word enrichment within the RefSeq (Pruitt et al., 2007) exons (Fig. 2a) and the cisRED (Robertson et al., 2006) TFBSs (Fig. 2b) in both human and mouse chromosomes. Noteworthy, a much weaker relation was found for the DNA repeats (Fig. 2c) detected by RepeatMasker (Jurka et al., 2005; Smit et al., 1996–2010) and no relation at all exist for the RefSeq (Pruitt et al., 2007) introns (Fig. 2d). The same behavior can be observed when varying the word length ($k=6, 7, 8$, Fig. 3a–d) and also for the remaining human and mouse chromosomes (see the word enrichment plots for all the chromosomes on the web supplement: <http://bioinfo2.ugr.es/DNAkeywords/>). The positive relation between σ_{nor} and word enrichment was found as well for other functional and/or evolutionarily conserved genome elements from both genomes: CpG islands (Hackenberg et al., 2006), TFBS conserved sites (Weirauch and Raney, 2011), phylogenetically conserved elements (PhastCons, Siepel et al., 2005), and insulators (Bao et al., 2008) (see the word enrichment plots for all these genome elements on the web supplement: <http://bioinfo2.ugr.es/DNAkeywords/>).

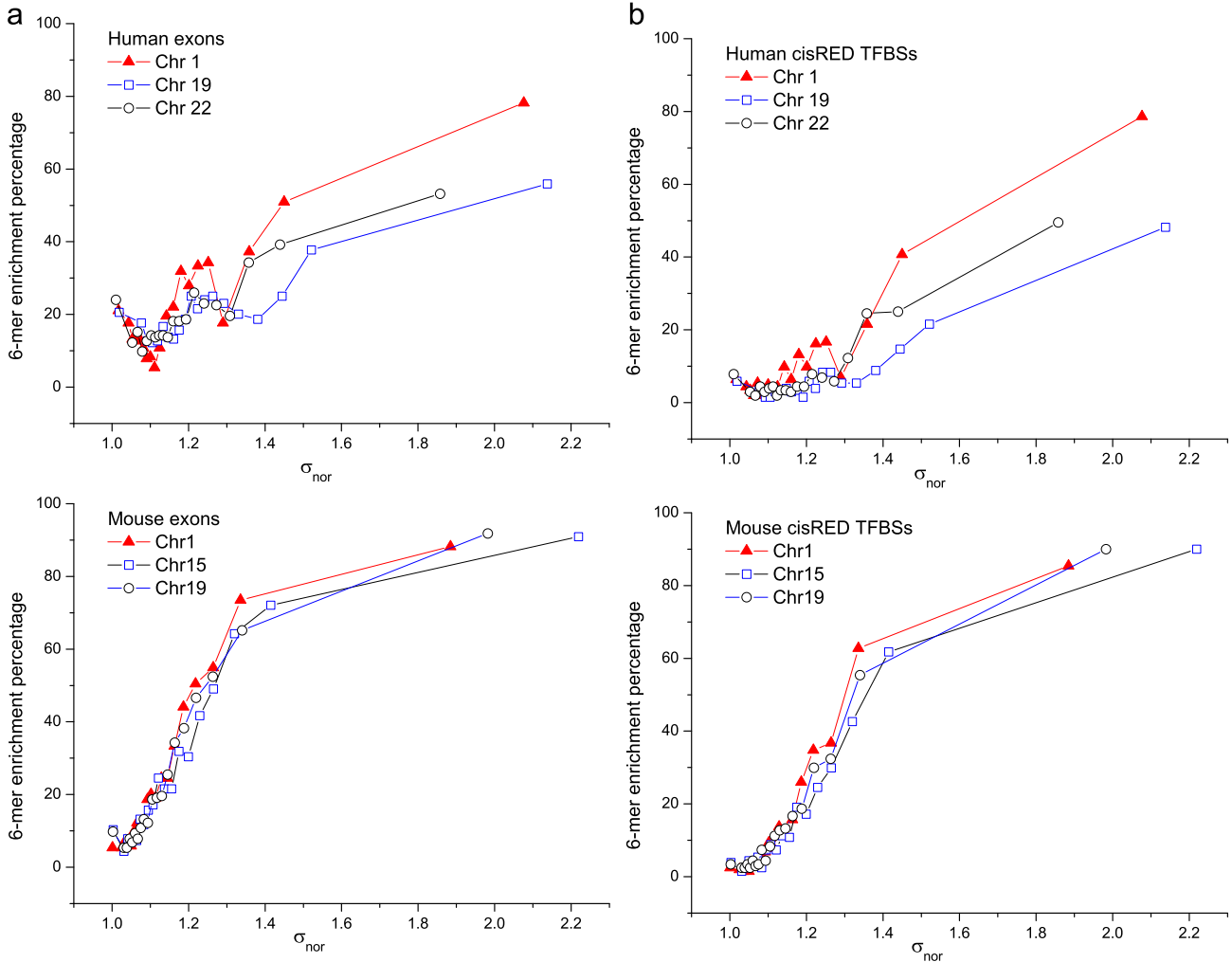


Fig. 2. 6-mer enrichment percentage versus clustering level in exons from the RefSeq database (Pruitt et al., 2007) (a), the cisRED TFBSs (Robertson et al., 2006) (b), repeats detected by RepeatMasker (Jurka et al., 2005; Smit et al., 1996–2010) (c), and introns from the RefSeq database (Pruitt et al., 2007) (d). The results for three humans (1, 19 and 22) and three mice (1, 5 and 9) chromosomes are shown. The same is true for longer words ($k=7, 8$) and also for the remaining human and mouse chromosomes (see the word enrichment plots for all the chromosomes and word-lengths on the web supplement: <http://bioinfo2.ugr.es/DNAkeywords/>). In the figure, the X axes show the mid-value of the word clustering level (partitioned on 20 bines), while the Y axes correspond to the percentage of words which are enriched ($r_i \geq 2$) within the particular genome element (exons, introns, TFBSs or repeats) in the given bin.

