Università di Pisa
Dipartimento di Informatica

Technical Report: TR-09-14

What a top-down linguistic analysis can tell about genomic sequences: The metabotropic Glutamate receptors 1 and 5 in Human and in Mouse as a case study

Giulia Menconi [a], 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

Address: Largo B. Pontecorvo 3, 56127 Pisa, Italy. Tel: +39 050 2212700 Fax: +39 050 2212726

What a top-down linguistic analysis can tell about genomic sequences: The metabotropic Glutamate receptors 1 and 5 in Human and in Mouse as a case study

Giulia Menconi [a], 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 bottom up strategy to detect features in genomic sequences. The strategys core is to exploit dictionary-based compression algorithms and analyze the content of the automatically generated dictionary. This paper preliminary try to classify the different over-represented words and to correlate them to experimentally identified or theoretically forecasted biological features. Among the various results obtained, we found a correlation between a particular class of words and the DNA flexibility and a strong anti-correlation between any over-represented word class and the high histone methylation signals. This could suggest that the DNA sequences located in correspondence of highly methylated sites are of the hypervariable origin, as they do not host any over-represented word.

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). To develop 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, as in the very bright analogy suggested by Robbins [15] and here recalled:
Consider the 3.3 gigabytes of a human genome as equivalent to 3.3 gigabytes of files on the mass-storage device of some computer system of unknown design. Obtaining the sequence is equivalent to obtaining an image of the contents of

that mass-storage device. Understanding the sequence is equivalent to reverse engineering that unknown computer system (both the hardware and the 3.3 gigabytes of software) all the way back to a full set of design and maintenance specifications.

Reverse engineering the sequence is complicated by the fact that the resulting image of the mass-storage device will not be a file-by-file copy, but rather a streaming dump of the bytes in the order they occupied on the device and the files are known to be fragmented. In addition, some of the device is known to contain erased files or other garbage. Once the garbage has been recognized and discarded and the fragmented files reassembled, the reverse engineering of the codes must be undertaken with only a partial, and sometimes incorrect understanding of the CPU on which the codes run. In fact, deducing the structure and function of the CPU is part of the project, since some of the 3.3 gigabytes are known to be the binary specifications for the computer-assisted-manufacturing process that fabricates the CPU. In addition, one must also consider that the huge database also contains code generated from the result of literally millions of maintenance revisions performed by the worst possible set of kludge-using, spaghetti-coding, opportunistic hackers who delight in clever tricks like writing self-modifying code and relying upon undocumented system quirks.

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 [20] . 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

[14] 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 [19], [17], [**?**].

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 [18].

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 [17] [19] [16], or to classify coding/non-coding sequences on the basis of the compression factor or similar indexes [10].

We chose four genomic sequences to test our approach, two paralogous couples that are, in their turn, orthologs: the Glutamate metabotropic (mGlu) receptors 1 and 5 in the Human and in the Mouse genome.

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 [10].

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. Then, we perform a statistical linguistic analysis focused on the nonexon part of the genes. Finally, we show whether and what the relationships are of the above results with known local biological properties, especially with known repetitive sequences.

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 [22]. 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 [8]. 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 [7]. 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 [7]. 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 rceptors 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

We shall consider the following four genes [12]: $GRM1$ and $GRM5$ in *Homo sapiens* and $GRM1$ and $GRM5$ in *Mus musculus* (hereafter indicated as $HGRM1$, $HGRM5$, $mGRM1$ and $mGRM5$, respectively). They all are metabotropic Glutamate receptors and they all share a low GC content. Human sequences were from NCBI Build 36.1 and mouse sequences from Build 37 (UCSC Genome Bioinformatics). 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 [7]). The DNA strand that has been analysed is that indicated by the UCSC browser as coding strand (plus strand). The analysed sequence includes the $5'$ 500000 bp upstrem to the first exon, and the 500000 bp 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.

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	length	GC-content
$HGRM1$	412965 bp	37%
$HGRM5$	563148 bp	36%
$mGRM1$	398962 bp	39%
$mGRM5$	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. [3].

We also 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

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 $\ell = 6$ and the words statistics is meaningful about up to length $\ell = 10-11$: 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 mGRM1, 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 $mGRM1$ after CASToRe analysis.

tgtgtgtgtgtgtgtgtgtgtg	acacacacacacacacacacac
ttttttttttt	agagagagagagagagagagagagagagag
cctgactgcagcagtgtg	atcctcctgactgcagcagtgtgaaggtggaggagt
aaggtggaggagtatcctcctgactgcagcagtgtg	ctatctatctatctatctatctatctatct

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 sequences, 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 3×10^{-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.

