

Computational Analysis of the Distribution of Short Repeated Motifs in Human Genomic Sequences

Abanish Singh* and Nikola Stojanovic*

Keywords: DNA repeated sequences, DNA functional elements, DNA sequence motifs

Abstract

Over the last several years the search for functional elements in human and other genomes by exploiting motif over-representation became increasingly popular. However, about half of the human genome consists of known repeated elements, and that is also the case with most higher eukaryotes. In this study we have shown that in addition to these known repeats, human genomic sequence features many short motifs which are significantly over-represented, and that their frequency varies only slightly between random repeat-masked sequences and regions located immediately upstream of the known genes. As the ongoing ENCODE project is set on the development of the techniques for high-throughput identification of the functional elements in the human genome, concentrating on about 1% of its entire DNA sequence, we have chosen these regions for a part of our study.

1 Introduction

The search for transcription factor binding sites is one of the most popular subfields of bioinformatics, and many algorithms have been developed over more than a decade of intensive research. The first approaches relied on a rather naive assumption that

the target sites for protein binding must feature information content sufficient for them to be recognized. Disillusionment soon followed, as any attempt to isolate functional elements in DNA resulted in an enormous number of false positives. Recent approaches have thus concentrated on the incorporation of additional information to the raw sequence data. They often relied on the phylogenetic conservation [20, 10, 18, 7] or a search for clusters of elements whose sequences match experimentally confirmed consensus motifs [10, 18] retrieved from databases such as TRANSFAC [15] or Jaspar [17]. The latter methods exploited the fact that proteins involved in the initiation of transcription rarely, if ever, act in isolation.

With the advances in microarray technology large sets of putatively co-expressed genes became available. This, in turn, stimulated the development of new techniques aiming to detect conserved motifs in the upstream sequences of these genes [9, 10, 4]. It is intuitive that if a group of genes is coordinately regulated, it should be controlled by the same transcription factors. From the hypothesis that protein binding is directed by DNA sequence motifs it follows that same motifs should be present in all observed upstream regulatory sequences, moreover as a cluster, or multiple clusters representing targets for the transcriptional initiation complexes [10, 11, 18]. This led to the exploitation of motif over-representation in the target regions. In addition, it has been observed in yeast that the promoter regions are often characterized by multiple occurrences of the same binding motif [23], and it has been postulated that it may also be the case in higher eukaryotes. Along with the expectation of a co-occurrence of motifs in differ-

*Department of Computer Science and Engineering, The University of Texas at Arlington, Arlington, TX 76019, USA; e-mail: {singh,nick}@cse.uta.edu

ent regulatory sequences, this postulate stimulated the search for over-represented, or “surprise”, motifs [2, 24].

As the search for non-coding functional elements in newly sequenced genomes intensified, a comprehensive study of the effectiveness of many motif-finding tools has been performed [22]. Not surprisingly, it has shown that, while there has been some success in the binding site recognition, the existing methods are not nearly satisfactory. There are several reasons for that. Spatial configuration of DNA, along with other epigenetic phenomena, may be a major factor in transcription factor binding, and no currently available tool incorporates this information. Transcription factors generally feature very non-specific binding preferences [3], and that permits for wide variations in the motif consensus. To make things worse, transcription factor binding sites are short, and their detection may be hampered by pieces of repeated or randomly conserved short sequences. Our own previous study has shown that it is indeed the case [20]. This, in turn, may force us to comparatively look at homologous or paralogous sequences that have diverged so much that the proteins that bind there are only similar, but no longer same. The solution to this problem may lie in the simultaneous study of a large number of closely related sequences, which are becoming increasingly available. This was one of the major methods of the ENCODE project [21], which in its pilot phase aimed at the development of high-throughput techniques for the classification of DNA elements within a set of target regions comprising approximately 1% of the human genome.

It is well known that even non-functional parts of a genome are not a random assembly of 4 letters. The analysis of statistical features of sequences often involves a non-trivial model, such as these based on Markov Models, or Hidden Markov Models. In this study we have done extensive simulations and analysis of real data in order to identify the short motif conservation patterns in the human genome, and in particular in the ENCODE target regions. While the number of repeated elements in randomly generated synthetic sequences was almost perfectly conforming to the Poisson expectation, the number of repeated substrings in repeat-masked random inter-

genic sequences was far greater than expected. This bias appears to be genome-wide, as it persisted even when we simultaneously considered many additional randomly collected human sequences, varying in size between 100 and 4,000 characters. In consequence, any search for conserved motifs is bound to return many results, and, depending on what we search for, most would likely be false positives.