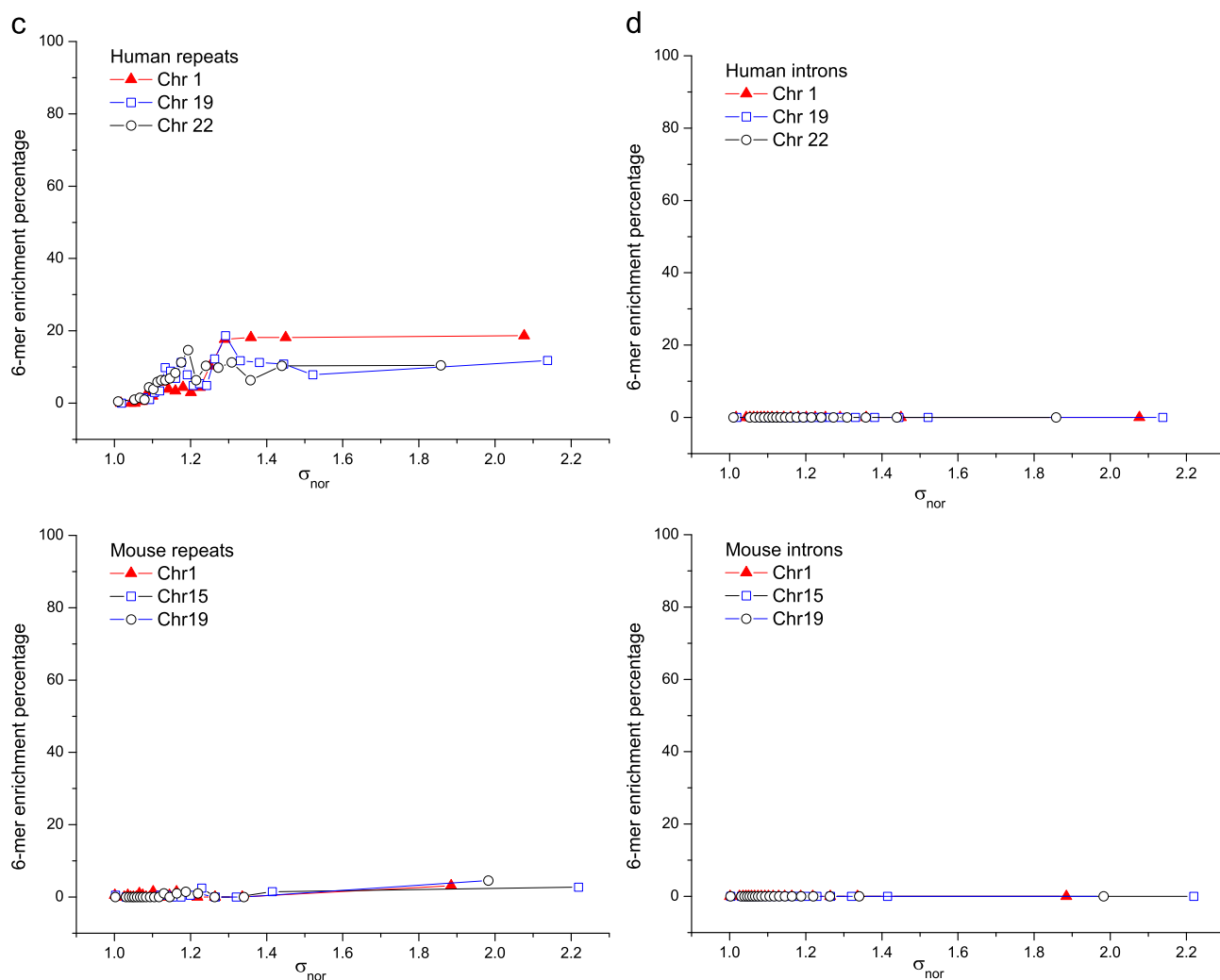


Fig. 2. (continued)

3.3. Many of the top-clustered words are significantly associated to functional elements

To further investigate the link between word-clustering and word-enrichment in a more direct and statistically significant way, we carried out extensive randomization tests to detect significant word enrichment or depletion in reference sequences like exons, introns and TFBSs (see Section 2 for details).

Given the high computational cost of such experiments, we restricted them to only two word subsets (or vocabularies). The first one contains the top-200 clustered 8-mers from each chromosome, while the second one, which we used as a control set, contains 200 randomly distributed, unclustered 8-mers from each chromosome (the list of words in both vocabularies are available on the web supplement). We first check that clustered and non-clustered words in these vocabularies are differentially represented in the databases of known motifs. Using the cisRED database for regulatory elements (Robertson et al., 2006), we found that 38.45% of the top-200 clustered words, but only 7.70% of the non-clustered words, are represented in this motif database. The difference is highly significant ($P \leq 0.001$), adding more evidence to the link between DNA word clustering and function. In the same way, when the JASPAR database (Sandelin et al., 2004) was scanned, we found that many of the top-200 clustered words are represented within known motifs (SP1, CREB, MEF2, MZF, p53, etc.). Furthermore, the word forming the core of SP1 sites (GGGCCG and its reverse complement) is clustered in each of

the 24 human chromosomes. Note however that, appealing as these relations may be, the most powerful way to show a statistically significant and genome-wide relation between the clustering of a particular word and biological function is comparing its richness inside vs. outside a given class of genome element.

Therefore, we carried out enrichment/depletion experiments for each of the words contained in both word subsets. Fig. 4 shows the percentages of enriched clustered and randomly distributed words ($p \leq 0.05$) in exons, introns and TFBSs in human and mouse chromosomes (see also the Enrichment/Depletion tables of top-200 clustered words on the web supplement). Higher percentages of enrichment in exons and TFBSs were observed for clustered words, as compared to random words, while the differences are lower for introns. With the only exception of chromosome Y, in the human genome the enrichments were always higher for clustered words than for random words. In the mouse, however, exceptions to this rule were also found for the exons of other chromosomes (9, 12 and X). The particular structure of the Y chromosome sequence (Skaletsky et al., 2003) and the lower quality of the mouse genome assembly may account for these exceptions.

Considering exons and TFBSs together, the 72.26% (in human) and 72.97% (in mouse) of the top-clustered 8-mers resulted associated to either exons or TFBSs, thus strongly supporting the link between word clustering and biological function. The remaining fraction of clustered words could constitute an important source of potential new functional markers.

Among the word subset containing the top-200 clustered 8-mers from each chromosome, there are many GC rich (and CpG rich) words, which can form part of CpG islands. This is why the plots corresponding to CpG islands (<http://bioinfo2.ugr.es/DNAkeywords/WordEnrichmentHistograms/hg18/CpGislands/>) show a clear relation between word clustering and the percentage of enrichment at these genome elements. Note that we cannot rule out that the known clustering of CpG dinucleotides affects the clustering of CpG containing k -mers even for high k . We are working on the elaboration of a more sophisticated vocabulary of functional words based on the clustering principle. The detection of spurious clustering caused by a “hitchhiking effect” of CpG dinucleotide clustering is one issue that needs to be tackled. The different compositional scales present in the genome (Carpena et al., 2007, 2011), and the consequent compartmentalization on the distribution of genome elements (Bernardi et al., 1985; Li, 2001; Oliver et al., 2001, 2004, 2002), may also impose specific constraints on the GC (and CpG) composition of clustered words.

