A top-down linguistic approach to the analysis of genomic sequences: The metabotropic glutamate receptors 1 and 5 in human and in mouse as a case study

Giulia Menconi, Aldamaria Puliti, Isabella Sbrana, Valerio Conti, Roberto Marangoni

www.elsevier.com/locate/yjtbi

A top-down linguistic approach to the analysis of genomic sequences: The metabotropic glutamate receptors 1 and 5 in human and in mouse as a case study

Giulia Menconi, Aldamaria Puliti, Isabella Sbrana, Valerio Conti, Roberto Marangoni

PII: S0022-5193(10)00605-3
DOI: doi:10.1016/j.jtbi.2010.11.020
Reference: YJTBI 6242

To appear in: *Journal of Theoretical Biology*

Received date: 15 June 2010
Revised date: 18 October 2010
Accepted date: 10 November 2010

Cite this article as: Giulia Menconi, Aldamaria Puliti, Isabella Sbrana, Valerio Conti and Roberto Marangoni, A top-down linguistic approach to the analysis of genomic sequences: The metabotropic glutamate receptors 1 and 5 in human and in mouse as a case study, *Journal of Theoretical Biology*, doi:10.1016/j.jtbi.2010.11.020

A top-down linguistic approach to the analysis of genomic sequences: The metabotropic Glutamate receptors 1 and 5 in Human and in Mouse as a case study

Giulia Menconi[1a], Aldamaria Puliti[b,c], Isabella Sbrana[d], Valerio Conti[b], Roberto Marangoni[e,f]

[a] *Istituto Nazionale di Alta Matematica, Roma, Italia*
[b] *Molecular genetics and Cytogenetics Unit, Gaslini Institute, Genova, Italia*
[c] *Dipartimento di Scienze Pediatriche, Università di Genova, Italia*
[d] *Dipartimento di Biologia, Università di Pisa, Italia*
[e] *Dipartimento di Informatica, Università di Pisa, Italia*
[f] *Istituto di Biofisica, CNR, Pisa, Italia*

Abstract

This paper presents a top-down strategy to detect features in genomic sequences. The strategy's core is to exploit dictionary-based compression algorithms and analyze the content of the automatically generated dictionary. We classify the different over-represented words and in the case study we correlate them to experimentally identified or theoretically forecasted biological features. A large spectrum analysis reveals that the only feature co-located with the *a priori* extracted words is the torsional flexibility of DNA, while non-B DNA configurations are anti-localized and other features are mostly independent of the extracted words. This analysis unravels complex relationships between the linguistic structures investigated under our approach and some known biological features.

Keywords: 68R15, 62P10, 92B05

[1]To whom correspondence should be addressed. Istituto Nazionale di Alta Matematica, Piazzale A. Moro 5 I-00185 Roma (Italy); Email: menconi@mail.dm.unipi.it. Phone: +39 050 2213128. Fax: +39 050 2212726.

1. Introduction

Genomes, and in particular eukaryotic genomes, are far to be homogeneous set of sequences, as they host several elements of different structure, functional role and even origin (e.g. exogenous elements). The development of strategies to recognize and classify these different kind of sequences is then a challenge for contemporary bioinformatics. This task can be resembled to a reverse engineering of an unknown operating system that sizes 3.3 gigabytes and host data, routines, re-assembled file fragments, etc., as in the very bright analogy suggested by Robbins [21].

The first step towards such a reverse engineering, is to get a (evenly rough) classification of the different elements existing on a genomic sequence: i.e., to distinguish between biological features carried by different sub-sequences (also called segments or words) of the DNA. Experimental methods have discovered a wide set of qualitatively different elements: coding regions, non-coding regions, introns, exons, promoters, enhancers, transcription factor binding sites, etc. Since experimental methods are very slow in adding new information, computational methods are searched in order to screen the whole genome searching for these features. Current computational approaches to discover functionally-associated elements along the genome are commonly based on a bottom-up philosophy (we could call it also *inductive* methods): on the basis of a known set of sequences that belong to a given class, suitable algorithms are designed to recognize unknown element. With suitable algorithm we refer to any approach belonging to approximate searches (e.g. consensus searches) or to machine learning strategies, as neural networks, support vector machines and similar.

Such bottom-up approaches are the most widely used and very useful, they anyhow soffer of a common problem: their efficiency is strictly bound to the composition of the training set. Since many classes show a very low common similarity, most of the classifications obtained by means of these methods are continuosly subject to revision.

From an opposite philosophy there are the so-called top-down methods (which could also be called *deductive* methods) where abstract tools are used to extract features from genomic sequences, without using any experimental data already known [2] . The importance of such approaches is clear: if we found an abstract formula able to correctly recognize some functional features in genomic sequence, this would represent a great advance in understanding DNA logic.

As of today, top-down approaches have grown slowly and less efficiently, if compared with bottom-up method, as the main problem is to recognize a good theoretical criterion to be applied to biological data.

The most important first attempt to do this, is represented by the application of general-purpose linguistic approaches on genomic sequences: the authors of [7] introduced the concept of "meaningful words" as an element of an organism-specific vocabulary in the DNA language. Since this pioneeristic work, several other papers have been produced where linguistic approaches have been applied to understand a wide variety of characteristics in genomes, from the identification of active genes to the large scale comparison [18], [22], [10].

In recent times, great importance is tributed to compression algorithms, as they provide, at the same time, both a linguistic tool to analyze sequence, and a method to store large sequences saving space. In fact, due to the exponential growing of biological databanks, a compression method able to efficently compress and allowing the sequence analysis directly on the compressed data is actively searched [14].

Dictionary-based compression algorithms, like those of the Lempel-Ziv family have been already used in the past to have an automatic selector of over-represented words, in order to select repeats along a genomic sequence [22] [18] [16], or to classify coding/non-coding sequences on the basis of the compression factor or similar indexes [19].

The analysis we are going to perform is general purpose and it is not focused on a specific biological feature. This makes it necessary to apply this kind of investigation on a well known case study, where biological features, structural characteristics and even phylogenetic relationships have been already described, being a first-glance genomic wide study almost impossible. Among the possible choices, we selected the glutamate metabotropic receptors genes 1 and 5 in the Human and in the Mouse. This selection has been based on multiple criteria: first, even if these sequences are referred to single genes, they are very long ($\sim$400 Kbp), showing long intronic sequences and, likely, they host most of the structural/functional characteristics found in the genome. Moreover, they exist in different subtypes, expressed in different organisms, and expressed not only in the nervous system but also in many other tissues, then it is possible to infer common features. Last, they have a great importance in cellular and synaptic activity, plasticity, cell death and survival, learning and memory, pain perception, and motor activity, and seem to have a role in the development of many complex pathologies. These

genes are either paralogous or orthologous each other and this allows us to study the eventual relationship between homology and over-expressed words.