In order to gain a better perspective on our ability to characterize significant over-represented motifs in different regions of the genome, we have looked at the number of short (less than 20 bp) repeats, both exact and approximate in various genomic environments and synthetic sequences. Although studies have been done regarding the distribution of tandem repeats [5], we are unaware of any systematic examination of the genome-wide occurrences of very short interspersed motifs.

2 Distribution of short exact repeated motifs

We started the analysis by creating 6 different datasets, each consisting of 100 sequences of 500 bases in length. Although we looked at other possible segment lengths, as short as 50, and as long as several thousand, the results on short exact sequences were not substantially different and length 500 was well suited for the consideration of regions immediately 5' to known genes. Although the issue is still unresolved, some studies have shown that most *cis*-acting regulatory elements appear to cluster in the gene upstream regions of about this length [14].

Four of our data sets were synthetic, containing sequences created by assembling of A's, C's, G's and T's using a random number generator on Unix, and sequences generated by second, third and fifth order Markov Models, trained on one million bases taken from human chromosome 2 obtained through the Ensembl [6] genome browser. The specific MMs have been selected because the second and the third order are widely used in the simulation of genetic sequences, and the fifth order is popular in gene-finding tools. We were especially interested in the behavior

Table 1: The mean numbers (μ) and standard deviations (σ) of repeated patterns of different lengths in different types of nucleotide sequences. Pattern counting has been done over 100 sequences of length 500 in each category.

Pattern Length	Expected Number	Random Synthetic	2 nd Order Markov M.	3 rd Order Markov M.	5 th Order Markov M.	Random Genomic	Upstream Regulatory
4	429.06	$\mu = 425.74$ $\sigma = 6.36$	$\mu = 437.99$ $\sigma = 8.12$	$\mu = 432.84$ $\sigma = 7.1$	$\mu = 432.23$ $\sigma = 6.91$	$\mu = 438.97$ $\sigma = 8.5$	$\mu = 433.92$ $\sigma = 9.94$
5	193.16	$\mu = 189.18$ $\sigma = 15.59$	$\mu = 237.83$ $\sigma = 17.0$	$\mu = 222.98$ $\sigma = 16.68$	$\mu = 222.27$ $\sigma = 15.83$	$\mu = 261.64$ $\sigma = 33.49$	$\mu = 260.11$ $\sigma = 30.67$
6	57.46	$\mu = 55.16$ $\sigma = 12.51$	$\mu = 84.33$ $\sigma = 15.16$	$\mu = 74.58$ $\sigma = 13.27$	$\mu = 75.88$ $\sigma = 14.66$	$\mu = 106.62$ $\sigma = 43.5$	$\mu = 115.31$ $\sigma = 37.72$
7	15.03	$\mu = 14.0$ $\sigma = 5.77$	$\mu = 24.5$ $\sigma = 9.97$	$\mu = 21.82$ $\sigma = 7.81$	$\mu = 23.3$ $\sigma = 9.48$	$\mu = 38.66$ $\sigma = 44.31$	$\mu = 47.54$ $\sigma = 29.88$
8	3.8	$\mu = 3.12$ $\sigma = 2.75$	$\mu = 7.05$ $\sigma = 5.15$	$\mu = 5.75$ $\sigma = 4.16$	$\mu = 6.87$ $\sigma = 5.19$	$\mu = 15.72$ $\sigma = 44.26$	$\mu = 21.3$ $\sigma = 21.62$
9	0.95	$\mu = 0.56$ $\sigma = 1.17$	$\mu = 1.94$ $\sigma = 2.42$	$\mu = 1.47$ $\sigma = 1.92$	$\mu = 1.97$ $\sigma = 2.25$	$\mu = 8.57$ $\sigma = 44.04$	$\mu = 11.33$ $\sigma = 15.67$

of the second order Markov Model, as it has been used to generate control sequences in a comprehensive evaluation of motif-finding tools [22]. Strings generated by even higher order MMs were considered in order to confirm trends, but not studied in detail. The remaining 2 sets were real DNA sequences: one was constructed from the upstream regions immediately 5' to annotated Ensembl human genes and the other consisted of random repeat-masked human intergenic sequences. The total length of sequences in each set was 50,000 bases (300,000 letters total).