The functionally relevant ‘motifs’ in exons are often domains or subdomains at the protein level. Therefore, part of the clustering of DNA words in coding exons (Figs. 2a and 3a) may be related to protein motifs. Repeated occurrences of a word in an exon

could result from repeated occurrences of a domain, as well as from codon usage and/or extreme GC levels, for example.

3.4. Conserved clustered words

We have also studied the shared words (among the top-200 clustered set by chromosome) in human and mouse genomes. We can interpret this set of shared words as “clustering-conserved” words. We found that human and mouse share a total of 668 clustered words. Out of these, 567 (84.9%) and 583 (87.3%) are associated to exons in human and mouse, respectively, corresponding to an increase in association of 16.1% in human and 14.3% in mouse. This finding is highly interesting as it hints to a relation between sequence and clustering conservation. Just like phylogenetic footprinting which is used to improve the prediction quality (less false positive predictions) of functional genome elements, this result suggests that the conservation of DNA word clustering can be used in the same way (a higher percentage is associated to exons). Sequence conservation is often used as an indicator for (often unknown) function, and therefore, the finding that clustering conservation behaves like sequencing conservation in this aspect adds more evidence that the clustering of DNA words has a biological meaning.

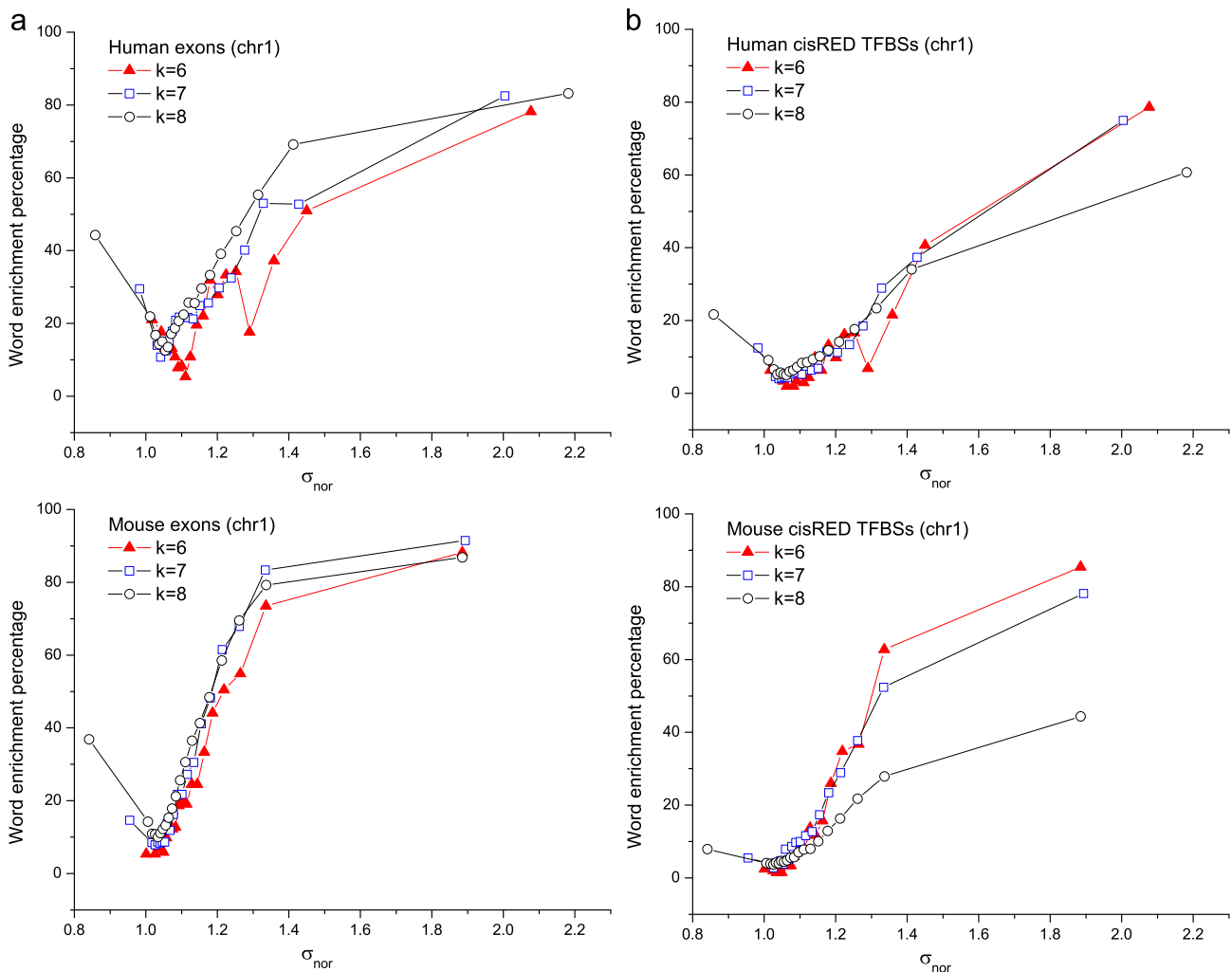


Fig. 3. Enrichment percentage versus clustering level in exons from the RefSeq database (Pruitt et al., 2007) (a), cisRED TFBSs (Robertson et al., 2006) (b), repeats detected by RepeatMasker (Jurka et al., 2005; Smit et al., 1996–2010) (c) and introns from the RefSeq database (Pruitt et al., 2007) (d) for human and mouse chromosome 1. Results for three word-sizes ($k=6, 7, 8$) are shown. In the figure, the X axes show the mid-value of the word clustering level (partitioned on 20 bins), while the Y axes correspond to the percentage of words which are enriched ($r_i \geq 2$) within the particular genome element (exons, introns, TFBSs or repeats) in the given bin.