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

n	over-represented n-mers				
4	4(2,2)				
5	5(2,3)				
6	6(3,3)				
7	7(3,4)			7(7,0)	
8	8(4,4)			8(8,0)	
9	9(4,5)			9(9,0)	
10	10(4,6)			10(9,1)	10(10,0)
11	11(5,6)			11(10,1)	11(11,0)
12	12(5,7)			12(11,1)	12(12,0)
...	.....				
...	.....				
...	.....				
24	24(8,16) 24(9,15) 24(10,14)	**24(11, 13)** balanced	24(12,12)	**24(19, 5)** omo-Weak	24(20,4) 24(21,3) 24(22,2)

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

For each of the 9 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 Bernoullian null hypothesis and the over-representation is defined with threshold 3×10^{-3}. 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 is around $\frac{1}{2}$. There is only one combinatorial structure of these words: they are $A_5C_8G_5T_6$.

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)	24(5,8,5,6)		
omoWeak			
24(19,5)	A	(6,3,2,13)	(7,2,3,12) (7,3,2,12)
	B	(8,2,3,11) (11,2,3,8)	(8,3,2,11) (11,3,2,8)
	C	(9,2,3,10) (10,2,3,9)	(9,3,2,10) (10,3,2,9)

OmoWeak over-represented 24-mers are 24(19,5). The number of G or C bases is only either $2+3$ or $3+2$, then the differences among omoWeak words may be identified by observing the symmetry in the $A+T$ content. Due to the fact that the 24-mers are grouped by their plain ACGT content (i.e. modulo site basis permutations), we are obviously associating reverse sequences to each other. Then, the combinatorial structures may be classified under three types: A-content not exceeding 7 nt (omoWeak A), A-content either 8 or 10 nt (omoWeak B), A-content either 9 or 10 nt (omoWeak C). Please notice that the last two types are complete (all the four possible combinatorial structures are present, given the above GC bias), while the first type has only a few elements.

Summing up, the combinatorial structures of omoWeak words are the following: omoWeak A $= \{A_6C_3G_2T_{13},\ A_7C_{3}^{2}G_{2}^{3}T_{12}\}$; omoWeak B $= \{A_{11}^{8}C_{3}^{2}G_{2}^{3}T_{8}^{11}\}$ and omoWeak C $= \{A_{10}^{9}C_{3}^{2}G_{2}^{3}T_{9}^{10}\}$.

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.

5 Discussion

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* [11]. 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 8-9% of balanced words had a match in (shorter) pyknons, while the fraction was negligible 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 summarized in Fig. 1: the average length is around 30 nt for every class of segments, while the longest segments reach 10^2 nt in the case of some omoWeak B and C. Moreover, the number of overlapping words within each segments seems to grow linearly w.r.t. segment length, at least for longest segments (Figure 2 shows the case of omoWeak C segments in HGRM1).

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

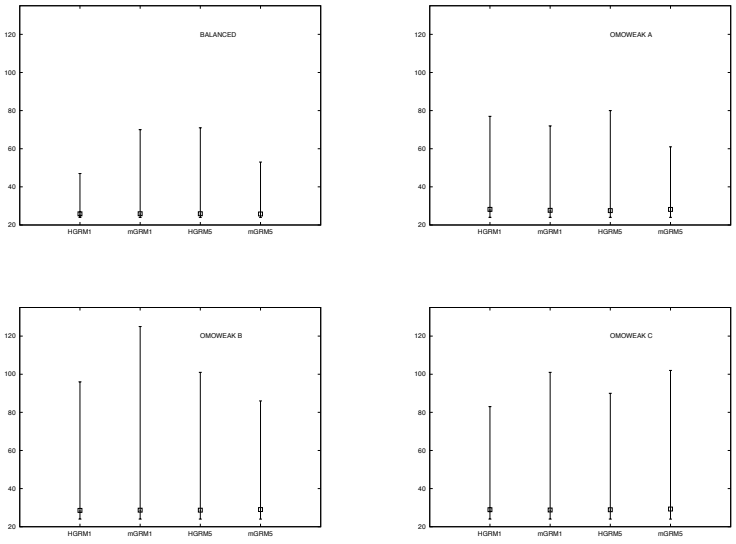

Figure 1: Minimum, average (□) and maximum length of overlap intervals.

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.

Only about less than 1% of the segments intersect an exon (minimum 0.24% for balanced segments in mGRM5, maximum 2% for omoWeak A segments in mGRM1).