2. Approach

The use of compression algorithms for genome data mining has been previously explored; in a previous work some of us proved that a discrimination between coding and non-coding regions in bacteria genomic sequences can be obtained *a priori* by studing the information content of a sequence [19].

The work is organised as follows.

A first information analysis exploits a compression on the genes and provides a dictionary of recurrent words. It is clear that recurrent subsequences share a symmetry in AT/CG content, which suggests an *ad hoc* deeper investigation. Second, we perform a statistical linguistic analysis on the complete gene sequences. Finally, we show whether and what the relationships are of the above results with experimentally found or computationally predicted local biological properties.

3. Materials and Methods

3.1. Metabotropic glutamate receptors 1 and 5

The mGlu1 and 5 receptors belong to the group I of metabotropic glutamate receptors which represent a family of eight G-protein coupled receptors distinguished on the basis of sequence diversity, expression profiles and pharmacology. The gene encoding for the mGlu1 receptor (locus name: $GRM1$ in humans and $Grm1$ in other species) has been mapped to chromosome 6q24 in humans, and chromosome 10, band 10a1, in mice, while the gene encoding for the mGlu5 receptor (locus name: $GRM5$ in humans and $Grm5$ in other species) has been mapped to chromosome 11 in humans and chromosome 7 in mice [28]. Exon/intron boundaries reveals that the human $GRM1$ spans about 410 kilobase pairs and consists of 10 exons and 9 introns. Exons vary from 85 (exon IX) to 3724 bp (exon X) in size, whereas intron sizes range from 149 to 1.3 kilobase pairs. The 10 different exons generate, by alternative splicing, more than 6 different splice variants [15]. Different protein isoforms have been described both in human and murine *GRM1*, among which the alpha and beta, of 1199 and 906 amino acids respectively, are the longest variants and represent the major forms expressed in the central nervous system. Comparison of the genomic structures of *GRM1* with *GRM5* reveals a

high degree of similarity in terms of exon/intron arrangement, both in human and mouse, which strongly suggests that group I mGlu receptors have been generated by gene duplication from a common ancestor. Analogies and/or diversities in their genomic sequence organization may reveal some biological features that the paralogous genes may share.

Concerning transcriptional regulation, functional studies indicate that the mGlu1 receptor gene, both in humans and mice, is driven by at least two alternative promoters located upstream from exons I and II, with the latter encoding the transcription initiation codon [12]. Functional analyses reveal the presence of a 57-bp core promoter from the first transcription initiation site, and two silencing elements, located between exons Ib and Ic, and the regulatory factor for X-box element found upstream from exon II [12]. Both silencing elements have a strong suppressive role in non-neuronal cells.

Main functions of mGlu1 and 5 receptors are in the regulation of neuronal excitability, synaptic plasticity, synapse selection, and neurotransmitter release, which are important for brain development and mechanisms of learning and neuroprotection. For their functions both mGlu1 and 5 receptors have been implicated in the pathophysiology of several neurological and psychiatric disorders, and represent possible targets for new therapeutic approaches. For all these reasons, a better comprension of mechanisms regulating *GRM1* and *GRM5* gene structures, activities and expression may be instrumental for the achievement of these goals.

3.1.1. DNA sequences

Human *GRM1* and *GRM5* sequences were from NCBI Build 36.1 and mouse sequences *Grm1* and *Grm5* from Build 37 (UCSC Genome Bioinformatics [26]). We based our analysis of human and mouse *GRM1* genes on the genomic structures obtained from UCSC data for all reported gene isoforms (obtained by [12]). The DNA strand that has been analysed is that indicated by the UCSC browser as coding strand (plus strand). The analysed sequence includes the $5'$ 5kb upstream to the first exon, and the 5kb downstream the end of the last $3'$ exons. All exons, including $5'$ and $3'$ UTR exons, were taken in consideration to get the final sequence to be analysed.

Some statistical features of the genes are shown in Table 1. Genes *GRM1* and *GRM5* in *Homo sapiens* and *Grm1* and *Grm5* in *Mus musculus* all share a low GC content.

This work aims at achieving a better understanding of oligonucleotide repetitive structures shared by the four genes.

Table 1: The four *grm* genes under study.

Gene	specie	Chromosome	length	GC-content
GRM1	*H. sapiens*	chr 6	412965 bp	37%
GRM5	*H. sapiens*	chr 11	563148 bp	36%
Grm1	*M. musculus*	chr 10	398962 bp	39%
Grm5	*M. musculus*	chr 7	552292 bp	37%

3.2. Algorithm and dictionaries

The proposed method is based on the use of CASToRe, a fast dictionary-based compression algorithm of the Lempel-Ziv family. We remark that definitions and indices may be equivalently defined for any reversible compression algorithm. We shall use CASToRe since it is useful in fast identification of some repeats.

The algorithm CASToRe selects a dictionary by exact matches and parses the input sequence σ in some variable-length recurrent words. Each new parsed word is the one that can be made with the longest prefix and the longest suffix already parsed. The input sequence is parsed in subwords belonging to the final dictionary relative to the sequence: $Dict(\sigma) = \{\phi_1, \ldots, \phi_t\}$.

For instance, the input sequence on alphabet $\{A, C, G, T\}$:

$$\sigma = AACACGCACGTCCGAGTCTGTC \tag{1}$$

has the following final dictionary after parsing:

$$Dict(\sigma) = \{A.A, C.A, C.G, CA.CG, T.C, CG.A, G.TC, T.GTC\}$$

where prefix and suffix are separated by a dot.

The main properties of the algorithm are shown in ref. [4].

We analysed each word in the final dictionary $Dict(\sigma) = \{\phi_1, \ldots, \phi_t\}$ by calculating the word score as follows. Each word ϕ_j is made of a prefix $\rho_p(j)$ and a suffix $\rho_s(j)$ both belonging to $\{\phi_1, \ldots, \phi_{j-1}\}$.