In order to count short exact repeated oligonucleotides we used a modification of the Karp-Rabin pattern matching algorithm [13], locating all repeats of specified length in time linear with the size of the sequence. The original Karp-Rabin method was based on numerical keys to code patterns, and we used such keys as indices to a hash table counting the number of occurrences of each motif. We ran our program separately for motif lengths varying between 4 and 9, and recorded the total numbers of repeated elements in Table 1. Since the repeats have been counted separately for each of the 100 sequences in each set, the recorded values include the mean (μ)

and the standard deviation (σ) for all runs. In addition to the empirically determined counts we have also recorded the expected numbers of the repeats, based on the Poisson model.

As it can be seen from Table 1, there were only insignificant differences between the models for motifs of length 4¹. Starting with length 5, an obvious pattern emerges, in which the number of repeats in sequences created by the random number generator correlates with Poisson predictions very well, but none of the other models do. Indeed, a chi-square test on the columns of Table 1 (μ values) confirmed with very high confidence (p -value within the 50% bound) that random synthetic sequence draws from the same distribution as Poisson prediction, but rejected all other models at 0.99 threshold. There were more repeats than expected in all Markov Models and they corresponded well to each other, confirmed by solid p -values. A weak similarity has been found between the second order Markov Model and random intergenic sequences. This, on one hand, justifies its

¹However, when the individual numbers of motifs occurring a particular number of times were taken into account there were considerable differences between the models, as outlined below.

use in modeling genomic environments, but it also advises caution concerning the use of the MMs in simulations.

It was surprising, and somewhat discouraging for the attempts of locating functional elements through over-representation, that random intergenic repeat-masked (and thus, presumably, reasonably unique) sequences featured about the same number of short repeated motifs as sequences taken upstream of the genes. The chi-square test was conclusive on this, with a p -value well within the 50% bound. As for the correspondence between the real sequences and the models, random genomic sequences appear to have somewhat similar number of repeats as the second order Markov Model, but are otherwise quite distinct from any other simulated data set. Overall, real genomic sequences were similar to each other, but not to the models, Markov Models mutually agreed well, but otherwise did not show significant similarity to other data sets, and synthetic sequences corresponded well to the Poisson prediction (a “sanity check”), but their composition was very different than that of Markov Model simulated data, as well as the real sequences.

We next analyzed these relationships at a finer granularity, looking separately at each motif length, and for each length separately at the number of motifs repeating n times, where n was varied between 2 and 10 or more (the latter counted together). Although we did full analysis for all motif lengths in our range, the results were similar, and we show the representative sample for motif lengths 4 and 9 in Table 2. As before, the sequences generated by using random numbers corresponded to Poisson predictions consistently well, while there was a discrepancy between these two and all other models. There was a somewhat weak mutual agreement between different Markov Models (at least two of the three corresponded to each other in every test), and occasionally between random genomic and gene upstream sequences, but the fit between the synthetic and real data was consistently poor.

In this round of testing we have applied the chi-square test on all combinations of models, separately for each motif length, using the sums of the num-

ber of repeats in 100 runs as our samples x_i , where i corresponded to the number of times the motifs have been repeated (so, for instance, in the test for motifs of length 6, x_3 was the count of motifs of length 6 repeated 3 times). This method provided us the sufficient sample size in each category i , which could have otherwise been a problem, having in mind the relative scarcity of long exact motifs repeated many times. Unfortunately, for all comparisons except these involving Poisson expectations we needed to estimate the expected values from the data, and thus substantially reduce the number of degrees of freedom, which has made our analysis of long repeats somewhat unreliable.

Interestingly, while there was a good agreement in the repeat distribution in random genomic and gene upstream sequences for motifs of length 4, the chi-square test failed for every other length. As it can be seen from Table 2 for length 9, gene upstream sequences appear to feature a preference to an increased number of moderately repeated motifs, while random genomic sequences are biased towards these repeated more dramatically (5 or more times). This pattern was consistent for all considered motif lengths, however the fewer motifs of higher repeat count compensated for the lower number of moderately repeated motifs, resulting in an overall similarity in the overall number of repeated sequences throughout our test sets, regardless of their proximity to the genes. Under any circumstances, the number of short repeated motifs in the genomic sequences was greater than in any of the synthetic models, and far greater (an order of magnitude for longer motifs) than the Poisson expectations.