The vast majority of segments are completely contained in noncoding regions. As an example, we show in picture 3 what the dispersion is of each class of segments within each UCSC intronic region, for $hGRM1$. The results are analogous for the other genes (data not shown). It is clear that the distribution of segments is almost uniform, according to the relative length of the introns. Longer introns contain most of all segments. A few exceptions are anyway notable: for instance, see balanced segments in intron 8 in HGRM1 or balanced segments in intron 2 in HGRM5.

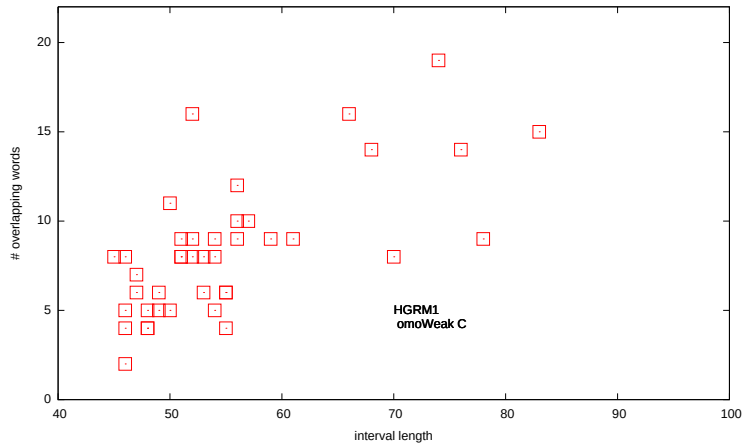

Figure 2: Longest segment lengths (more than 30 *nt*) vs nr. of overlapping words in HGRM1 omoWeak C class.

HGRM1

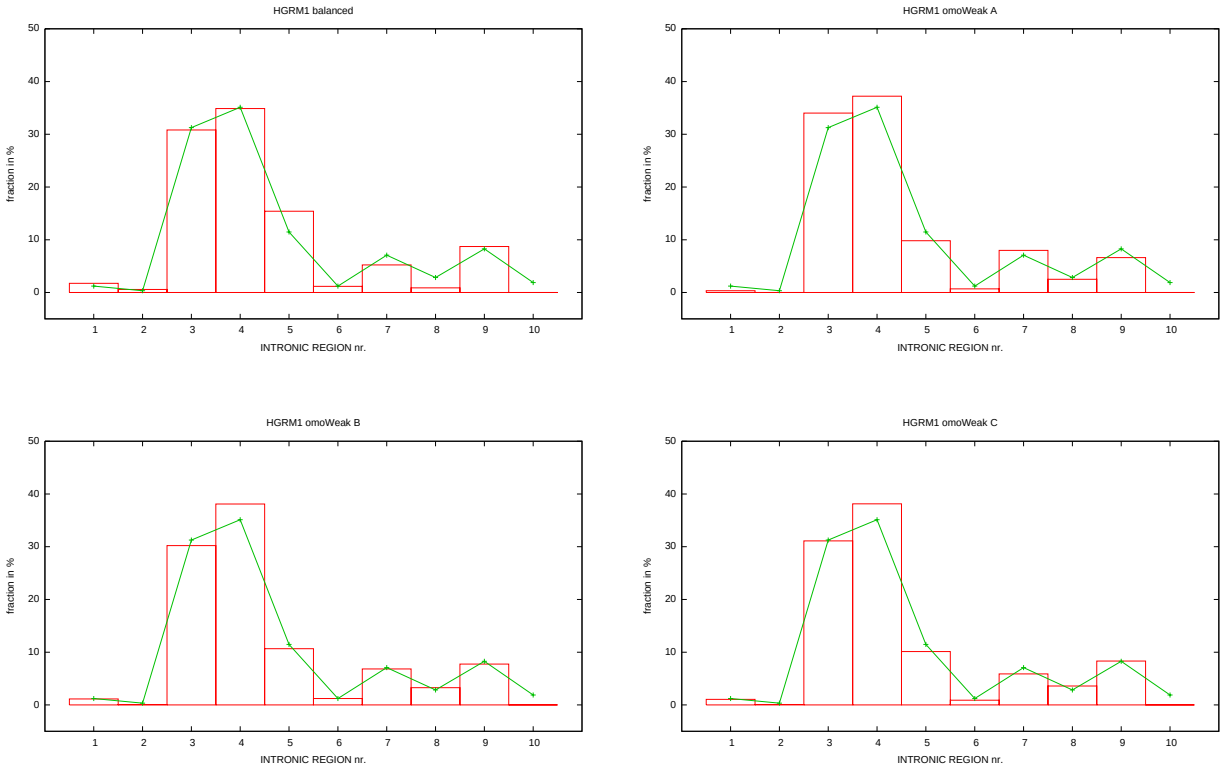

Figure 3: *HGRM1: Fraction of segments within each intron (this gene has 9 coding exons in UCSC genomic sequence). The dotted line refers to the fraction of complete gene occupied by each intron.*