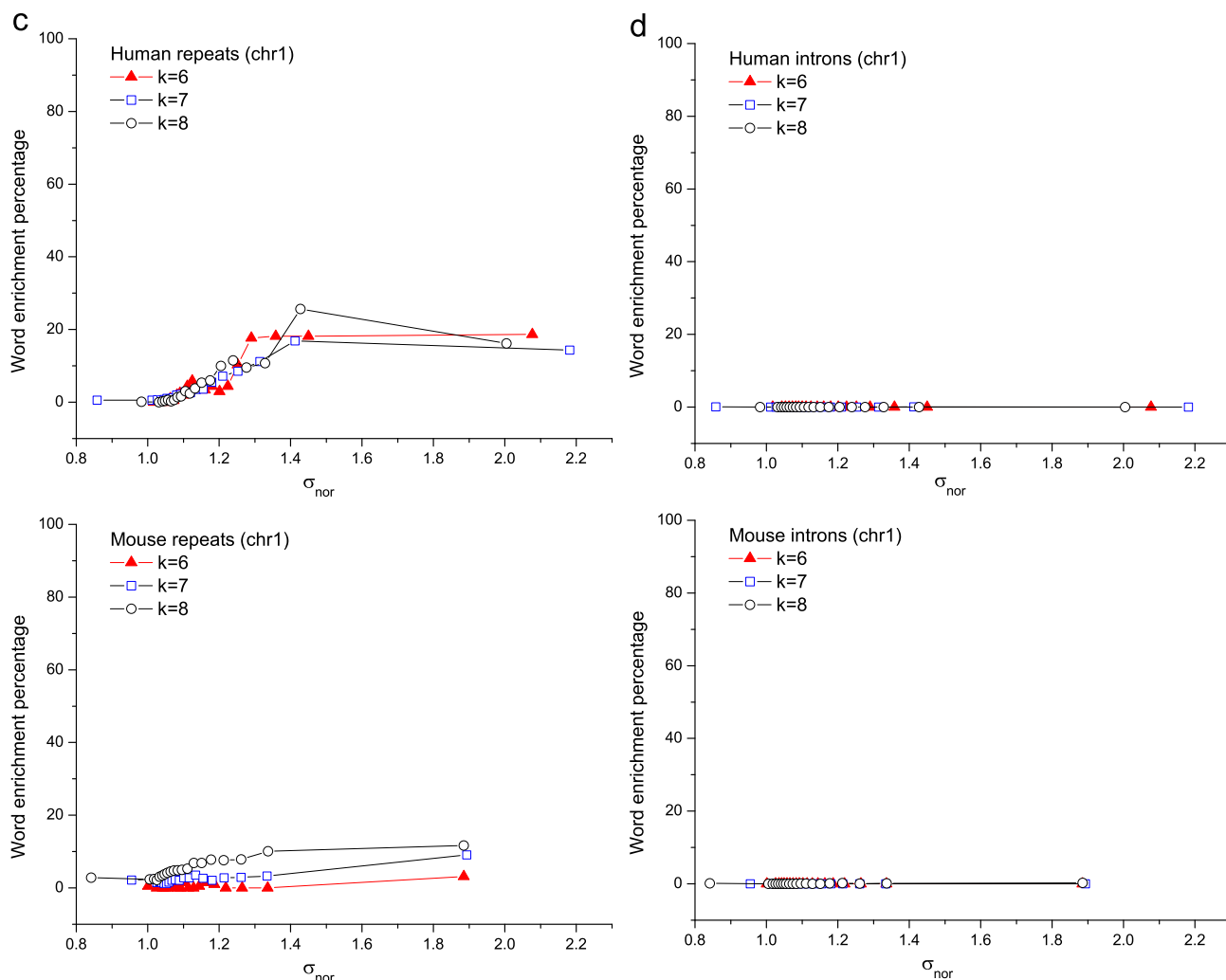


Fig. 3. (continued)

3.5. Diagnostic words: enrichment/depletion of clustered words in exons/introns

Finally, we have identified 1014 (in humans) and 424 (in mice) clustered words that are enriched in exons but significantly depleted ($p \leq 0.05$) in introns (see the tables of diagnostic words ($k=8$) on the web supplement). These words could be used, therefore, as diagnostic words to discriminate exons from introns. Interestingly, when randomly distributed, unclustered words were analyzed in the same way, only 8 words in humans, and none in mouse, were found to be enriched in exons but depleted in introns.

Diagnostic words could help to improve those gene prediction algorithms that are based on differential frequencies of k -mers (Burge and Karlin, 1997; Guigo et al., 1992; Staden and McLachlan, 1982). By preferentially focusing on diagnostic words, such algorithms could be substantially improved to best discriminate exons from sequence background.

In the last years, a great deal of work has been made to exhaustively annotate protein-coding genes; however, new exons follow to be discovered when more powerful studies are carried out (Lindblad-Toh et al., 2011). Diagnostic clustered words, not relying on evolutionary conservation, might be of help to uncover further exons.

4. Discussion

Besides the genetic code specifying the proteins, other information layers exist in the genome, being the most important the regulatory code—i.e. non-protein-coding transcripts and genomic elements that temporally and spatially regulate gene expression (Birney et al., 2007).

The use of language analogies has often proven useful in biology (Kay, 2000; Li, in press). We described here that another principle derived from linguistics may be useful in elucidating the genome code. Our finding that clustered words are enriched in the functional part of human and mouse genomes, suggests that the clustering principle may be of help in the genome-wide search of function in genome sequences. Three distinctive advantages can be drawn for our approach. First, any type of biological function, either regulatory or structural, can be potentially discovered. Second, training is not needed (i.e. it is an *ab initio* method), so that it could uncover both known and new functions. Third, it can be applied to individual genomes (i.e. no comparison with other related genomes is required), thus being able to uncover phylogenetically conserved as well as species-specific functional elements. Given that half of all functional sequence is specific to individual lineages (Ponting et al., 2011), the last feature looks particularly appealing.