Then, even if the ϕ_j's are pairwise distinct (with the possible exception of the last one, ϕ_t), each ϕ_j occurs $occ(j) \geq 1$ times within the sequence σ when used as a prefix or a suffix of a subsequent word (with the possible exception of ϕ_t). For instance, given the sequence σ in above example (1), the words in the dictionary occur differently: $occ(A) = 6$, $occ(CG) = 3$, $occ(GTC) = 2$, etc.

4. Results and discussion

4.1. Word usage

First steps concern the compression of CASToRe algorithm on the *complete* gene sequences. The dictionaries resulting from that compression have been analysed and compared in order to extract common features to be helpful as a preliminary filter in the statistical linguistic investigation. We remark that this analysis is completely biologically blind, therefore it highlights structures whose importance (in recurrence, length, etc) is given by intrinsic features, typical of the sequence and not derived from external knowledge.

In our genes the most frequent word length is $\ell = 6$bp and the words statistics is meaningful about up to length $\ell = 10 - 11$bp: longer words occur only 2-3 times. It is remarkable that when the algorithm has parsed coding regions, words are mostly of length 3 and codons.

A crucial remark concerns the structure of words in the dictionaries of GRMs. The parsed words show some peculiar recurrences in the $\{A, T\} - \{G, C\}$ content. This naturally leads to the linguistic analysis we have performed. From the final dictionary relative to the complete sequences of the four genes, we extracted some *interesting* words as the words that have length ranging from 12 to 36 nt and occur from 5 to 25 times within the dictionary. Such words are well-known microsatellites (see part of them in Table 2 for *Grm1*, but the same happens for the other genes). This supports the idea that such an analysis may be extremely helpful in mining gene linguistics. Moreover, all the selected patterns belong to nonexon regions within the gene and show an evident symmetry between the occurrence of weak and strong chemical bonds among nucleotides.

Table 2: Part of interesting words selected as the longest and recurrent in the dictionary gene *Grm1* after CASToRe analysis.

atcctcctgactgcagcagtgtgaaggtggaggagt	cctgactgcagcagtgtg
aaggtggaggagtatcctcctgactgcagcagtgtg	aaaatatt
aattgaaaat	cccgagaaggaaag
tatcctcctgactgcagcagtgtg	aacacagggcccccaatggagaag

4.2. Over-represented oligonucleotides

Due to the manifest symmetry of interesting words in AT/GC bonds in the dictionary of complete genes and since such words belong to nonexon se-

quences, we built the nucleotide supersequence $Nex(g)$ made of only nonexon fragments inside each gene g and took under consideration the binary filtering of the nucleotide sequence, based on the Weak/Strong bond: $[A, T] = w,\ [G, C] = s$.

Let us assume the following null hypothesis: the $\{w, s\}$ nucleotides are distributed following a Bernoullian distribution of parameter $p = f_w$, that is the frequency of symbol 'w'. Then an oligonucleotide x of length n (n-mer, in the following) containing exactly k w-bases should appear with probability

$$p(x) = \binom{n}{k} p^k (1-p)^{n-k} \tag{2}$$

Any n-mer whose frequency exceeds that estimated probability by 5×10^{-3} is defined as over-represented. The threshold value is optimized in order to keep over-represented words both in common among genes and sufficiently long. Notice that over-represented n-mers may not be the most frequent n-mers, since being over-represented is related to a comparison with random expectation, not with frequency. The n-mer containing exactly k w-nucleotides and $h = (n - k)$ s-nucleotides shall be denoted by $n(k, h)$.

We shall extract the over-represented $\{w, s\}$-oligonucleotides of length 4 to 24 in *Nex* sequences and selct only the ones shared by the four genes, then we shall analyse only the selected 24-mers corresponding onto the *4 bases* alphabet $\{A, C, G, T\}$ on *complete* sequences.

4.2.1. Linguistic analysis

From the analysis we exploited on n-mers frequency for $n = 4, 5, \ldots, 24$, we may observe that for $n \geq 7$ the over represented m-mers contain the over represented $(m - 1)$-mers and also omo-m-mers become frequent.

The over-representation chain is shown in Table 3. At each length n, we selected *only* the over-represented n-mers that were in common for the 4 genes.

The over-represented 24-mers under the binary filter $\{w, s\}$ belong to 6 different classes, which may be roughly distinguished into two categories: from $\frac{1}{3}$ to $\frac{1}{2}$ of w =AT-content and more than $\frac{3}{4}$ of AT-content.

We remark that some classes also include some words found in the dictionary build by CASToRe (see Table 2, last two rows).

We shall now investigate such 24-mers in the light of the $\{A, C, G, T\}$ alphabet.

Table 3: Over-represented n-mers ($n = 4, \ldots, 24$) shared by the 4 genes under binary filter.

n	over-represented n-mers		
4		4(2,2)	
5		5(2,3)	
6		6(3,3)	
7		7(3,4)	
8	8(8,0)		
9	9(9,0)	9(4,5)	
10	10(10,0)		
11	11(10,1)	11(5,6)	
12	12(11,1)		
13	13(12,1) 13(11,2)	13(6,7)	
14	14(13,1) 14(12,2)		
...		.....	
...	.....		
...	.....		
24	24(9,15) 24(10,14)	24(11,13)	24(19,5) 24(20,4) 24(21,3)

For each of the 6 above $\{w, s\}$ classes, we performed the same statistical analysis on the *complete* gene sequences and on the *4 bases* alphabet $\{A, C, G, T\}$. Again, we used a multinomial null hypothesis and the over-representation is defined with threshold 2.2×10^{-4} (that is the order of the empirical frequecies of 24-mers over the 4-bases alphabet). They are grouped w.r.t. the (former) $\{w, s\}$ content and denoted by (n_A, n_C, n_G, n_T).

The over-represented common 24-mers on $\{A, C, G, T\}$ alphabet are given by two groups: balanced and omoWeak. In table 4 we show the list of omoWeak and balanced over-represented 24-mers on $\{A, C, G, T\}$ alphabet shared by the four genes.

