A study of the repetitive structure and distribution of short motifs in human genomic sequences

Abanish Singh

Department of Computer Science and Engineering,
University of Texas at Arlington,
TX 76019, Arlington, USA
E-mail: singh@cse.uta.edu

Cedric Feschotte*

Department of Biology,
University of Texas at Arlington,
TX 76019, Arlington, USA
E-mail: cedric@uta.edu
*Corresponding author

Nikola Stojanovic

Department of Computer Science and Engineering,
University of Texas at Arlington,
TX 76019, Arlington, USA
E-mail: nick@cse.uta.edu

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 sequences feature 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.

Author: Please reduce abstract to not more than 100 words.

Keywords: DNA; repeated sequences; functional elements; sequence motifs.

Reference to this paper should be made as follows: Singh, A., Feschotte, C. and Stojanovic, N. (xxxx) 'A study of the repetitive structure and distribution of short motifs in human genomic sequences', *Int. J. Bioinformatics Research and Applications*, Vol. x, No. x, pp.xxx–xxx.

Biographical notes: Abanish Singh received his ME Degree in Computer Science and Engineering in 2000 from the Motilal Nehru National Institute of Technology, Allahabad, India. Before joining the Doctoral Program at the University of Texas at Arlington in 2004, he served on the faculty of the Department of Computer Science and Engineering at the Sant Longowal

Institute of Engineering and Technology, Longowal, India, since 1993. His current PhD Thesis work is in bioinformatics and computational biology, specifically in the genomic sequence analysis and motif discovery.

Cedric Feschotte received his PhD Degree in Biology in 2001 from the University of Paris VI, France. He was a postdoctoral research associate at the Departments of Plant Biology and Genetics at the University of Georgia from 2001 to 2004, where he worked with Dr. Susan Wessler on transposable elements in plant genomes. Since 2004, he is an Assistant Professor in the Department of Biology at the University of Texas at Arlington. His current research focus is on the evolutionary history and genomic impact of mobile genetic elements, with an emphasis on the human genome.

Nikola Stojanovic received his PhD Degree in Computer Science and Engineering in 1997 from the Pennsylvania State University, University Park, PA. After five years of working on the Human Genome Project at the Whitehead Institute/MIT Center for Genome Research in Cambridge, MA, he joined the faculty of the University of Texas at Arlington in 2003, as an Assistant Professor. His research interests are in algorithms for genomic sequence analysis, phylogenetic studies and sequence alignments.

1 Introduction

The search for transcription factor binding sites is one of the most popular sub-fields 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 recognised. 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 (Stojanovic et al., 1999; Jegga et al., 2002; Sharan et al., 2003; Corcoran et al., 2005) or a search for clusters of elements whose sequences match experimentally confirmed consensus motifs (Jegga et al., 2002; Sharan et al., 2003) retrieved from databases such as TRANSFAC (Matys et al., 2006) or Jaspar (Sandelin et al., 2004). 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 (Hughes et al., 2000; Jegga et al., 2002; Bannai et al., 2004). 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 (Jegga et al., 2002; Johansson et al., 2003; Sharan et al., 2003). 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 characterised by multiple occurrences of the same binding motif (van Helden et al.,

1998), 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 different regulatory sequences, this postulate stimulated the search for over-represented, or 'surprise', motifs (Apostolico et al., 2000; van Helden, 2004).

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 (Tompa et al., 2005). 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. Transcrip-tion factors generally feature non-specific binding preferences (Balhoff and Wray, 2005), and that permits 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 (Stojanovic et al., 1999). This, in turn, may force us to comparatively look at homologous or paralogous sequences which 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 (The ENCODE Project Consortium, 2004), 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 four letters. The analysis of statistical features of sequences often involves a non-trivial background 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 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 intergenic 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 characterise significant over-represented motifs in different regions of the genome, we have looked at the number of short (<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 (Bilgen et al., 2004) and larger interspersed repeats (International Human Genome Sequencing Consortium, 2001), 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 six different data sets, 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 (Khambata-Ford et al., 2003).