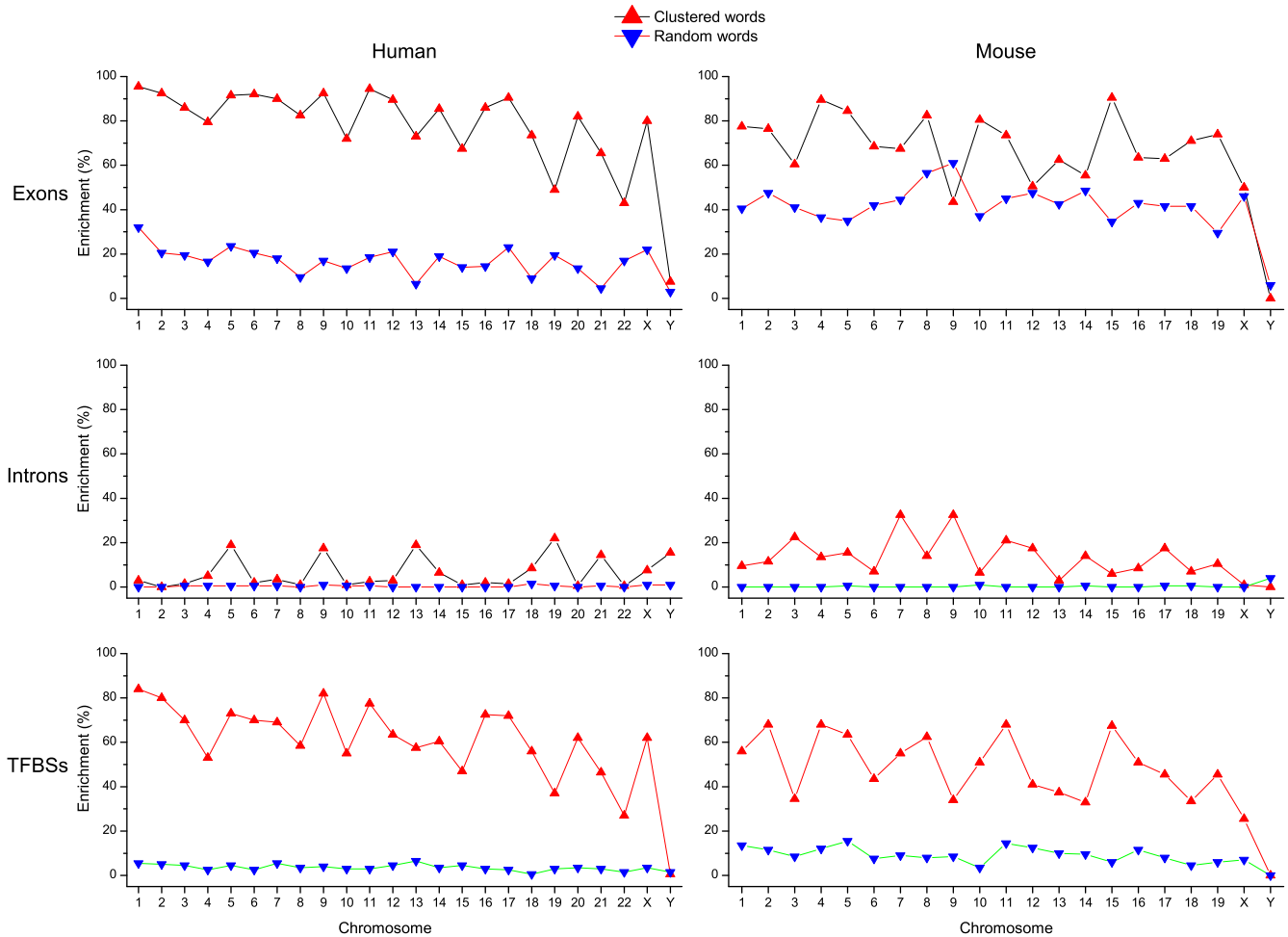


Fig. 4. Percentage of enrichment (p -value ≤ 0.05) of clustered and non-clustered words ($k=8$) in exons and introns from the RefSeq database (Pruitt et al., 2007), and cisRED TFBSs (Robertson et al., 2006) in human and mouse chromosomes.

Several functional studies (FitzGerald et al., 2004; Salisbury et al., 2006) have previously exploited local word clustering. Note, however, that these studies focus on a particular region (promoter, 3' flank) of an annotated genome. On the contrary, the approach used here is radically different: we first compute global word clustering in an entire non-annotated chromosome sequence, then trying to relate clustered words to different functional genome elements (exons, TFBSs, CpG islands, insulators, etc.). Thus, both approaches are clearly distinct. Through the positional genome approach it was observed that some words are clustered within functional genome elements, while through our approach we found that some genome-wide clustered words are enriched in given functional genome elements.

A limitation of our study is that it only consider word clustering along the one-dimensional chromosome sequence, thus ignoring the spatial clustering that may result from the 3D organization of the chromosomes within the nucleus, which can put together genetic elements actually far in the chromosome or even located on different chromosomes. Indeed, there exist functional elements, as enhancers or Locus Control Regions, which may become spatial neighbors within the nucleus, being however far away located on the lineal sequence of the chromosome. The 3D vicinity of genetic elements is surely most relevant to gene function (Pennisi, 2011), but unfortunately the sequence data we used do not allow us to address this interesting problem.

Once the link between global word-clustering and biological function has been established, a series of open questions appear.

Some of them may be worth of an in depth exploration to assess the potentialities and limitations of the new principle, also envisaging some practical applications. First, we have demonstrated that the top-200 clustered 8-mers of each chromosome show a clear relation with biological function. Preliminary data indicated, however, that a similar relation exists also for smaller word-lengths (i.e. $k=6$). Given the inclusion of shorter words in the larger ones, the first task would be to ascertain if within each word lineage exists a characteristic length carrying the signal for function. Only then, a vocabulary of true 'DNA keywords' (i.e. a curate set of words of different lengths and free of inclusions and redundancies) can be generated. Second, only 'exact' word copies have been searched for in this work. However, regulatory motifs of eukaryotic genes often contain variable binding sites for transcription factors (Boeva et al., 2007). It is also known that nucleotide variation of regulatory motifs may lead to distinct expression patterns (Segal et al., 2007). Therefore, it seems reasonable that degenerated, 'fuzzy' copies of each word lineage should also be included in the analysis. Third, given the combinatorial nature of gene regulation (Lemon and Tjian, 2000), it would be also interesting to determine the chromosome location and extension of statistically significant clusters of different words (Hackenberg et al., 2011), then determining the keyword composition of each cluster (i.e. homo- or heterotypic word clusters) and other compositional features. The hypothesis to explore here might be that, as occurs in *cis*-regulatory modules (Berman et al., 2004, 2002), the particular arrangement of words

in a homotypic or heterotypic cluster may be not random, and that the word arrangement within such clusters may be important for its functionality. Lastly, some of the clustered words we found may be useful to the development of epigenome markers. An example is the word TACAG, which we found highly clustered in many chromosomes. The non-CpG site for cytosine methylation contained within this word have been found to be heavily methylated in human embryonic stem cell lines, as well as strongly conserved and enriched in splice sites, which may suggest a mechanistic connection between DNA methylation and splicing (Chen et al., 2011; Laurent et al., 2010). Further work is in progress to explore and substantiate all these possibilities.