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

	HGRM1	mGRM1	HGRM5	mGRM5
# balanced words	573	635	715	704
# segments	344	387	410	424
% covered	2.15%	2.51%	1.88%	1.97%
# omoWeak A words	2315	1395	2986	2667
# segments	876	570	1231	1036
% covered	5.97%	3.95%	6.02%	5.28%
# omoWeak B words	3225	2158	5082	4358
# segments	1407	939	2151	1820
% covered	9.73%	6.75%	10.95%	9.55%
# omoWeak C words	3703	2429	5659	4889
# segments	1225	826	1901	1576
% covered	8.57%	5.95%	9.76%	8.35%

Second, 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 biological properties of these inter-exon sequences:

- flexibility and stability: by means of the calculations developed by Sarai and coll. in 2005 [2], we have studied the sequence dependence of flexibility and its anisotropy along various conformational variables of DNA base pairs. Flexibility and stability are anticorrelated: flexibility is higher in AT-rich (omoWeak) regions and consequently stability is lower.

- CpG-islands: CpG islands are associated with genes, particularly housekeeping genes, in vertebrates. CpG islands are typically common near transcription start sites, and may be associated with promoter regions. Normally a C (cytosine) base followed immediately by a G (guanine) base (a CpG) is rare in vertebrate DNA because the Cs in such an arrangement tend to be methylated. This methylation helps distinguish the newly synthesized DNA strand from the parent strand, which aids in the final stages of DNA proofreading after duplication. However, over evolutionary time, methylated Cs tend to turn into Ts because of spontaneous deamination. The result is that CpGs are relatively rare unless there is selective pressure to keep them or a region is not methylated for some other reason, perhaps having to do with the regulation of gene expression. CpG islands are regions where CpGs are present at significantly higher levels than is typical for the genome as a whole.

- histones methylation: We refer to the data reported by Barski et al. [1] For each histone protein and each Lysine that can be methylated and for every methylation degree (1, 2 or 3 methyles) there is a distinct signal.

- DNA-polymerase II binding sites: We refer again to Barski et al. [1]. The measures were similar to that of Methylation.

The above two characteristics (methylation and pol II binding sites) are available only for HGRM1.

5.1 Flexibility and stability

We selected the flexibility peaks as the regions whose flexibility value is not lower than a threshold (we used $mean\ value + 2 \cdot stand\ dev$). According to the above remark on the fact that stability should be lower in omoWeak segments, balanced segments show poor matches to regions with high flexibility, while omoWeak segments are highly correlated and (especially omoWeak B and C) selected peak regions are significantly covered by such AT-rich over-represented segments (see figures about "% peaks matching" in Table 6).

Table 6: Segments matching with flexibility peaks: the fraction of peaks matching in each class of segments is w.r.t. total number of selected peaks, while the fraction of matching segments is w.r.t. total amount of segments.

	HGRM1	mGRM1	HGRM5	mGRM5
# peaks	89	101	123	127
% peaks matching in balanced	26.97%	47.52%	18.70%	28.35%
% peaks matching in omoWeak A	58.43%	55.45%	69.92%	55.91%
% peaks matching in omoWeak B	85.39%	69.31%	81.30%	63.78%
% peaks matching in omoWeak C	73.03%	60.40%	78.86%	62.20%
% matching balanced segments	4.94%	5.94%	3.90%	4.95%
% matching omoWeak A segments	7.31%	6.84%	7.31%	4.73%
% matching omoWeak B segments	7.39%	5.96%	6.65%	4.29%
% matching omoWeak C segments	6.45%	5.93%	7.21%	4.51%

5.2 CpG islands

We downloaded the CpG-island maps for all the considered genes from the UCSC genome database. We discovered that these genes have a very low number of islands (minimum 1 for 3 genes, maximum 2 in HGRM1). There are multiple matches of omoWeak segments onto the islands:

- in HGRM1, CpG-44 intersects 3 omoWeak A, 3 omoWeak B and 2 omoWeak C segments, while CpG-22 intersects 1omoWeak C segment.
- CpG-17 in mGRM1 intersects once with an omoWeak C segment.
- CpG-89 in HGRM5 intersects 5 omoWeak B and 2 omoWeak C segments.
- CpG-48 in mGRM5 intersects 2 omoWeak A, 5 omoWeak B and 4 omoWeak C segments.