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 (Birney et al., 2006) genome browser. These 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 behaviour of the second order Markov Model, as it has been used to generate control sequences in a comprehensive evaluation of motif-finding tools (Tompa et al., 2005). Strings generated by even higher order MMs were considered in order to confirm trends, but not studied in detail. The remaining two 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 (Karp and Rabin, 1987), 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 4.[1] Starting with length five, 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), whose results are shown in Table 2, confirmed with very high confidence that random synthetic sequence draws from the same distribution as Poisson prediction, but rejected other sets (except, weakly, the higher order MMs). 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 also been found between the second order Markov Model and random intergenic sequences. This, on one hand, justifies its use in modeling genomic environments, but it also advises caution concerning the use of the MMs in simulations.

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

Table 2 Chi-square confidence levels for the compared data sets, indicating the likelihood that sequences in the compared set pairs (column-wise in Table 1) have indeed been drawn from the same distribution. MMn abbreviates *nth* order Markov model

	Expected number	*Random synthetic*	*2nd order Markov M.*	*3rd order Markov M.*	*5th order Markov M.*	*Random genomic*	*Upstream regulatory*
Expected	1.0	>0.995	<0.005	≈0.02	<0.005	<<0.005	<<0.005
Random	>0.995	1.0	<0.01	≈0.2	≈0.1	<<0.005	<<0.005
MM2	<0.005	<0.01	1.0	≈0.6	≈0.8	≈0.025	<0.005
MM3	≈	≈0.2	≈0.6	1.0	>0.995	<0.005	<0.005
MM5	<	≈0.1	≈0.8	>0.995	1.0	<0.005	<0.005
Genomic	<<	<<0.005	≈0.025	<0.005	<0.005	1.0	≈0.8
Regulatory	<<	<<0.005	<0.005	<0.005	<0.005	≈0.8	1.0

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 the *p*-value indicating a strong agreement. 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 different than that of Markov Model simulated data, and dramatically different than that of the real sequences.

We next analysed 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 two and ten 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, 7 and 9 in Table 3. 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 number 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 three 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 longer repeats somewhat unreliable.

Table 3 The mean numbers (μ) and standard deviations (σ) of repeated patterns of length 4, 7 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 two and ten or more. Pattern counting has been done over 100 sequences of length 500 in each category