5. Conclusions

We have shown that the clustering of a DNA word significantly correlates with its association to functional elements: the higher the clustering of a word, the more likely it is enriched in functional elements. Almost the 73% of the top-clustered 8-mers resulted associated to exons or TFBSs in human and mouse genomes, thus strongly supporting the link between word clustering and biological function. This strong tendency was not observed in introns or in unclustered, randomly distributed words (non-functional background). The behavior in human and mouse genomes was virtually the same, what makes us believe that the link between word clustering and function might be a general principle. This new concept might help to uncover species-specific functionality and to improve the prediction of important functional elements, as exons or TFBSs.

Supporting information

The following supplementary materials related to this article can be found online at <http://bioinfo2.ugr.es/DNAkeywords/>:

- Word clustering statistics ($k=2-9$)
- Word enrichment plots ($k=6-8$)
- Top-200 clustered words by chromosome ($k=8$)
- Enrichment/depletion analyses of top-200 clustered words ($k=8$)
- Diagnostic words ($k=8$)

Acknowledgments

This work was supported by the Ministry of Innovation and Science of the Spanish Government [BIO2008-01353 to JLO and BIO2010-20219 to MH], 'Juan de la Cierva' grant (M.H.) and Basque country 'AE' grant (G.B.). We gratefully acknowledge the valuable comments of two anonymous referees, which significantly improved the manuscript. We thank Ángel M. Alganza for help with system administration and database support.

References

- Arnau, V., Gallach, M., Marin, I., 2008. Fast comparison of DNA sequences by oligonucleotide profiling. *BMC Res. Notes* 1, 5.
- Bao, L., Zhou, M., Cui, Y., 2008. CTCFBSDB: a CTCF-binding site database for characterization of vertebrate genomic insulators. *Nucleic Acids Res.* 36, D83–D87.
- Berman, B.P., Pfeiffer, B.D., Laverty, T.R., Salzberg, S.L., Rubin, G.M., Eisen, M.B., Celniker, S.E., 2004. Computational identification of developmental enhancers: conservation and function of transcription factor binding-site clusters in *Drosophila melanogaster* and *Drosophila pseudoobscura*. *Genome Biol.* 5, R61.
- Berman, B.P., Nibu, Y., Pfeiffer, B.D., Tomancak, P., Celniker, S.E., Levine, M., Rubin, G.M., Eisen, M.B., 2002. Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the *Drosophila* genome. *Proc. Natl. Acad. Sci. USA* 99, 757–762.
- Bernardi, G., Olofsson, B., Filipski, J., Zerial, M., Salinas, J., Cuny, G., Meunier-Rotival, M., Rodier, F., 1985. The mosaic genome of warm-blooded vertebrates. *Science* 228, 953–958.
- Bird, A.P., 1986. CpG-rich islands and the function of DNA methylation. *Nature* 321, 209–213.
- Birney, E., Stamatoyannopoulos, J.A., Dutta, A., Guigo, R., Gingeras, T.R., Margulies, E.H., Weng, Z., Snyder, M., Dermitzakis, E.T., Thurman, R.E., Kuehn, M.S., Taylor, C.M., Neph, S., Koch, C.M., Asthana, S., Malhotra, A., Adzhubei, I., Greenbaum, J.A., Andrews, R.M., Flicek, P., Boyle, P.J., Cao, H., Carter, N.P., Clelland, G.K., Davis, S., Day, N., Dhami, P., Dillon, S.C., Dorschner, M.O., Fiegler, H., Giresi, P.G., Goldy, J., Hawrylycz, M., Haydock, A., Humbert, R., James, K.D., Johnson, B.E., Johnson, E.M., Frum, T.T., Rosenzweig, E.R., Karnani, N., Lee, K., Lefebvre, G.C., Navas, P.A., Neri, F., Parker, S.C., Sabo, P.J., Sandstrom, R., Shafer, A., Vetrie, D., Weaver, M., Wilcox, S., Yu, M., Collins, F.S., Dekker, J., Lieb, J.D., Tullius, T.D., Crawford, G.E., Sunyaev, S., Noble, W.S., Dunham, I., Denoeud, F., Reymond, A., Kapranov, P., Rozowsky, J., Zheng, D., Castelo, R., Frankish, A., Harrow, J., Ghosh, S., Sandelin, A., Hofacker, I.L., Baertsch, R., Keefe, D., Dike, S., Cheng, J., Hirsch, H.A., Sekinger, E.A., Lagarde, J., Abril, J.F., Shahab, A., Flamm, C., Fried, C., Hackermuller, J., Hertel, J., Lindemeyer, M., Missal, K., Tanzer, A., Washietl, S., Korb, J., Emanuelsson, O., Pedersen, J.S., Holroyd, N., Taylor, R., Swarbreck, D., Matthews, N., Dickson, M.C., Thomas, D.J., Weirauch, M.T., Gilbert, J., et al., 2007. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* 447, 799–816.
- Boeva, V., Clement, J., Regnier, M., Roytberg, M.A., Makeev, V.J., 2007. Exact p -value calculation for heterotypic clusters of regulatory motifs and its application in computational annotation of cis-regulatory modules. *Algorithms Mol. Biol.* 2, 13.
- Brown, C.D., Johnson, D.S., Sidow, A., 2007. Functional architecture and evolution of transcriptional elements that drive gene coexpression. *Science* 317, 1557–1560.
- Burge, C., Karlin, S., 1997. Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* 268, 78–94.
- Bussemaker, H.J., Li, H., Siggia, E.D., 2000. Building a dictionary for genomes: identification of presumptive regulatory sites by statistical analysis. *Proc. Natl. Acad. Sci. USA* 97, 10096–10100.
- Carmack, C.S., McCue, L.A., Newberg, L.A., Lawrence, C.E., 2007. PhyloScan: identification of transcription factor binding sites using cross-species evidence. *Algorithms Mol. Biol.* 2, 1.
- Carpena, P., Bernaola-Galván, P., Coronado, A.V., Hackenberg, M., Oliver, J.L., 2007. Identifying characteristic scales in the human genome. *Phys. Rev. E* 75, 032903.
- Carpena, P., Bernaola-Galván, P., Hackenberg, M., Coronado, A.V., Oliver, J.L., 2009. Level statistics of words: finding keywords in literary texts and DNA. *Phys. Rev. E* 79, 035102–4.
- Carpena, P., Oliver, J.L., Hackenberg, M., Coronado, A.V., Barturen, G., Bernaola-Galván, P., 2011. High-level organization of isochores into gigantic superstructures in the human genome. *Phys. Rev. E Stat. Nonlinear Soft Matter Phys.* 83, 031908.
- Chakravarty, A., Carlson, J.M., Khetani, R.S., DeZiel, C.E., Gross, R.H., 2007. SPACER: identification of cis-regulatory elements with non-contiguous critical residues. *Bioinformatics* 23, 1029–1031.
- Chen, P.-Y., Feng, S., Joo, J.W.J., Jacobsen, S.E., Pellegrini, M., 2011. A comparative analysis of DNA methylation across human embryonic stem cell lines. *Genome Biol.* 12, R62.
- Durand, D., Sankoff, D., 2003. Tests for gene clustering. *J. Comput. Biol.* 10, 453–482.
- Eden, E., Lipson, D., Yogev, S., Yakhini, Z., 2007. Discovering motifs in ranked lists of DNA sequences. *PLoS Comput. Biol.* 3, e39.
- FitzGerald, P.C., Shlyakhtenko, A., Mir, A.A., Vinson, C., 2004. Clustering of DNA sequences in human promoters. *Genome Res.* 14, 1562–1574.
- Giardine, B., Riemer, C., Hardison, R.C., Burhans, R., Elnitski, L., Shah, P., Zhang, Y., Blankenberg, D., Albert, I., Taylor, J., Miller, W., Kent, W.J., Nekrutenko, A., 2005. Galaxy: a platform for interactive large-scale genome analysis. *Genome Res.* 15, 1451–1455.
- Guigo, R., Knudsen, S., Drake, N., Smith, T., 1992. Prediction of gene structure. *J. Mol. Biol.* 226, 141–157.
- Hackenberg, M., Previti, C., Luque-Escamilla, P.L., Carpena, P., Martínez-Aroza, J., Oliver, J.L., 2006. CpGcluster: a distance-based algorithm for CpG-island detection. *BMC Bioinform.* 7, 446.
- Hackenberg, M., Barturen, G., Carpena, P., Luque-Escamilla, P.L., Previti, C., Oliver, J.L., 2010. Prediction of CpG-island function: CpG clustering vs. sliding-window methods. *BMC Genom.* 11, 327.
- Hackenberg, M., Carpena, P., Bernaola-Galván, P., Barturen, G., Alganza, A.M., Oliver, J.L., 2011. WordCluster: detecting clusters of DNA words and genomic elements.
- Hampson, S., Kibler, D., Baldi, P., 2002. Distribution patterns of over-represented k -mers in non-coding yeast DNA. *Bioinformatics* 18, 513–528.
- Han, L., Zhao, Z., 2009. CpG islands or CpG clusters: how to identify functional GC-rich regions in a genome? *BMC Bioinform.* 10, 65.
- Herold, J., Kurtz, S., Giegerich, R., 2008. Efficient computation of absent words in genomic sequences. *BMC Bioinform.* 9, 167.
- Hughes, J.D., Estep, P.W., Tavazoie, S., Church, G.M., 2000. Computational identification of cis-regulatory elements associated with groups of functionally related genes in *Saccharomyces cerevisiae*. *J. Mol. Biol.* 296, 1205–1214.
- Jurka, J., Kapitonov, V.V., Pavlicek, A., Klonowski, P., Kohany, O., Walichiewicz, J., 2005. Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet. Genome Res.* 110, 462–467.