Balanced over-represented 24-mers are 24(11,13), whose ratio AT/GC

Table 4: Over-represented 24-mers on $\{A, C, G, T\}$ alphabet shared by the 4 genes. The nucleotide content is shown as n_A, n_C, n_G, n_T.

balanced			
24(11,13)	$\begin{cases} (3,9,4,8) \\ (8,4,9,3) \end{cases}$	(4,9,4,7)	(4,10,3,7)

omoWeak			
24(19,5)	$\begin{cases} (6,4,1,13) \\ (13,1,4,6) \end{cases}$		$\begin{cases} (5,3,2,14) \\ (14,2,3,5) \end{cases}$
	(4,2,3,15)	(6,2,3,13)	(5,2,3,14)
	(5,4,1,14)	(6,3,2,13)	(12,1,4,7)
24(20,4)	$\begin{cases} (7,3,1,13) \\ (13,1,3,7) \end{cases}$	(5,3,1,15)	(6,2,2,14)
		(6,3,1,14)	(7,2,2,13)
24(21,3)		(10,1,2,11)	

is around $\frac{1}{2}$. There are only three combinatorial structure of these words: they are $A_3C_9G_4T_8$ and its reverse complement $A_8C_4G_9T_3$ and two singletons $A_4C_{\underset{10}{9}}G_{\underset{3}{4}}T_7$.

OmoWeak over-represented 24-mers are 24(19,5), 24(20,4) and 24(21,3). Within such three subclasses, several combinatorial structures occur. Notice that for some of them, also the reverse complement one is over-represented: in 24(19,5), (6,4,1,13) and (5,3,2,14) and in 24(20,4), (7,3,1,13). For the others this is not true.

As a first remark, please note that, since the gene sequences are definitely AT-rich (more than 60%), the results regarding the over-representation of balanced (GC-rich) 24-mers are anyway surprising.

We located the above two classes of over-represented *acgt*-24-mers on the complete sequence of each gene and investigated some of their features in comparison with known biological structures.

We are aware of a recent oligonucleotide analysis concerning *pyknons* [23]. The authors identified recurrent variable-length sequences (most of them is $16nt$ long) in human and mouse genomes and linked them to properties of intronic regions. Only a list of human and mouse pyknons is available and no information is given about occurrences or location of each pyknon. Therefore, the only way to make a comparison seemed to match each pyknon to our words. We found that only around 10% of balanced words had a match in (shorter) pyknons, while the fraction was about 2-4% for omoWeak words.

Anyway, we followed a further step, by observing that the selected words frequently overlap onto eachother and they result to be clustered in overlap intervals.

From now on, we shall focus on those collections of consecutive overlapping words (not on individual words) and denote them as -either balanced or omoWeak- *segments*.

The extent of segments is on average around 30 nt for every class of segments, while the longest segments reach 10^2 nt in the case of omoWeak and an average of 90 nt for balanced segments.

Some quantitative results are summarized on Table 5. In particular, the fraction of gene sequence covered by the segments of each class is a conserved property, expecially for balanced segments.

4.3. Gene organization and conservation

First of all, we located the segments on the gene sequences, w.r.t. introns and exons, making reference to the genomic sequence available on UCSC. Exons cover from 0.7% to 1% of the genes.

For what concerns balanced segments, about less than 2.5% of the segments intersect an exon in human and mouse *GRM1*, while about 1.3% in *GRM5*. Finally, omoweak segments located within exons do not exceed 0.10% in the four genes.

The vast majority of segments are completely contained in noncoding regions. When analysing what the dispersion is of each class of segments within each UCSC intronic region, we have that for the four genes the distribution of segments is almost uniform, according to the relative length of the introns (see supplementary material, section 1). Longer introns contain most of all segments. Anyway, a few exceptions are notable. For *GRM1*, there is a slight

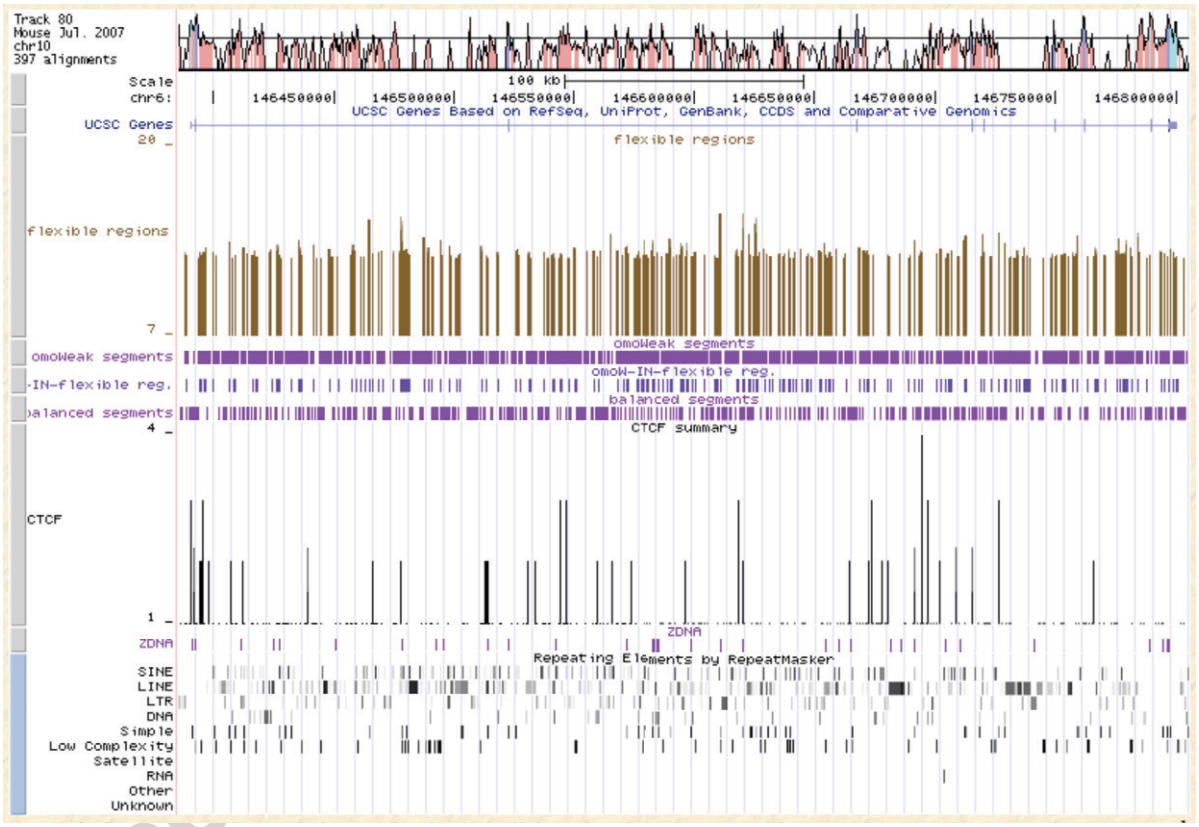

Figure 1: *Example of results for GRM1 gene: homology w.r.t.* Grm1*; gene organization; flexible regions; localization of omoWeak segments; omoWeak segments co-localizing with torsional flexibility measures; localization of balanced segments; CTCF methylation data; Z-DNA sequences; repeat classes. The coordinates of single segments and of flexible regions have been included in BED files and displayed as custom tracks on the UCSC browser.*

Table 5: Classes of over-represented 24-mers on $\{A, C, G, T\}$ alphabet shared by the 4 genes: words found, segments, and relative sequence fraction covered.