3 Analysis of most common short degenerate motifs

After experiencing this dramatic micro-repetitive structure of the human genome we were interested to find the most common short motifs significantly repeated in the ENCODE regions. Although the program used above was capable of locating short exact motifs in sequences of any length, in linear time, con-

Table 2: The mean numbers (μ) and standard deviations (σ) of repeated patterns of length 4 and 9 in different types of nucleotide sequences. For each motif length, the corresponding rows represent the numbers of motifs repeated n times, where n varies between 2 and 10 or more. Pattern counting has been done over 50 sequences of length 500 in each category.

Length/ Repeats	Expected Number	Random Synthetic	2 nd Order Markov M.	3 rd Order Markov M.	5 th Order Markov M.	Random Genomic	Upstream Regulatory
4/2	69.25	$\mu = 70.28$ $\sigma = 6.2$	$\mu = 50.05$ $\sigma = 6.97$	$\mu = 55.42$ $\sigma = 6.77$	$\mu = 55.16$ $\sigma = 7.94$	$\mu = 45.17$ $\sigma = 8.38$	$\mu = 47.19$ $\sigma = 9.42$
4/3	45.09	$\mu = 44.86$ $\sigma = 5.84$	$\mu = 37.64$ $\sigma = 5.53$	$\mu = 39.04$ $\sigma = 5.32$	$\mu = 38.93$ $\sigma = 4.99$	$\mu = 31.68$ $\sigma = 8.1$	$\mu = 31.05$ $\sigma = 7.69$
4/4	22.02	$\mu = 21.36$ $\sigma = 3.9$	$\mu = 24.39$ $\sigma = 4.34$	$\mu = 22.7$ $\sigma = 4.7$	$\mu = 23.21$ $\sigma = 4.37$	$\mu = 20.13$ $\sigma = 4.73$	$\mu = 19.0$ $\sigma = 5.27$
4/5	8.6	$\mu = 8.17$ $\sigma = 2.8$	$\mu = 12.8$ $\sigma = 3.18$	$\mu = 11.69$ $\sigma = 3.14$	$\mu = 11.12$ $\sigma = 3.08$	$\mu = 11.65$ $\sigma = 3.12$	$\mu = 11.0$ $\sigma = 3.46$
4/6	2.8	$\mu = 2.83$ $\sigma = 1.73$	$\mu = 5.44$ $\sigma = 2.34$	$\mu = 4.52$ $\sigma = 1.83$	$\mu = 4.56$ $\sigma = 1.86$	$\mu = 6.41$ $\sigma = 2.38$	$\mu = 5.81$ $\sigma = 2.28$
4/7	0.78	$\mu = 0.73$ $\sigma = 0.8$	$\mu = 2.51$ $\sigma = 1.47$	$\mu = 2.16$ $\sigma = 1.38$	$\mu = 2.09$ $\sigma = 1.39$	$\mu = 3.67$ $\sigma = 2.02$	$\mu = 3.17$ $\sigma = 1.9$