- Karolchik, D., Kuhn, R.M., Baertsch, R., Barber, G.P., Clawson, H., Diekhans, M., Giardine, B., Harte, R.A., Hinrichs, A.S., Hsu, F., Kober, K.M., Miller, W., Pedersen, J.S., Pohl, A., Raney, B.J., Rhead, B., Rosenbloom, K.R., Smith, K.E., Stanke, M., Thakkapallayil, A., Trumbower, H., Wang, T., Zweig, A.S., Haussler, D., Kent, W.J., 2008. The UCSC Genome Browser Database: 2008 update. *Nucleic Acids Res.* 36, D773–D779.
- Kay, L.E., 2000. Who Wrote the Book of Life? A History of the Genetic Code. Stanford University Press, Palo Alto, CA.
- Kendal, W.S., 2004. A scale invariant clustering of genes on human chromosome 7. *BMC Evol. Biol.* 4, 3.
- Laurent, L., Wong, E., Li, G., Huynh, T., Tsigos, A., Ong, C.T., Low, H.M., Kin Sung, K.W., Rigoutsos, I., Loring, J., Wei, C.L., 2010. Dynamic changes in the human methylome during differentiation. *Genome Res.* 20, 320–331.
- Lemon, B., Tjian, R., 2000. Orchestrated response: a symphony of transcription factors for gene control. *Genes Dev.* 14, 2551–2569.
- Li, W., 2001. Delineating relative homogeneous G+C domains in DNA sequences. *Gene* 276, 57–72.
- Li, W., in press. Menzerath's Law at the gene-exon level in the human genome. *Complexity*, doi:10.1002/cplx.20398. In press.
- Li, X., Wong, W.H., 2005. Sampling motifs on phylogenetic trees. *Proc. Natl. Acad. Sci. USA* 102, 9481–9486.
- Lindblad-Toh, K., Garber, M., Zuk, O., Lin, M.F., Parker, B.J., Washietl, S., Kheradpour, P., Ernst, J., Jordan, G., Mauceli, E., Ward, L.D., Lowe, C.B., Holloway, A.K., Clamp, M., Gnerre, S., Alfoldi, J., Beal, K., Chang, J., Clawson, H., Cuff, J., Di Palma, F., Fitzgerald, S., Flicek, P., Guttman, M., Hubisz, M.J., Jaffe, D.B., Jungreis, I., Kent, W.J., Kostka, D., Lara, M., Martins, A.L., Masingham, T., Moltke, I., Raney, B.J., Rasmussen, M.D., Robinson, J., Stark, A., Vilella, A.J., Wen, J., Xie, X., Zody, M.C., Worley, K.C., Kovar, C.L., Muzny, D.M., Gibbs, R.A., Warren, W.C., Mardis, E.R., Weinstock, G.M., Wilson, R.K., Birney, E., Margulies, E.H., Herrero, J., Green, E.D., Haussler, D., Siepel, A., Goldman, N., Pollard, K.S., Pedersen, J.S., Lander, E.S., Kellis, M., 2011. A high-resolution map of human evolutionary constraint using 29 mammals. *Nature* advance online publication.
- Neel, J.V., 1961. The hemoglobin genes: a remarkable example of the clustering of related genetic functions on a single mammalian chromosome. *Blood* 18, 769–777.
- Nussinov, R., 1981. Eukaryotic dinucleotide preference rules and their implications for degenerate codon usage. *J. Mol. Biol.* 149, 125–131.
- Oliver, J.L., Bernaola-Galván, P., Carpena, P., Román-Roldán, R., 2001. Isochore chromosome maps of eukaryotic genomes. *Gene* 276, 47–56.
- Oliver, J.L., Carpena, P., Hackenberg, M., Bernaola-Galván, P., 2004. IsoFinder: computational prediction of isochores in genome sequences. *Nucleic Acids Res.* 32, W287–W292.
- Oliver, J.L., Carpena, P., Roman-Roldan, R., Mata-Balaguer, T., Mejias-Romero, A., Hackenberg, M., Bernaola-Galvan, P., 2002. Isochore chromosome maps of the human genome. *Gene* 300, 117–127.
- Ortuño, M., Carpena, P., Bernaola-Galván, P., Muñoz, E., Somoza, A., 2002. Keyword detection in natural languages and DNA. *Europhys. Lett.* 57, 759–764.
- Pennisi, E., 2011. Does a gene's location in the nucleus matter? *Science* 334, 1050–1051.
- Ponting, C.P., Nellaker, C., Meader, S., 2011. Rapid turnover of functional sequence in human and other genomes. *Annu. Rev. Genomics Hum. Genet.* 12, 275–299.
- Pruitt, K.D., Tatusova, T., Maglott, D.R., 2007. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.* 35, D61–D65.