Length/ repeats	*Expected number*	*Random synthetic*	*2nd order Markov M.*	*3rd order Markov M.*	*5th order Markov M.*	*Random genomic*	*Upstream regulatory*
4/2	69.25	$\mu = 70.28$	$\mu = 50.05$	$\mu = 55.42$	$\mu = 55.16$	$\mu = 45.17$	$\mu = 47.19$
		$\sigma = 6.2$	$\sigma = 6.97$	$\sigma = 6.77$	$\sigma = 7.94$	$\sigma = 8.38$	$\sigma = 9.42$
4/3	45.09	$\mu = 44.86$	$\mu = 37.64$	$\mu = 39.04$	$\mu = 38.93$	$\mu = 31.68$	$\mu = 31.05$
		$\sigma = 5.84$	$\sigma = 5.53$	$\sigma = 5.32$	$\sigma = 4.99$	$\sigma = 8.1$	$\sigma = 7.69$
4/4	22.02	$\mu = 21.36$	$\mu = 24.39$	$\mu = 22.7$	$\mu = 23.21$	$\mu = 20.13$	$\mu = 19.0$
		$\sigma = 3.9$	$\sigma = 4.34$	$\sigma = 4.7$	$\sigma = 4.37$	$\sigma = 4.73$	$\sigma = 5.27$
4/5	8.6	$\mu = 8.17$	$\mu = 12.8$	$\mu = 11.69$	$\mu = 11.12$	$\mu = 11.65$	$\mu = 11.0$
		$\sigma = 2.8$	$\sigma = 3.18$	$\sigma = 3.14$	$\sigma = 3.08$	$\sigma = 3.12$	$\sigma = 3.46$
4/6	2.8	$\mu = 2.83$	$\mu = 5.44$	$\mu = 4.52$	$\mu = 4.56$	$\mu = 6.41$	$\mu = 5.81$
		$\sigma = 1.73$	$\sigma = 2.34$	$\sigma = 1.83$	$\sigma = 1.86$	$\sigma = 2.38$	$\sigma = 2.28$
4/7	0.78	$\mu = 0.73$	$\mu = 2.51$	$\mu = 2.16$	$\mu = 2.09$	$\mu = 3.67$	$\mu = 3.17$
		$\sigma = 0.8$	$\sigma = 1.47$	$\sigma = 1.38$	$\sigma = 1.39$	$\sigma = 2.02$	$\sigma = 1.9$
4/8	0.19	$\mu = 0.23$	$\mu = 0.94$	$\mu = 0.69$	$\mu = 0.77$	$\mu = 1.8$	$\mu = 1.79$
		$\sigma = 0.47$	$\sigma = 1.01$	$\sigma = 0.81$	$\sigma = 1.0$	$\sigma = 1.39$	$\sigma = 1.37$
4/9	0.04	$\mu = 0.03$	$\mu = 0.38$	$\mu = 0.33$	$\mu = 0.41$	$\mu = 1.32$	$\mu = 1.13$
		$\sigma = 0.17$	$\sigma = 0.64$	$\sigma = 0.57$	$\sigma = 0.62$	$\sigma = 1.43$	$\sigma = 1.35$
4/10+	0.01	$\mu = 0$	$\mu = 0.16$	$\mu = 0.12$	$\mu = 0.23$	$\mu = 0.79$	$\mu = 0.74$
		$\sigma = 0$	$\sigma = 0.39$	$\sigma = 0.32$	$\sigma = 0.51$	$\sigma = 1.56$	$\sigma = 1.09$
7/2	7.4	$\mu = 6.97$	$\mu = 11.8$	$\mu = 10.39$	$\mu = 10.86$	$\mu = 15.85$	$\mu = 18.51$
		$\sigma = 2.87$	$\sigma = 4.85$	$\sigma = 3.64$	$\sigma = 4.12$	$\sigma = 6.58$	$\sigma = 9.25$
7/3	0.07	$\mu = 0.02$	$\mu = 0.3$	$\mu = 0.3$	$\mu = 0.39$	$\mu = 0.75$	$\mu = 1.52$
		$\sigma = 0.14$	$\sigma = 0.61$	$\sigma = 0.59$	$\sigma = 0.72$	$\sigma = 1.62$	$\sigma = 2.7$

Table 3 The mean numbers (μ) and standard deviations (σ) of repeated patterns of length 4, 7 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 two and ten or more. Pattern counting has been done over 100 sequences of length 500 in each category (continued)