Gene	# balanced words	# segments	% covered
GRM1	1167	542	3.51%
Grm1	1210	572	3.82%
GRM5	1391	669	3.14%
Grm5	1285	592	2.88%

Gene	# omoWeak words	# segments	% covered
GRM1	6066	1532	11.15%
Grm1	3887	972	7.28%
GRM5	8050	2149	11.34%
Grm5	7182	1879	10.16%

excess of balanced segments within intron II w.r.t. intron fraction over the complete gene (1.10% of segments against relative intron length of 0.33%). For *Grm5*, its balanced segments are not contained in intron V (about 10^4 *nt* long) and for *Grm1* its omoWeak segments are not contained in intron IX (about 1.4×10^3 *nt* long). The distribution within the other introns is almost uniform.

Second, we co-localized segments w.r.t. homologous regions, in order to investigate how much such segments (which have been selected as the ones shared by the four GRM genes) are influenced by the homologies among the genomic sequences under study.

We took under consideration homology level by calculating identity by means of GenomeVista [13, 5]. Results among paralogous genes are 16% identity for *GRM1*/*GRM5* and 13% identity for *Grm1*/*Grm5* (where results are read w.r.t. *GRM1*) and homology among orthologous genes ranges from 18% identity for *GRM1*/*Grm1* to 12% identity for *GRM5*/*Grm5* (where results are read w.r.t. human genes). An example of the performed analysis is shown in figure 1.

The fraction of balanced segments localized in homology regions (w.r.t. all balanced segments) ranges from around 45% for paralogous genes to $12-15$% for orthologous genes. The fraction of omoWeak segments localized in homology regions ranges from 50% to 70% (for paralogous genes, more concentrated in human genes; for orthologous genes, more concentrated in

GRM1 genes).

We also compared the expected fraction Exp of genomic sequence covered by the balanced/omoWeak segments localized in homology regions to the observed value Obs. The value Exp was calculated as $Exp = C \times f$, where C is the fraction of conserved regions in the reference gene (i.e., the fraction of *GRM1* with homology w.r.t. *GRM5*) and f is the fraction of reference gene covered by balanced/omoWeak segments (taken from Table 5). The value Obs is the fraction of reference gene covered by balanced/omoWeak segments localized within homology regions.

This analysis resulted in an overall coincidence of the two measures (see figures on Table 6), which supports the idea of the absence of an explicit bias of the sequence conservation on the segments' distribution.

Table 6: Expected and observed genomic sequence covered by segments in conserved regions. Paralogy figures must be read w.r.t. *GRM1* genes, orthology figures w.r.t. human genes.

Homology	Exp(bal)	Obs(bal)	Exp(omoW)	Obs(omoW)
GRM1/GRM5	0.56%	1.60%	1.78%	2.63%
Grm1/Grm5	0.50%	1.57%	0.95%	1.81%
GRM1/Grm1	0.63%	0.54%	2.00%	2.59%
GRM5/Grm5	0.38%	0.30%	1.36%	1.30%

Then, we investigated whether there are any relationships among the balanced and omoWeak segments and the biological features.

We took under consideration some physical features of the DNA helix and experimentally found or computationally predicted biological properties of these inter-exon sequences.

4.4. DNA sequence and chromatin structure

DNA sequence shows different patterns throughout genomes associated with alternative conformations and diversely able to sustain chemical interactions.

In the following tables, we shall show the results by focusing on the co-localization of balanced/omoWeak segments with the features listed above. To this aim, we shall denote (for instance) the balanced segments co-localizing to some feature (say, Z-DNA configurations) by label "bal-IN-feature" and

the features co-localizing to some balanced segments by label "feature-IN-bal". Therefore, Zdna-IN-bal is the fraction of Z-DNA configurations (for each gene) co-localizing to balanced segments is w.r.t. total number of configurations, while the fraction of balanced segments co-localizing to Z-DNA configurations w.r.t. total amount of segments is the fraction bal-IN-Zdna.

4.4.1. Flexibility

The torsional flexibility of DNA helix is a sequence-dependent property expressed as fluctuations of twist angle; the potential local variations in the DNA structure may be estimated by analysing the dinucleotide degree values calculated by algorithm `stabflex` [29] based on work by Sarai and coll. [1] in overlapping windows, summed and averaged by the window length (we used windows $100nt$-long and overlapping for $99nt$). Flexible regions are known to mediate DNA-protein binding processes and are suggested to play a role in specific interactions of DNA metabolism; evidence has been reported on a relationship with DNA replication origin, DNA repair, DNase I cleavage and regulatory elements binding. DNA flexibility is involved in chromosome instability [20].

We selected the flexibility regions as follows. We selected the measures of deviation of the twist angle which are not lower than a threshold. We used $\theta = mean\ value + 2 \cdot stand\ dev$ which ranges from 12.27 to 12.46 for the different genes. Finally, we aggregate such windows (when overlapping) into wider flexible regions to which the results refer. Balanced segments show negligible matches to regions with high flexibility. Differently, flexible regions are significantly covered by omoWeak segments (see Table 7), as expected by their frequent association to AT-rich regions; more than 70% of flexible regions match some omoWeak segment, with the only exception of *Grm1* where this phenomenon is more diluted.

The co-localized flexible-omoWeak regions are distributed along the whole gene, independently of level of sequence identity.

Since the omoWeak segments co-localizing are around 20%, we investigated whether some specific omoWeak words (w.r.t. table 4) co-localize with flexible regions. More than 40% of the omoWeak21 words (21(10,1,2,11) only) match flexible regions (in *GRM1*, 47%); they represent around the 10% out of all co-localizing omoWeak words. The other combinatorial structures within omoWeak19 and omoWeak20 classes are represented for a smaller fraction.

The co-localized flexible-omoWeak regions are distributed along the whole gene, independently of the level of sequence identity.

Table 7: Segments co-localizing with flexibility peaks.