4/8	0.19	$\mu = 0.23$ $\sigma = 0.47$	$\mu = 0.94$ $\sigma = 1.01$	$\mu = 0.69$ $\sigma = 0.81$	$\mu = 0.77$ $\sigma = 1.0$	$\mu = 1.8$ $\sigma = 1.39$	$\mu = 1.79$ $\sigma = 1.37$
4/9	0.04	$\mu = 0.03$ $\sigma = 0.17$	$\mu = 0.38$ $\sigma = 0.64$	$\mu = 0.33$ $\sigma = 0.57$	$\mu = 0.41$ $\sigma = 0.62$	$\mu = 1.32$ $\sigma = 1.43$	$\mu = 1.13$ $\sigma = 1.35$
4/10+	0.01	$\mu = 0$ $\sigma = 0$	$\mu = 0.16$ $\sigma = 0.39$	$\mu = 0.12$ $\sigma = 0.32$	$\mu = 0.23$ $\sigma = 0.51$	$\mu = 0.79$ $\sigma = 1.56$	$\mu = 0.74$ $\sigma = 1.09$
9/2	0.48	$\mu = 0.28$ $\sigma = 0.58$	$\mu = 0.97$ $\sigma = 1.21$	$\mu = 0.72$ $\sigma = 0.96$	$\mu = 0.92$ $\sigma = 1.07$	$\mu = 2.07$ $\sigma = 3.18$	$\mu = 3.81$ $\sigma = 5.72$
9/3	0	$\mu = 0$ $\sigma = 0$	$\mu = 0$ $\sigma = 0$	$\mu = 0.01$ $\sigma = 0.01$	$\mu = 0.03$ $\sigma = 0.17$	$\mu = 0.21$ $\sigma = 1.6$	$\mu = 0.33$ $\sigma = 0.9$
9/4	0	$\mu = 0$ $\sigma = 0$	$\mu = 0$ $\sigma = 0$	$\mu = 0$ $\sigma = 0$	$\mu = 0.01$ $\sigma = 0.01$	$\mu = 0.05$ $\sigma = 0.26$	$\mu = 0.15$ $\sigma = 0.57$
9/5	0	$\mu = 0$ $\sigma = 0$	$\mu = 0$ $\sigma = 0$	$\mu = 0$ $\sigma = 0$	$\mu = 0$ $\sigma = 0$	$\mu = 0.13$ $\sigma = 1.01$	$\mu = 0.07$ $\sigma = 0.35$
9/6	0	$\mu = 0$ $\sigma = 0$	$\mu = 0$ $\sigma = 0$	$\mu = 0$ $\sigma = 0$	$\mu = 0$ $\sigma = 0$	$\mu = 0.13$ $\sigma = 1.29$	$\mu = 0.06$ $\sigma = 0.28$
9/7	0	$\mu = 0$ $\sigma = 0$	$\mu = 0$ $\sigma = 0$	$\mu = 0$ $\sigma = 0$	$\mu = 0$ $\sigma = 0$	$\mu = 0.09$ $\sigma = 0.8$	$\mu = 0.07$ $\sigma = 0.29$
9/8	0	$\mu = 0$ $\sigma = 0$	$\mu = 0$ $\sigma = 0$	$\mu = 0$ $\sigma = 0$	$\mu = 0$ $\sigma = 0$	$\mu = 0.03$ $\sigma = 0.3$	$\mu = 0.01$ $\sigma = 0.01$
9/9	0	$\mu = 0$ $\sigma = 0$	$\mu = 0$ $\sigma = 0$	$\mu = 0$ $\sigma = 0$	$\mu = 0$ $\sigma = 0$	$\mu = 0.1$ $\sigma = 0.1$	$\mu = 0.01$ $\sigma = 0.1$
9/10+	0	$\mu = 0$ $\sigma = 0$	$\mu = 0$ $\sigma = 0$	$\mu = 0$ $\sigma = 0$	$\mu = 0$ $\sigma = 0$	$\mu = 0.04$ $\sigma = 0.4$	$\mu = 0.01$ $\sigma = 0.01$