Length/ repeats	*Expected number*	*Random synthetic*	*2nd order Markov M.*	*3rd order Markov M.*	*5th order Markov M.*	*Random genomic*	*Upstream regulatory*
7/4	0.001	$\mu = 0$	$\mu = 0$	$\mu = 0.01$	$\mu = 0.05$	$\mu = 0.06$	$\mu = 0.47$
		$\sigma = 0$	$\sigma = 0$	$\sigma = 0.01$	$\sigma = 0.22$	$\sigma = 0.24$	$\sigma = 1.11$
7/5	0	$\mu = 0$	$\mu = 0$	$\mu = 0.02$	$\mu = 0.03$	$\mu = 0.13$	$\mu = 0.18$
		$\sigma = 0$	$\sigma = 0$	$\sigma = 0.14$	$\sigma = 0.17$	$\sigma = 0.91$	$\sigma = 0.52$
7/6	0	$\mu = 0$	$\mu = 0$	$\mu = 0$	$\mu = 0.01$	$\mu = 0.13$	$\mu = 0.13$
		$\sigma = 0$	$\sigma = 0$	$\sigma = 0$	$\sigma = 0.01$	$\sigma = 1.01$	$\sigma = 0.44$
7/7	0	$\mu = 0$	$\mu = 0$	$\mu = 0$	$\mu = 0$	$\mu = 0.06$	$\mu = 0.03$
		$\sigma = 0$	$\sigma = 0$	$\sigma = 0$	$\sigma = 0$	$\sigma = 0.42$	$\sigma = 0.17$
7/8	0	$\mu = 0$	$\mu = 0$	$\mu = 0$	$\mu = 0$	$\mu = 0.07$	$\mu = 0.06$
		$\sigma = 0$	$\sigma = 0$	$\sigma = 0$	$\sigma = 0$	$\sigma = 0.6$	$\sigma = 0.24$
7/9	0	$\mu = 0$	$\mu = 0$	$\mu = 0$	$\mu = 0$	$\mu = 0.14$	$\mu = 0.03$
		$\sigma = 0$	$\sigma = 0$	$\sigma = 0$	$\sigma = 0$	$\sigma = 1.3$	$\sigma = 0.17$
7/10+	0	$\mu = 0$	$\mu = 0$	$\mu = 0$	$\mu = 0$	$\mu = 0.08$	$\mu = 0.03$
		$\sigma = 0$	$\sigma = 0$	$\sigma = 0$	$\sigma = 0$	$\sigma = 0.8$	$\sigma = 0.17$
9/2	0.48	$\mu = 0.28$	$\mu = 0.97$	$\mu = 0.72$	$\mu = 0.92$	$\mu = 2.07$	$\mu = 3.81$
		$\sigma = 0.58$	$\sigma = 1.21$	$\sigma = 0.96$	$\sigma = 1.07$	$\sigma = 3.18$	$\sigma = 5.72$
9/3	0	$\mu = 0$	$\mu = 0$	$\mu = 0.01$	$\mu = 0.03$	$\mu = 0.21$	$\mu = 0.33$
		$\sigma = 0$	$\sigma = 0$	$\sigma = 0.01$	$\sigma = 0.17$	$\sigma = 1.6$	$\sigma = 0.9$
9/4	0	$\mu = 0$	$\mu = 0$	$\mu = 0$	$\mu = 0.01$	$\mu = 0.05$	$\mu = 0.15$
		$\sigma = 0$	$\sigma = 0$	$\sigma = 0$	$\sigma = 0.01$	$\sigma = 0.26$	$\sigma = 0.57$
9/5	0	$\mu = 0$	$\mu = 0$	$\mu = 0$	$\mu = 0$	$\mu = 0.13$	$\mu = 0.07$
		$\sigma = 0$	$\sigma = 0$	$\sigma = 0$	$\sigma = 0$	$\sigma = 1.01$	$\sigma = 0.35$
9/6	0	$\mu = 0$	$\mu = 0$	$\mu = 0$	$\mu = 0$	$\mu = 0.13$	$\mu = 0.06$
		$\sigma = 0$	$\sigma = 0$	$\sigma = 0$	$\sigma = 0$	$\sigma = 1.29$	$\sigma = 0.28$
9/7	0	$\mu = 0$	$\mu = 0$	$\mu = 0$	$\mu = 0$	$\mu = 0.09$	$\mu = 0.07$
		$\sigma = 0$	$\sigma = 0$	$\sigma = 0$	$\sigma = 0$	$\sigma = 0.8$	$\sigma = 0.29$
9/8	0	$\mu = 0$	$\mu = 0$	$\mu = 0$	$\mu = 0$	$\mu = 0.03$	$\mu = 0.01$
		$\sigma = 0$	$\sigma = 0$	$\sigma = 0$	$\sigma = 0$	$\sigma = 0.3$	$\sigma = 0.01$
9/9	0	$\mu = 0$	$\mu = 0$	$\mu = 0$	$\mu = 0$	$\mu = 0.01$	$\mu = 0.01$
		$\sigma = 0$	$\sigma = 0$	$\sigma = 0$	$\sigma = 0$	$\sigma = 0.01$	$\sigma = 0.1$
9/10+	0	$\mu = 0$	$\mu = 0$	$\mu = 0$	$\mu = 0$	$\mu = 0.04$	$\mu = 0.01$
		$\sigma = 0$	$\sigma = 0$	$\sigma = 0$	$\sigma = 0$	$\sigma = 0.4$	$\sigma = 0.01$