5.3 Histone methylation

The available data are 22 signals for HGRM1 histone methylation and one regarding the DNA-polymerase II binding sites for the same gene. Again, since the measures came from a sliding window experiment, we selected only the methylation values exceeding $mean\ value + 2 \cdot stand\ dev$.

As a result, only in 3 signals (all relative to histone protein $H3$) we found very scarce matches: $H3K4me3$ (2 omoWeak A segments matching over around 300 peaks), $H3K9me2$ (1 balanced segment matching over around 500 peaks) and $H3K9me3$ (1 balanced,1 omoWeak B and 2 omoWeak C segments matching over around 300 peaks).

These results clearly show that over-represented segments and methylation peaks occurences are mutually exclusive. Histone methylation has been correlated to heterochromatin formation, and then indirectly linked to the regulation of gene expression. The question about a possible correlation between high methylation level in the histones and the DNA sequence bound on them has not been solved yet. In a recent paper [4], the application of computational methods lead to a strong indication that the sequence content may have a prominent role in heterochromatin formation, and, then, in histone methylation. Nevertheless, the nature of these DNA sequences remains unclassified. Now, the results we obtained seem to suggest that they belong to a hypervariable nature, as they never host over-represented segments of any type.

5.4 Repeats

As a second step, we focused on known 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 [13]. 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.

We compared the segments to the repeats listed in the database and we grouped our balanced and omoWeak words as either *inside-known-repeats* or *outside-known-repeats*.

Table 7 shows the quantitative results. For each segment, multiple matches can be found and *viceversa* (this motivates that fact that the number of matches may exceed the number of segments as in previous Table 5). The fraction of inside-known-repeats w.r.t. known repeats is around 20%, on average for balanced words but ranges from 30% to 50%, on average for omoWeak words. This means that we select only a few part of known repeats. Moreover, segments not matching to any known repeat are around 50% of the whole collection. Figures from 4 to 6 show the composition of the inside RepeatMasker sequences for the GRM genes for the complete collection, for long segments (longer than $mean\ value + 2 \cdot stand\ dev$) and for short segments.

Some comments are due. First, it is a common behaviour that omoWeak classes are extremely similar to each other, for any gene. Therefore, since we introduced those 3 classes trying to extract any characterization of omoWeak segments, we may conclude that those classes do not suggest to relate to different specific known repeats, but to the repeat structure characteristic of the gene under examination.

Second, any further distinction we tried (long, short, RepeatMasker class, etc.) showed no relation with neither homology nor paralogy relationships: it seems that, with respect to overrepresented intervals, each gene behaves on its own.

Table 7: Inside- and outside-known-repeats: the amount of matches, the fraction of matched known repeats and the fraction of not matching segments.

	HGRM1	mGRM1	HGRM5	mGRM5
balanced segments				
# inside repeats	174	153	210	240
% matching repeats/known repeats	20.4%	17.96%	16.23%	17.52%
# outside repeats	177	240	203	192
% outside/balanced	51.45%	62.02%	49.51%	45.28%
omoWeak A segments				
# inside repeats	462	246	641	514
% matching/known	41.32%	24.55%	36.95%	33.04%
# outside repeats	430	332	606	547
% outside/omoWeak A	49.09%	58.25%	49.23%	52.80%
omoWeak B segments				
# inside repeats	643	333	1052	872
% matching/known	51.26%	32.34%	50.66%	44.42%
# outside repeats	784	624	1124	985
% outside/omoWeak B	55.72%	66.45%	52.25%	54.12%
omoWeak C segments				
# inside repeats	533	287	936	732
% matching/known	45.87%	29.04%	50.44%	40.85%
# outside repeats	705	546	998	880
% outside/omoWeak C	57.55%	66.10%	52.50%	55.84%

6 Final remarks

We investigated what a top-down analysis of four genes may suggest about their biological 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 were concentrated on non-exon sequences. We compare the balanced and omoWeak segments to some physical features of the DNA helix and experimentally found biological properties of these inter-exon sequences. The resulting matches show that half of the complete collection of over-represented segments may be related to some already established property. Challenging is the reverse: half of the complete collection does not match with any analysed properties (see Table 8). Again, the dispersion w.r.t. each intron is definitely comparable to the relative extent of the intron w.r.t the compete gene sequence (an example is plotted on Figure 7 for HGRM1). The results for the other genes are analogous (data not shown). The meaning of these segments is the base for further investigations, expecially from the experimental point of view.