Table 3: Longest repeated consensus motifs in the ENCODE regions, after filtering known repeats, simple sequence and tandem repeats. Uppercase letters indicate positions which were perfectly conserved, lowercase letters represent positions where there was a clear majority character, and N’s indicate non-conserved positions. Only 100% TRANSFAC hits are listed. If there was such hit within human factors, others, if any, were omitted. In the absence of a human hit, other factors are indicated by (*) following the factor name.

Motif consensus	Effective length	Expected count in ENCODE	Occurrences in ENCODE	Matches in RepBase	Matches in TRANSFAC
TTTaNaAAAGAAA	11	9.7	135	Multiple (5)	NF-AT1
ATGTNtNTTAAA	9.5	77.5	327	Multiple (5)	MIG (*)
CTGTTTNaNTTT	9.5	77.5	281	Multiple (5)	HNF-3 α , HNF-3B
AAAATgNcTTTT	10	38.8	125	Multiple (6)	YY1

centrating on perfect conservation appeared to be too restrictive. We have thus used our tool for finding short variable repeated motifs, described in detail in [19].

Briefly, our software for finding significant short variable motifs locates all exact repeated patterns in given sequences, starting with dinucleotides. This seeding step is done in time slightly super-linear to the length of the sequences, using the suffix tree data structure [25], which has recently been applied to problems similar to ours [1, 4]. After building the original list of repeats, we use an indexing scheme to quickly locate all neighbors of a given (seed) motif, and search for all pairs that appear to be substantially repeated together, at a fixed distance. Whenever such pairs are found we build the tentative consensus of the approximate motif, and recursively try to extend it with additional seed elements and overlaps. The consensus building continues until a certain quality threshold, usually set to 0.9 (90%) or 0.95, cannot be maintained any longer. We label all positions of absolute conservation with uppercase letters, and assign them the weight based on the number of sites participating in the construction of the consensus. Positions featuring a majority character, but occasionally broken with a mismatch are labeled with a lowercase character, whose weight is determined based on the number of sites which agree. When there is no agreement at a position it is signified by character ‘N’. The final consensus motif is reported

based on the probabilistic evaluation of its length, weight and the number of occurrences.

Since the program assumes that it has been given a set of sequences, rather than a single one, it considerably reduces the search space by filtering out the motifs which do not appear in the minimal number of distinct segments (a settable parameter). Unfortunately, when the sequences in the input are very similar (such as highly conserved sequences from *Hox* gene clusters, or homologous sequences from closely related species), this causes a theoretically exponential explosion of the recursive refinement step. However, such situations are rare (after repeat masking, most sequences do not exhibit good conservation of motifs longer than about a dozen bases), and easily detectable. On average, our software is capable of locating all significantly repeated variable motifs quickly and accurately.

We ran this motif-detection program on the entire set of ENCODE regions, obtained through Ensembl, after masking the repeats. The repeat masking step was done since there was little purpose in trying to find common short repeated motifs in the presence of known long repeat elements. We performed 150,000 runs on 5 to 10 randomly chosen segments of length 1000, setting the program parameters so that only elements which have been found in all segments were reported. We have not excluded known exons from our test data — it simplified the selection and, since exons generally comprise less than 2% of the human

Table 4: Five most significant repeated consensus motifs in the ENCODE regions, after filtering known repeats, simple sequence, tandem repeats and GC-poor repeats. Uppercase letters indicate positions which were perfectly conserved, lowercase letters represent positions where there was a clear majority character, and N’s indicate non-conserved positions. Only 100% TRANSFAC hits are listed. If there was such hit within human factors, others, if any, were omitted. In the absence of a human hit, other factors are indicated by (*) following the factor name.

Motif consensus	Effective length	Expected count in ENCODE	Occurrences in ENCODE	Matches in RepBase	Matches in TRANSFAC
CCCAG _g NNCTG	7.5	310.1	2692	Multiple (47)	SV40 – unknown (*)
GGGNNcTGGG	7.5	310.1	2621	Multiple (37)	SV40 – unknown (*)
AGANNcAGAA	7.5	310.1	2586	Multiple (81)	—
CTGNNtTCCT	7.5	310.1	2251	Multiple (56)	E74A (*)
AGGNNtGGGG	7.5	310.1	2240	Multiple (42)	—

genome we believed that they would not significantly affect our results.

Although we recorded only motifs of length 7 and above, their number was in thousands even after filtering these which were nearly identical or inverse complements of each other, and these featuring extremely simple sequence (all A’s, for instance) or tandem repeats. All these motifs do deserve further classification, but at this time we have limited our study only to about two dozen which were statistically least likely to occur by chance. Since our repeat masked ENCODE sequences contained 40,645,510 bases, counting both strands, in a completely random string of this length a motif of size 10, for instance, would be expected to be found about 39 times. We used such considerations as a basis for the selection of the top choices, where the effective length of the string was calculated by assigning different weights to differently conserved positions (1 for an uppercase letter, 0.5 for lowercase and 0 for an ‘N’ — this could have been further refined by using the exact weights of the identified motifs, but for this study that was not necessary).

In order to do a tentative characterization of the discovered motifs, we have checked them against the human entries in RepBase [12] for possible membership in a known repeat family, and TRANSFAC [15] for a possible functional role. Table 3 summarizes this

information for 4 longest motifs on our list (interestingly, the fifth one of that size was (CTG)₄, which we filtered out), and Table 4 provides the same account for the top 5 motifs after these with two or less G’s or C’s (plus these already included in Table 3) were removed. The remaining top motifs were either degenerated poly-A’s (or poly-T’s), or a combination of A’s and T’s. Although they are also potentially significant, we have not studied them in detail, since poly-A tails are known to be present in many copies in genomic sequences.