Gene	# flexible regions	flex-IN-bal	flex-IN-omoW
GRM1	275	2.2%	72.36%
Grm1	206	3.88%	56.80%
GRM5	327	3.80%	72.48%
Grm5	284	3.88%	71.48%

Gene	bal-IN-flex	omoW-IN-flex
GRM1	1.11%	21.02%
Grm1	1.57%	18.11%
GRM5	1.35%	18.43%
Grm5	1.86%	18.68%

4.4.2. Non-B DNA conformations

We consider Z-DNA and G-quadruplex structures. Z-DNA patterns, whose potential is predictable by the algorithm `zhunt` [30], occur at sequences with alternating pyrimidines and purines, -such as (CG:CG)n and (CA:TG)n- that may wind the double helix into a left-handed zigzag pattern. Therefore, unlike B-form DNA, which possesses one major groove and one minor groove, Z-DNA has only one deep and narrow groove with 12 bp per helical turn. G-quadruplex DNA is a four-stranded structure consisting of a square co-planar array of four guanines formed by a stretch of guanine-rich DNA. Each guanine acts as a donor and acceptor of Hoogsteen hydrogen bonds in a cyclic arrangement involving N-1, N-2, O-6, and N-7. G-quadruplex conformations have been predicted by `quadparser` algorithm [31]. Non-B DNA conformations show non random distributions in genomes and association with unstable regions.

Resulting data on configurations are poor (with the exception of murine Z-DNA), therefore also co-localization is weak, especially for Z-DNA conformations. Table 8 do not show figures for bal-IN- and omoW-IN- since the fraction is negligible.

4.4.3. Histones' modifications

DNA binding proteins define functional domains and chromatin activity. A high resolution map for histone modification distribution and CTCF bind-

Table 8: Segments co-localizing with non-B DNA configurations. We set 0% even when the fraction is not greater than 10^{-3}%.

Gene	# Zdna regions	Zdna-IN-bal	Zdna-IN-omoW
GRM1	33	0%	3.03%
Grm1	110	4.54%	2.73%
GRM5	68	1.47%	10.29%
Grm5	93	2.15%	2.15%

Gene	# Gquadr. regions	Gquadr-IN-bal	Gquadr-IN-omoW
GRM1	24	12%	0%
Grm1	60	13.33%	0%
GRM5	28	14.29%	0%
Grm5	88	12.5%	1.14%

ing sites is available for human genome, provided by Barski et al. [3] and Wang et al. [24].

The available data are 21 signals for human histone methylations and 18 for human histone acetylations. Each signal contains a list of peaks coming from the experimental outputs. The data has been compressed by the authors with a minor loss in resolution: for instance, if the worst case is 0.0625, the signals contained only values bigger than 1.

Methylation files range from around 200 to around 10^3 data per signal, while acetylation from 200 to 700 data per signal.

We checked whether such peaks intersect balanced or omoWeak segments.

We shall show the average results over the complete data sets provided by Barski et collaborators. The histone methylation patterns in GRM genes under study are comparable to those reported by Barsky for the transcriptional regions of active genes; high levels of H3K27me1 e H3K36me3 are present and also, even if more limited, of H3K4me1, H3K4me2, H3K9me1, H2BK5me1 e H4K20me1. The same is true for results concerning acetylation.

A limited co-localization of histone modifications with both balanced and omoWeak segments is detectable (table 9); it indicates that segments have an histone-independent pattern.

Table 9: Segments co-localizing with histone methylation and acetylation data (average results).

Gene	meth-IN-bal	meth-IN-omoW
GRM1	28.29%	47.50%
GRM5	24.71%	51.75%

Gene	bal-IN-meth	omoW-IN-meth
GRM1	41.26%	31.30%
GRM5	29.19%	22.56%

Gene	acet-IN-bal	acet-IN-omoW
GRM1	24.66%	48.25%
GRM5	25.95%	50.62%

Gene	bal-IN-acet	omoW-IN-acet
GRM1	12.80%	11.10%
GRM5	19.00%	14.06%

4.4.4. CTCF

CTCF is a multi-zinc finger protein in vertebrates that binds the insulators, i.e. DNA elements defining different chromatin domains. Thus it plays an important role in chromatin remodeling. It can dimerize when it is bound to different DNA sequences, mediating long-range chromatin looping. It mediates interchromosomal association and may direct distant DNA segments to a common transcription factor. It causes local loss of histone acetylation and gain of histone methylation. When bound to chromatin, it provides an anchor point for nucleosomes positioning. Experimental data are taken from [3].

CTCF target sites generally cluster at boundaries of chromatin domains and are scantily represented inside gene sequences; this occurs even in our genes, where low signals are present. By considering all of them for the correlation analysis with balanced and omoWeak words we found multiple matches (table 10) and a slight excess of CTCF binding sites in omoWeak (Figure 2). This may be representative of an interaction between them.

Table 10: Segments co-localizing with CTCF.

Gene	CTCF-IN-bal	CTCF-IN-omoW
GRM1	39.19%	72.89%
GRM5	33.74%	75.50%

Gene	bal-IN-CTCF	omoW-IN-CTCF
GRM1	28.60%	29.76%
GRM5	20.24%	22.21%

4.5. Repeats

We refer to the interspersed repeat databases screened by RepeatMasker that are based on the repeat databases (Repbase Update) copyrighted by the Genetic Information Research Institute [27]. We considered repeats classified on both the DNA strands, since we disregard the relationships with the transciption activity, focusing our interest only on the genomic features.

Table 11: Segments co-localizing with repeats: cumulative results.

Gene	repeats-IN-bal	repeats-IN-omoW
GRM1	28.33%	49.40%
Grm1	23.05%	28.14%
GRM5	25.32%	53.40%
Grm5	24.11%	46.99%

Gene	bal-IN-repeats	omoW-IN-repeats
GRM1	48.16%	41.09%
Grm1	41.08%	29.02%
GRM5	60.54%	45.97%
Grm5	64.87%	37.47%

We compared the segments to the repeats listed in the database and we show in table 11 the quantitative results taking into account all the repeats as a whole, w.r.t. balanced/omoWeak segments. For each segment, multiple matches can be found and *viceversa*. The fraction of repeats-IN is around 25%, on average for balanced segments but ranges from around 30% to more