Figure 4: *classification of all the segments shared with marked sequences in ReapeatMasker.*

Acknowledgment

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.

References

[1] 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. 2007 May 18;129(4):823-37.

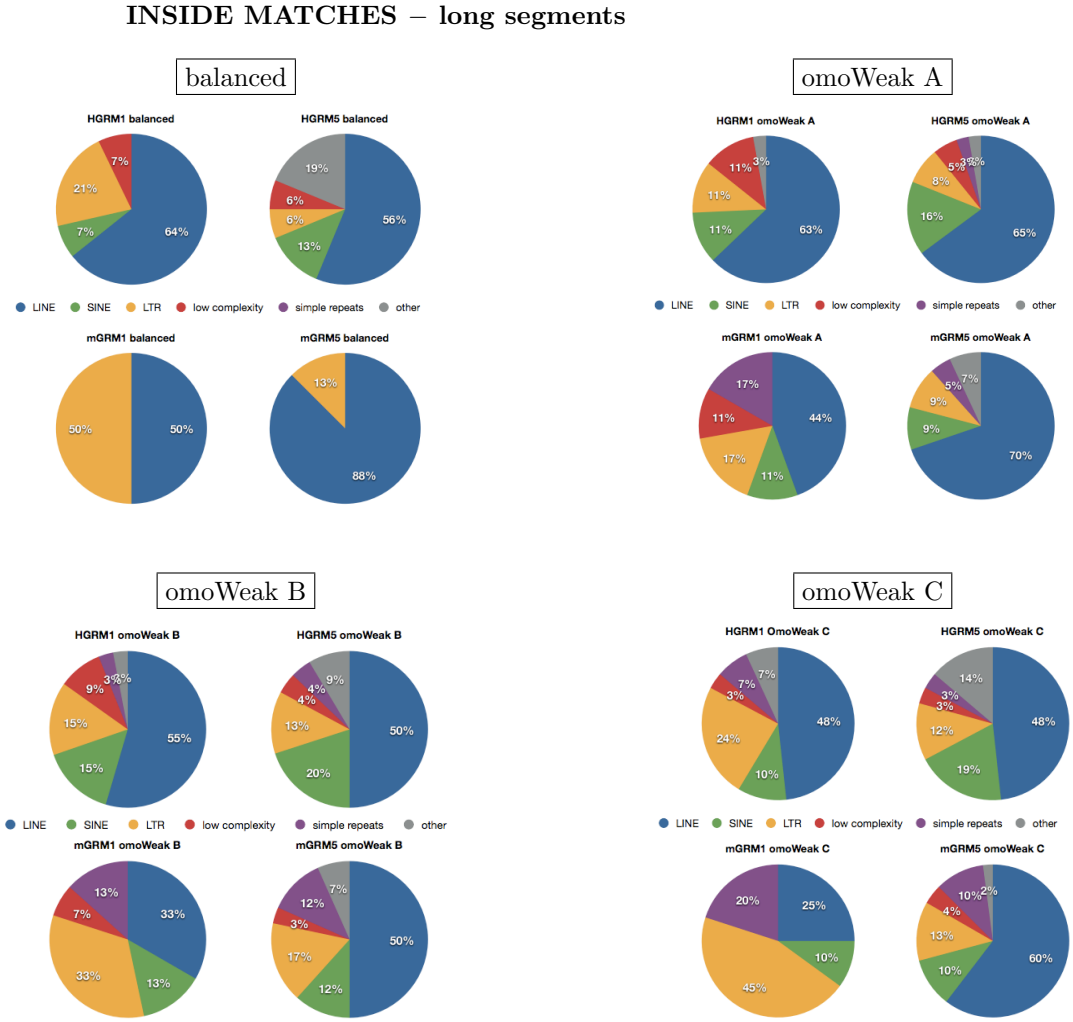

Figure 5: *only long intervals analyzed here.*

[2] 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. 2005 Nov 23;127(46):16074-89.

[3] Bonanno C., Menconi G., "Computational Information for the logistic map at the chaos threshold", *Discrete and Continuous Dynamical Systems - B*, **2**, 3, 415–431 (2002).