- Robertson, G., Bilenky, M., Lin, K., He, A., Yuen, W., Dagpinar, M., Varhol, R., Teague, K., Griffith, O.L., Zhang, X., Pan, Y., Hassel, M., Sleumer, M.C., Pan, W., Pleasance, E.D., Chuang, M., Hao, H., Li, Y.Y., Robertson, N., Fjell, C., Li, B., Montgomery, S.B., Astakhova, T., Zhou, J., Sander, J., Siddiqui, A.S., Jones, S.J., 2006. cisRED: a database system for genome-scale computational discovery of regulatory elements. *Nucleic Acids Res.* 34, D68–D73.
- Salisbury, J., Hutchison, K.W., Graber, J.H., 2006. A multispecies comparison of the metazoan 3'-processing downstream elements and the CstF-64 RNA recognition motif. *BMC Genomics* 7, 55.
- Sandelin, A., Alkema, W., Engstrom, P., Wasserman, W.W., Lenhard, B., 2004. JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res.* 32, D91–D94.
- Sandve, G.K., Gundersen, S., Rydbeck, H., Glad, I.K., Holden, L., Holden, M., Liestol, K., Clancy, T., Ferkingstad, E., Johansen, M., Nygaard, V., Tostesen, E., Frigessi, A., Hovig, E., 2010. The Genomic HyperBrowser: inferential genomics at the sequence level. *Genome Biol.* 11, R121.
- Sargsyan, K., Lim, C., 2010. Arrangement of 3D structural motifs in ribosomal RNA. *Nucleic Acids Res.* 38, 3512–3522.
- Segal, L., Lapidot, M., Solan, Z., Ruppim, E., Pilpel, Y., Horn, D., 2007. Nucleotide variation of regulatory motifs may lead to distinct expression patterns. *Bioinformatics* 23, i440–i449.
- Siddharthan, R., Nimwegen, E., 2007. Detecting regulatory sites using PhyloGibbs. *Methods Mol. Biol.* 395, 381–402.
- Siepel, A., Bejerano, G., Pedersen, J.S., Hinrichs, A.S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L.W., Richards, S., Weinstock, G.M., Wilson, R.K., Gibbs, R.A., Kent, W.J., Miller, W., Haussler, D., 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* 15, 1034–1050.
- Sinha, S., Blanchette, M., Tompa, M., 2004. PhyME: a probabilistic algorithm for finding motifs in sets of orthologous sequences. *BMC Bioinform.* 5, 170.
- Skaletsky, H., Kuroda-Kawaguchi, T., Minx, P.J., Cordum, H.S., Hillier, L., Brown, L.G., Repping, S., Pyntikova, T., Ali, J., Bieri, T., Chinwalla, A., Delehaunty, A., Delehaunty, K., Du, H., Fewell, G., Fulton, L., Fulton, R., Graves, T., Hou, S.F., Latrielle, P., Leonard, S., Mardis, E., Maupin, R., McPherson, J., Miner, T., Nash, W., Nguyen, C., Ozersky, P., Pepin, K., Rock, S., Rohlffing, T., Scott, K., Schultz, B., Strong, C., Tin-Wollam, A., Yang, S.P., Waterston, R.H., Wilson, R.K., Rozen, S., Page, D.C., 2003. The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. *Nature* 423, 825–837.
- Smit, A., Hubley, R., Green, P., 1996–2010. RepeatMasker Open-3.0 <<http://www.repeatmasker.org/>>.
- Staden, R., McLachlan, A.D., 1982. Codon preference and its use in identifying protein coding regions in long DNA sequences. *Nucleic Acids Res.* 10, 141–156.
- Subirana, J.A., Messegue, X., 2010. The most frequent short sequences in non-coding DNA. *Nucleic Acids Res.* 38, 1172–1181.
- Trifonov, E.N., Brendel, V., 1986. *Gnomic. A Dictionary of Genetic Codes*. Balaban Publishers, Rehovot, Philadelphia.
- Tsonis, A.A., Elsner, J.B., Tsonis, P.A., 1997. Is DNA a language? *J. Theor. Biol.* 184, 25–29.
- Wang, G., Yu, T., Zhang, W., 2005. WordSpy: identifying transcription factor binding motifs by building a dictionary and learning a grammar. *Nucleic Acids Res.* 33, W412–W416.
- Wang, T., Stormo, G.D., 2003. Combining phylogenetic data with co-regulated genes to identify regulatory motifs. *Bioinformatics* 19, 2369–2380.
- Weirauch, M., Raney, B., 2011. HMR Conserved Transcription Factor Binding Sites, vol. 2011. <<http://genome.ucsc.edu/cgi-bin/hgTrackUi?hgsid=98599487&c=chrX&g=tfbsConsSites>>.
- Wilson, M.D., Barbosa-Morais, N.L., Schmidt, D., Conboy, C.M., Vanes, L., Tybulewicz, V.L., Fisher, E.M., Tavare, S., Odom, D.T., 2008. Species-specific transcription in mice carrying human chromosome 21. *Science* 322, 434–438.
- Zhou, H., Slater, G.W., 2003. A metric to search for relevant words. *Physica A* 329, 309–327.