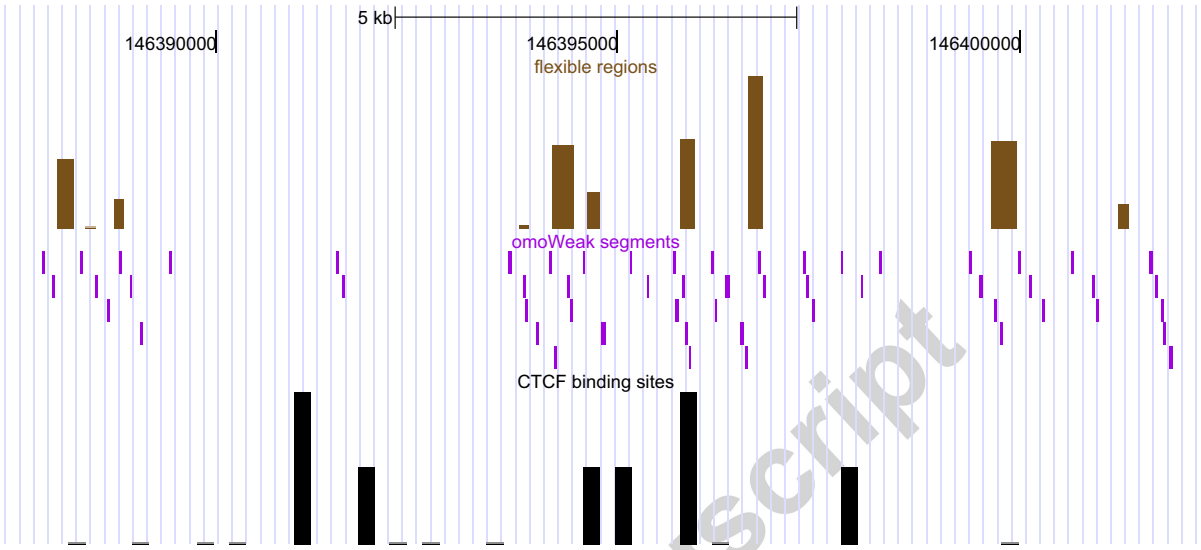

Figure 2: *Example of results for 15kb within GRM1 gene displayed as custom tracks on the UCSC browser (chr6:146,388,224-146,403,223): flexible regions; localization of omoWeak segments; CTCF methylation data.*

than 50%, on average for omoWeak segments. This means that we select only a few part of known repeats. Moreover, segments not matching to any known repeat range from 30% to 70% of the whole collection, more evident for omoWeak segments.

In supplementary material (section 3) pictures can be found, showing the composition of the co-localizing RepeatMasker sequences for the GRM genes for the complete collection of repeat classes: SINE, LINE, LTR, low complexity, simple repeats and other repeats.

Some comments are due. First, it is a common behaviour that balanced classes are extremely similar to each other, with the only exception of *Grm5*. The dispersion of RepeatMasker classes w.r.t. omoWeak matching segments seems to follow rather a specie-specific trend than to be a conserved property, expecially for the different role played by simple repeats in mouse and human genes.

5. Final remarks

We investigated what a top-down analysis of four genes, linked by either paralogy or orthology relationhips, may suggest about their biological

Table 12: Summing up what the relation is between some properties and omoWeak/balanced segments. When the relation is co-localization (resp. anti-localization), we mean that the fraction of feature-IN-omoWeak/bal is definitely greater (resp. lower) than 50%. The arrow is an additional tool to visualize such relation. When figures are around 50% we say that the feature and the linguistic class are indifferent to each other, meaning that there are no suggestions of any existing bias among the two occurrences.

Feature	$\rightarrow$?	Segments	co-localized	indifferent to eachother	anti-localized
Conservation		omoWeak		$\surd$	
Conservation		balanced		$\surd$	
Flexibility	$\rightarrow$	omoWeak	$\surd$		
Flexibility		balanced		$\surd$	
non-B DNA	$\nrightarrow$	omoWeak			$\surd$
non-B DNA	$\nrightarrow$	balanced			$\surd$
Histone meth.		omoWeak		$\surd$	
Histone meth.		balanced		$\surd$	
Histone acet.		omoWeak		$\surd$	
Histone acet.		balanced		$\surd$	
CTCF	$\rightarrow$	omoWeak	$\surd$		
CTCF		balanced		$\surd$	
repeats		omoWeak		$\surd$	
repeats		balanced		$\surd$	

features.

Starting from some combinatorial hints on recurrent words given by a preliminary compression analysis, we built a collection of DNA segments almost uniformly located along the genes, that were over-represented with respect to some biologically blind rule. We found that they are concentrated on non-exon sequences. We compared the balanced and omoWeak segments in the {w, s} alphabet, to some physical features of the DNA helix and experimentally found biological properties of these inter-exon sequences.

The resulting matches show that more than half of the complete collection of over-represented segments may be related to some already described biological property. The reverse is challenging: part of the collection shows (in human genes) meaningful relations with histone methylation and acethylation signals, whose biological interpretation remains vague since the mea-

surement values for the studied genes are extremely low.

As a general evaluation, we observe that the non negligible relationship between these segments and several biological properties, very different each other, along with their statistical significance, suggest that these segments should have a role in the functional information content of the genome. This conjecture is strongly supported by the conservation we have found for these over-expressed words in our comparisons between orthologous and paralogous genes. Even though the very low similarity in their sequences (less than 20% in the average case) the conservation of the over-expressed words is very high.

The over-expressed words would play some structural/functional role which is important and then conserved beyond the plain sequence preservation. The very similar outcome of co-localization analysis for omoWeak with CTCF signals and flexible regions would suggest a relationship of these words with some aspects of chromatin structure. Indeed CTCF binding sites regulate chromatin domains activity [8] whilst flexible regions mediate DNA-protein binding processes and specific interactions of DNA metabolism [17]. Interestingly, both these sequence-dependent parameters that are not related to specific sequences, colocalize with words along the whole genes independently of sequence identity. These findings may be suggestive of a conserved long-range structure, with a possible functional role, and of a higher-order organization. The impossibility of further focusing the words role is probably due to the present lacking of more detailed biological knowledge.

This paper would represent a suggestion directed to genomists to further investigate, from an experimental point of view, the possible role of such segments.

Supplementary material

We provide some additional data about gene organization, co-localization w.r.t. CpG islands and w.r.t. individual classes of repeats.

ACKNOWLEDGEMENTS

The work of G.M. was supported by a post-doc research scholarship "Compagnia di San Paolo" awarded by the Istituto Nazionale di Alta Matematica "F. Severi". The Authors want to thank Andrea Bedini for having made available the software used to compute DNA flexibility and Francesco Morandin for useful suggestions.