4 Discussion

The micro-repetitive structure of the human genome we have described advises us caution when using over-representation for the determination of functional DNA elements. At present, most motif-finding tools do not solely rely on a single motif frequency, taking into account other features of biological relevance, such as clustering of the motifs, matching against experimentally confirmed consensus patterns or evolutionary conservation. While each of these approaches has merits, they all have weaknesses. As we have shown, clustering of over-represented motifs may very well be a consequence of their common source in an ancient repeat (i.e. transposed sequence) not recognized by the RepeatMasker or other repeat-

finding tools, such as RepeatScout [16]. Almost all most frequent motifs we looked at in ENCODE had significant hits in the RepBase, and often in TRANSFAC, too. On the other hand, methods based on phylogenetic footprints may be overly dependent on positional conservation. The ultimate classification of any DNA segment must be done in the laboratory, and the construction of a detailed map of the repetitive landscape of the genome can be of great help in this process.

The fact that for exact motifs of length 5 or more there was no good chi-square agreement between random genomic and gene upstream sequences is potentially significant. The noticed bias towards more copies of single motifs in intergenic regions may indicate the same origin of the sequences, presumably by many overlapping insertions of DNA mobile elements, but different selection pressures after some of the regions acquired a functional role. Under this scenario, further insertions, which would lead to even more random motif copies in non-functional segments would be harmful in regulatory regions, and thus selected against. TRANSFAC hits in organisms other than human would also point in this direction, as parts of ancient repeats shared among the species may have acquired function in some, but not all of them.

Even if database lookups for functional patterns often result in a large number of false positives, the fact that not every one of our top motifs had a match indicates that a selection of sensible candidates for further study is possible. In another project (manuscript in preparation) we have collected 7 putative direct targets for the MLL (Mixed Lineage Leukemia) transcriptional complexes [8], and identified 11 significant common motifs which are currently being investigated in the laboratory. While we expect that some of these, and possibly most, would be noise, we are confident that some would indeed carry a real signal.

Our study can be made more complete in many ways. We need to map the discovered motifs back to their original genomic locations, looking at their groupings and experimental evidence. If most of the short frequent motifs indeed originate in layered ancient transpositional activity, some of them will show significant patterns of co-occurrence. On the other

hand, we may have been too far reaching when taking on the analysis of the human genome first. It would be worthwhile to also look at a simpler genome, like that of a worm, and we shall direct some of our efforts in that direction in the future.

5 Acknowledgments

The authors would like to thank Dr. Cedric Feschotte of the UTA Department of Biology for valuable comments on our results and the general nature of repeated elements in DNA, and Dr. Subhrangsu Mandal of the UTA Department of Chemistry and Biochemistry for help in classifying potential protein binding motifs. This work has been partially supported by UTA REP grant 14 7487 62 to NS.

References

- [1] Adebisi, E.F., Jiang, T. and Kaufmann, M., An efficient algorithm for finding short approximate non-tandem repeats. *Bioinformatics*, 17:S5-S12, 2001.
- [2] Apostolico, A., Bock, M.E., Lonardi, S. and Xu, X., Efficient detection of unusual words. *J. Comput. Biol.*, 7:71-94, 2000.
- [3] Balhoff, J.P. and Wray, G.A., Evolutionary analysis of the well characterized *endo16* promoter reveals substantial variation within functional sites. *PNAS*, 102:8591-8596, 2005.
- [4] Bannai, H., Inenaga, S., Shinohara, A., Takeda, M. and Miyano, S., Efficiently finding regulatory elements using correlation with gene expression. *J. Bioinform. Comput. Biol.*, 2:273-288, 2004.
- [5] Bilgen, M., Karaca, M., Onus, A.N. and Ince, A.G., A software program combining sequence motif searches with keywords for finding repeats containing DNA sequences, *Bioinformatics*, 20(18):3379-3386, 2004.
- [6] Birney, E., Andrews, D., Caccamo, M. *et al.* Ensembl 2006. *Nucleic Acids Res.*, 34:D561-D453, 2006.