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 3 for lengths 7 and 9, gene upstream

sequences appear to feature a preference to an increased number of moderately repeated motifs, while random genomic sequences are biased towards smaller numbers of these repeated more dramatically (five 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, concentrating on perfect conservation appeared to be too restrictive. We have thus used our tool for finding short variable repeated motifs, described in detail in Singh and Stojanovic (2006).

Briefly, our software starts by locating all exact repeated patterns in given sequences, including dinucleotides. This seeding step is done in time slightly super-linear to the length of the sequences, using the suffix tree data structure (Weiner, 1973), which has recently been applied to problems similar to ours (Adebiyi et al., 2001; Bannai et al., 2004). After building the original list of repeats, we use an indexing scheme to quickly locate all neighbours 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, can not 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 labelled 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 vertebrate ultra-conserved sequences, or even just 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 intergenic 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 5–10 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 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 characterisation of the discovered motifs, we have checked them against the human entries in RepBase (Jurka, 2000) for possible membership in a known repeat family, and TRANSFAC (Matys et al., 2006) for a possible functional role. Table 4 summarises this information for four longest motifs in our list (the fifth one of that size was (CTG)4, which we filtered out), and Table 5 provides the same account for the top five motifs after these with two or less *G*'s or *C*'s 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. One explanation for their prevalence is in that they are derived from the terminus of non-LTR retrotransposon repeats. These elements are abundant in the human genome (over 2.3 million copies spanning over one third of the genome) and they are characterised by a stretch of poly-*A*'s at their 3' end (International Human Genome Sequencing Consortium, 2001). Because of its variable length and rapid mutational degradation, part or all of the 3' poly-*A* terminus of non-LTR retrotransposons may often remain after repeat masking.

Even as the number of matches our top motifs had in RepBase generally exceeded what would be expected by chance only, these hits were not concentrated in a single repeat class, and thus probably do not represent remnants of a particular mobile element, at least not one of a known classification. Similarly, their TRANS-FAC matches do not appear to lend strong support to the hypothesis that they may be functional protein binding sites – the examined sequence was human, but most of the hits were in non-human elements, or elements which are common in repeated sequences throughout the mammalian lineage (or even broader). While these motifs are clearly strongly repetitive, and some also likely functional, further studies are needed in order to characterise their nature and origins.

Table 4 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-3α, HNF-3B
AAAATgNcTTTT	10	38.8	125	Multiple (6)	YY1

Table 5 Five most significant repeated consensus motifs in the ENCODE regions, after filtering known repeats, simple sequence, tandem repeats and GC-poor repeats. Upper case 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*
CCCAgNNCTG	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)	–

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. Clustering of over-represented motifs may very well be a consequence of their common source in an ancient repeat (i.e., transposed sequence) not recognised by the RepeatMasker or other repeat-finding tools, such as Re-peatScout (Price et al., 2005). Almost all most frequent motifs we looked at in ENCODE had hits in 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 seven putative direct targets for the Mixed Lineage Leukemia (MLL) transcriptional complexes (Hess, 2004), 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 a few would indeed carry a 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.

Acknowledgements

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

References

Adebiyi, E.F., Jiang, T. and Kaufmann, M. (2001) 'An efficient algorithm for finding short approximate non-tandem repeats', *Bioinformatics*, Vol. 17, pp.S5–S12.

Apostolico, A., Bock, M.E., Lonardi, S. and Xu, X. (2000) 'Efficient detection of unusual words', *J. Comput. Biol.*, Vol. 7, pp.71–94.