References

[1] Arazo-Bravo MJ, Fujii S, Kono H, Ahmad S, Sarai A., Sequence-dependent conformational energy of DNA derived from molecular dynamics simulations: toward understanding the indirect readout mechanism in protein-DNA recognition, J Am Chem Soc. 127 (2005) 16074-16089.

[2] Arvey A.J., Azad R.K., Raval A., Lawrence J.G., Detection of genomic islands via segmental genome heterogeneity, Nucl Acids Res, 16 (2009) 5255-5266.

[3] Barski A, Cuddapah S, Cui K, Roh TY, Schones DE, Wang Z, Wei G, Chepelev I, Zhao K., High-resolution profiling of histone methylations in the human genome, Cell,18 (2007) 823-37.

[4] Bonanno C., Menconi G., Computational Information for the logistic map at the chaos threshold, Discrete and Continuous Dynamical Systems - B, 2 (2002)415-43.

[5] Bray, N., Dubchak, I. and Pachter, L. , AVID: A Global Alignment Program, Genome Research, 13 (2003) 97.

[6] Bultrini E., Pizzi E., Del Giudice P., Frontali C., Pentamer vocabularies characterizing introns and intron-like intergenic tracts from C. elegans and D. melanogaster, Gene, 304 (2003)183-192.

[7] Brendel V., Beckmann J.S., Trifonov E.N., Linguistics of nucleotide sequences: morphology and comparison of vocabulary. J. Biomol. Struct. and Dynamic, 4 (1986) 11-21.

[8] Burke LJ et al., CTCF binding and higher order chromatin structure of the H19 locus are maintained in mitotic chromatin,The EMBO Journal 24 (2005) 3291 - 3300

[9] Champ PC, Maurice S, Vargason JM, Camp T, Ho PS, Distributions of Z-DNA and nuclear factor I in human chromosome 22: a model for coupled transcriptional regulation, Nucleic Acids Research, 32 (2004) 6501-10.

[10] Corá D, Di Cunto F, Caselle M, Provero P., Identification of candidate regulatory sequences in mammalian 3' UTRs by statistical analysis of oligonucleotide distributions, BMC Bioinformatics, 8 (2002) 174.

[11] Corti C., R.W. E. Clarkson, L. Crepaldi, C.F. Sala,J.H. Xuereb, F. Ferraguti, Gene Structure of the Human Metabotropic Glutamate Receptor 5 and Functional Analysis of Its Multiple Promoters in Neuroblastoma and Astroglioma Cells, J. Biol. Chem, 278 (2003) 33105-33119.

[12] Crepaldi L., C. Lackner, C. Corti, F. Ferraguti, Transcriptional Activators and Repressors for the Neuron-specific Expression of a Metabotropic Glutamate Receptor, J. Biol. Chem., 282 (2007) 17877-17889.

[13] Couronne O, Poliakov A, Bray N, Ishkhanov T, Ryaboy I, Rubin E, Pachter R, Dubchak I, Strategies and Tools for Whole-Genome Alignments, Genome Research, 13 (2003) 7.

[14] Ferragina P., Giancarlo R., Greco V., Manzini G., Valiente G., Compression-based classification of biological sequences and structures via the Universal Similarity Metrix: experimental assessment, BMC Bioinformatics, 8 (2007) 252.

[15] Ferraguti F, Crepaldi L, Nicoletti F., Metabotropic glutamate 1 receptor: current concepts and perspectives, Pharmacol Rev., 60 (2008) 536-81.

[16] Gabrielian A., Bolshoy A., Sequence complexity and DNA curvature, Computers and Chemistry, 23 (1999) 263-274.

[17] Grove A., Galeone A., Mayol L, Geiduschek P.E., Localized DNA Flexibility Contributes to Target Site Selection by DNA-bending Proteins, J. Mol. Biol. 260 (1996) 120-125

[18] Kirzhner V., Nevo E., Korol A., Bolshoy A., A large-scale comparison of genomic sequences: one promising approach, Acta Biotheoretica, 51 (2002) 73-89.

[19] Menconi G., Marangoni R., A compression-based approach for coding sequences identification I. Prokaryotic genomes, J. Comput. Biol., 13 (2006) 1477-1488.

[20] Puliti A., Rizzato C., Conti V., Bedini A., Gimelli G., Barale R., Sbrana I. (2010) Low-copy repeats on chromosome 22q11.2 show replication timing switches, DNA flexibility peaks and stress inducible asynchrony, sharing instability features with fragile sites, Mutat. Res. 686 (2010) 74-83.

[21] Robbins R.J., Challenges in the Human genome project, IEEE Engineering in Medicine and Biology, 11(1992) 25-34 .

[22] Stern L., Allison L., Coppel R.L., Dix T.I., Discovering patterns in Plasmodium falciparium genomic DNA, Molecular and Biochemical Parasitology, 118 (2001) 175-186.

[23] Tsirigos A., Rigoutsos I., Human and mouse introns are linked to the same processes and functions through each genomes most frequent non-conserved motifs, Nucl Acids Res, 36 (2008) 3484-3493.

[24] Wang Z, Zang C, Rosenfeld JA, Schones DE, Barski A, Cuddapah S, Cui K, Roh TY, Peng W, Zhang MQ, Zhao K. Combinatorial patterns of histone acetylations and methylations in the human genome, Nat. Genet., 40 (2008) 897-903.

[25] Wheeler B.S., Blau J.A., Willard H.F., Scott K.C., The Impact of Local Genome Sequence on Defining Heterochromatin Domains, PLoS Genet., 5(4): e1000453. doi:10.1371/journal.pgen.1000453

[26] http://genome.ucsc.edu/
homo $GRM1$: hg18_dna range=chr6:146385472-146805427;
homo $GRM5$: hg18_dna range=chr11:87872389-88443761;
mouse $GRM1$: mm9_dna range=chr10:10403555-10807154;
mouse $GRM5$: mm9_dna range=chr7:94727889-95287655.

[27] RepeatMasker by A.F.A. Smit, R. Hubley and P. Green at http://repeatmasker.org and
Genetic Information Research Institute at http://www.girinst.org/

[28] http://www.ncbi.nlm.nih.gov/unigene

[29] `stabflex` available at http://78.40.125.79/stabflex/

[30] `zhunt` available at http://gac-web.cgrb.oregonstate.edu/zDNA/

[31] `quadparser` available at http://www.quadruplex.org/?view=quadparser_web