- [7] Corcoran, D.L., Feingold, E., Dominick, J., Wright, M., Harnaha, J., Trucco, M., Gianoukakis, N. and Benos, P.V., Footer: A quantitative comparative genomics method for efficient recognition of *cis*-regulatory elements. *Genome Res.*, 15:840–847, 2005.
- [8] Hess, J.L., MLL: a histone methyltransferase disrupted in leukemia. *Trends Mol. Med.* 10:500–507, 2004.
- [9] Hughes, J.D., Estep, P.W., Tavazoie, S. and Church, G.M., Computational identification of *cis*-regulatory elements associated with groups of functionally related genes in *Saccharomyces cerevisiae*. *J. Mol. Biol.*, 296:1205–1214, 2000.
- [10] Jegga, A.G., Sherwood, S.P., Carman, J.W., Pinski, A.T., Phillips, J.L., Pestian, J.P. and Aronow, B.J., Detection and visualization of compositionally similar *cis*-regulatory element clusters in orthologous and coordinately controlled genes. *Genome Res.*, 12:1408–1417, 2002.
- [11] Johansson, O., Alkema, W., Wasserman, W.W. and Lagergren, J., Identification of functional clusters of transcription factor binding motifs in genome sequences: the MSCAN algorithm. *Proceedings of the 11th International Conference on Intelligent Systems in Molecular Biology*, 169–176, 2003.
- [12] Jurka, J., Repbase Update: a database and an electronic journal of repetitive elements. *Trends Genet.*, 9:418–420.
- [13] Karp, R. and Rabin, M., Efficient randomized pattern matching algorithms. *IBM J. Res. Development*, 31:249–260, 1987.
- [14] Khambata-Ford, S., Liu, Y., Gleason, C., Dickson, M., Altman, R.B., Batzoglou, S. and Myers, R., Identification of promoter regions in the human genome by using a retroviral plasmid library-based functional reporter gene assay. *Genome Res.*, 13:1765–1774, 2003.
- [15] Matys, V., Kel-Margoulis, O.V., Fricke, E. *et al.* TRANSFAC® and its module TRANSCompel®: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res.*, 34:D108–D110, 2006.
- [16] Price, A.L., Jones, N.C. and Pevzner, P.A., *De novo* identification of repeat families in large genomes. *Proceedings of the 13th International Conference on Intelligent Systems in Molecular Biology*, 351–358, 2005.
- [17] Sandelin, A., Alkema, W., Engstrom, P., Wasserman, W. and Lenhard, B., JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res.*, 32:D91–D94, 2004.
- [18] Sharan, R., Ovcharenko, I., Ben-Hur, A. and Karp, R.M., CREME: a framework for identifying *cis*-regulatory modules in human-mouse conserved segments. *Proceedings of the 11th International Conference on Intelligent Systems in Molecular Biology*, 283–291, 2003.
- [19] Singh, A. and Stojanovic, N., An efficient algorithm for the identification of repetitive variable motifs in the regulatory sequences of co-expressed genes. To appear in *Proceedings of the 21st International Symposium on Computer and Information Sciences, ISCIS'06.*, Springer-Verlag LNCS.
- [20] Stojanovic, N., Florea, L., Riemer, C., Gumucio, D., Slightom, J., Goodman, M., Miller, W. and Hardison, R., Comparison of five methods for finding conserved sequences in multiple alignments of gene regulatory regions. *Nucleic Acids Res.*, 27:3899–3910, 1999.
- [21] The ENCODE Project Consortium. The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science*, 306:636–640, 2004.
- [22] Tompa, M., Li, N., Bailey, T.L. *et al.* Assessing computational tools for the discovery of transcription factor binding sites. *Nature Biotechnology*, 23:137–144, 2005.
- [23] van Helden, J., Andre, B. and Collado-Vides, J., Extracting regulatory sites from the upstream

region of yeast genes by computational analysis of oligonucleotide frequencies. *J. Mol. Biol.*, 281:827–842, 1998.

[24] van Helden, J., Metrics for comparing regulatory sequences on the basis of pattern counts. *Bioinformatics*, 20:399-406, 2004.

[25] Weiner, P., Linear pattern matching algorithms. *Proceedings of the 14th IEEE Symposium on Switching and Automata Theory*, 1–11, 1973.