[4] 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

[5] Bultrini E., Pizzi E., Del Giudice P., Frontali C., Pentamer vocabularies characterizing introns and intron-like intergenic tracts from C. elegans and

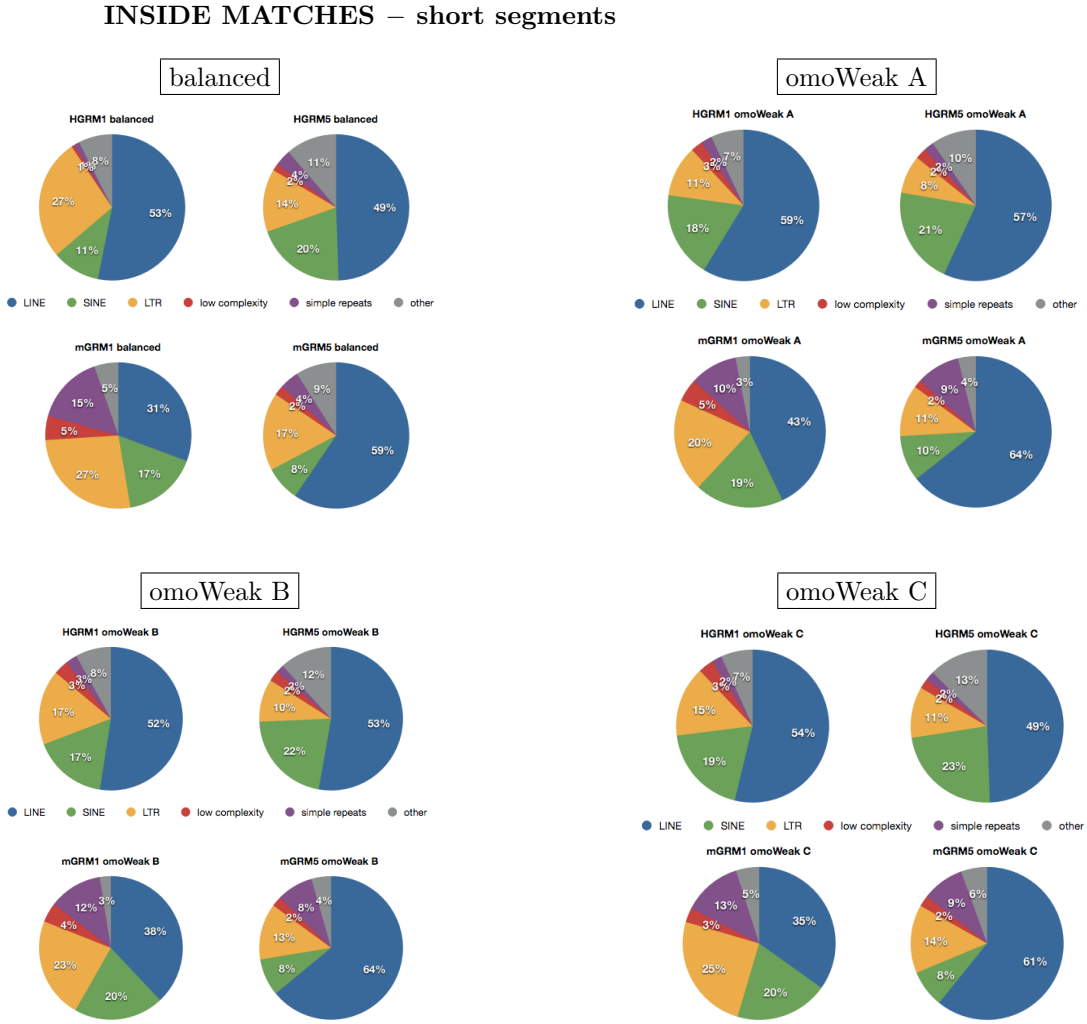

Figure 6: *Only short intervals analyzed here.*

D. melanogaster, Gene, 304, 183-192 (2003).

[6] Corti and coll. (2003) J. Biol. Chem, 278(35)

[7] Crepaldi and coll. (2007) J. Biol. Chem., 282 (24)

[8] Ferraguti F, Crepaldi L, Nicoletti F., Metabotropic glutamate 1 receptor: current concepts and perspectives, Pharmacol Rev. 2008 Dec;60(4):536-81.

[9] R.N. Mantegna, et al., Linguistic Features of Noncoding DNA Sequences, Phys. Rev. Lett. 73 (23) (1994) 31693172.

[10] Menconi G., Marangoni R., "A compression-based approach for coding sequences identification I. Prokaryotic genomes", *Journal of Computational Biology* , **13**,8 (2006).

Table 8: No-match-at-all segments.

	HGRM1	mGRM1	HGRM5	mGRM5
% no-match balanced segments	48.84%	58.66%	47.32%	42.92%
% within exons	1.20%	0.44%	0.52%	0.56%
% no-match omoWeak A segments	45.32%	53.51%	45.00%	49.32%
% within exons	1.00%	2.62%	1.99%	1.37%
% no-match omoWeak B segments	51.60%	62.30%	48.30%	50.93%
% within exons	1.10%	1.53%	1.54%	1.19%
% no-match omoWeak C segments	53.71%	62.47%	48.50%	52.54%
% within exons	1.06%	1.36%	1.41%	1.21%

HGRM1

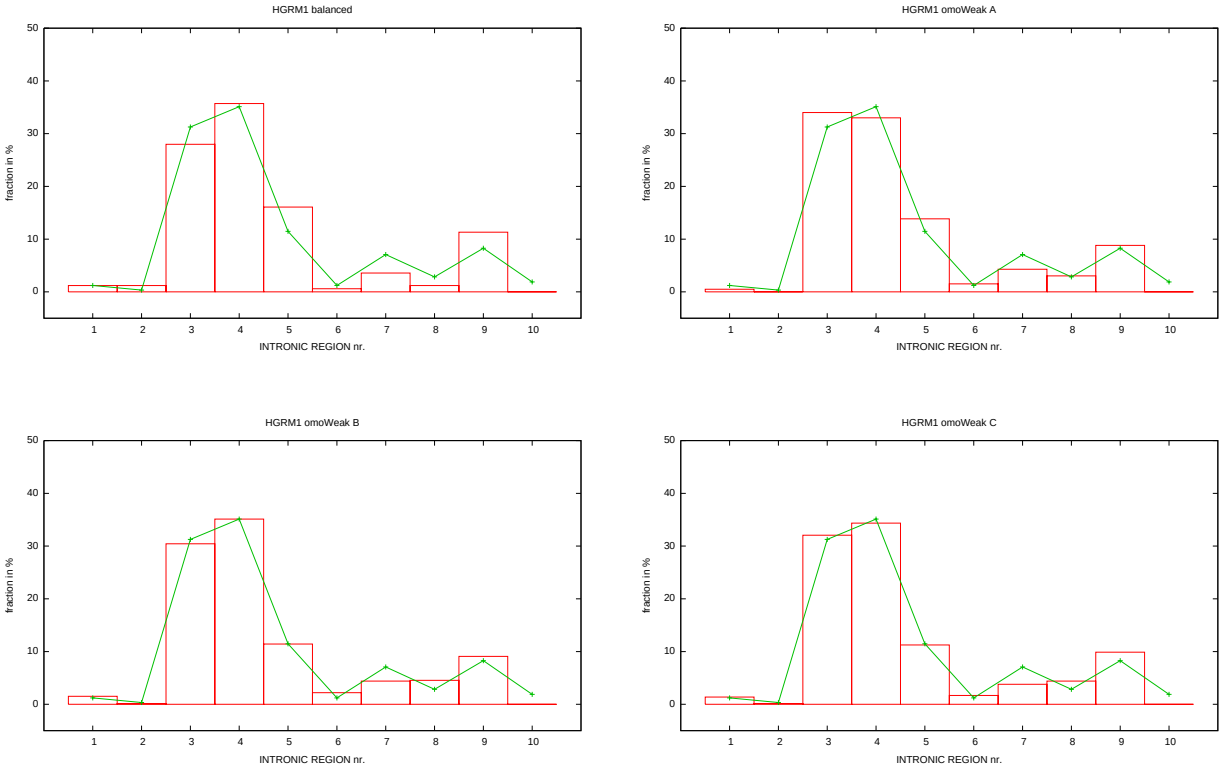

Figure 7: *HGRM1: Fraction of no-match-at-all segments within each intron (this gene has 9 coding exons in UCSC genomic sequence). The dotted line refers to the fraction of complete gene occupied by each intron.*

[11] Tsirigos A., Rigoutsos I., "Human and mouse introns are linked to the same processes and functions through each genomes most frequent non-conserved motifs", Nucleic Aicds Research (2008), 36,10, 3484-3493.

[12] 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.

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

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

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

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

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

[18] 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**, 252 (2007).

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

[20] Arvey A.J., Azad R.K., Raval A., Lawrence J.G., "Detection of genomic islands via segmental genome heterogeneity", *NAR*, **1-12**, (2009)

[21] 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. 2007 May 24;8:174.

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