Balhoff, J.P. and Wray, G.A. (2005) 'Evolutionary analysis of the well characterized *endo16* promoter reveals substantial variation within functional sites', *PNAS*, Vol. 102, pp.8591–8596.

Bannai, H., Inenaga, S., Shinohara, A., Takeda, M. and Miyano, S. (2004) 'Efficiently finding regulatory elements using correlation with gene expression', *J. Bioinform.Comput. Biol.*, Vol. 2, pp.273–288.

Bilgen, M., Karaca, M., Onus, A.N. and Ince, A.G. (2004) 'A software program combining sequence motif searches with keywords for finding repeats containing DNA sequences', *Bioinformatics*, Vol. 20, pp.3379–3386.

Birney, E., Andrews, D., Caccamo, M. *et al.* (2006) 'Ensembl 2006', *Nucleic Acids Res.*, Vol. 34, pp.D561–D453. AUTHOR PLEASE PROVIDE REMAINING AUTHORS NAME.

Corcoran, D.L., Feingold, E., Dominick, J., Wright, M., Harnaha, J., Trucco, M., Giannoukakis, N. and Benos, P.V. (2005) 'Footer: a quantitative comparative genomics method for efficient recognition of *cis*-regulatory elements', *Genome Res.*, Vol. 15, pp.840–847.

Hess, J.L. (2004) 'MLL: a histone methyltransferase disrupted in leukemia', *Trends Mol. Med.*, Vol. 10, pp.500–507.

Hughes, J.D., Estep, P.W., Tavazoie, S. and Church, G.M. (2000) 'Computational identification of cis-regulatory elements associated with groups of functionally related genes in *Saccharomyces cerevisiae*', *J. Mol. Biol.*, Vol. 296, pp.1205–1214.

International Human Genome Sequencing Consortium (2001) 'Initial sequencing and analysis of the human genome', *Nature*, Vol. 409, pp.860–921.

Jegga, A.G., Sherwood, S.P., Carman, J.W., Pinski, A.T., Phillips, J.L., Pestian, J.P. and Aronow, B.J. (2002) 'Detection and visualization of compositionally similar *cis*-regulatory element clusters in orthologous and coordinately controlled genes', *Genome Res.*, Vol. 12, pp.1408–1417.

Johansson, O., Alkema, W., Wasserman, W.W. and Lagergren, J. (2003) '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*, pp.169–176.

Jurka, J. (2000) 'Repbase Update: a database and an electronic journal of repetitive elements', *Trends Genet.*, Vol. 9, pp.418–420.

Karp, R. and Rabin, M. (1987) 'Efficient randomized pattern matching algorithms', *IBM J. Res. Development*, Vol. 31, pp.249–260.

Khambata-Ford, S., Liu, Y., Gleason, C., Dickson, M., Altman, R.B., Batzoglou, S. and Myers, R. (2003) 'Identification of promoter regions in the human genome by using a retroviral plasmid library-based functional reporter gene assay', *Genome Res.*, Vol. 13, pp.1765–1774.

Matys, V., Kel-Margoulis, O.V., Fricke, E. *et al.* (2006) 'TRANSFAC® and its module TRANS compel®: transcriptional gene regulation in eukaryotes', *Nucleic Acids Res.*, Vol. 34, pp.D108–D110. AUTHOR PLEASE PROVIDE REMAINING AUTHORS NAME.

Price, A.L., Jones, N.C. and Pevzner, P.A. (2005) 'De novo identification of repeat families in large genomes', *Proceedings of the 13th International Conference on Intelligent Systems in Molecular Biology*, pp.351–358.

Sandelin, A., Alkema, W., Engstrom, P., Wasserman, W. and Lenhard, B. (2004) 'JASPAR: an open-access database for eukaryotic transcription factor binding profiles', *Nucleic Acids Res.*, Vol. 32, pp.D91–D94.

Sharan, R., Ovcharenko, I., Ben-Hur, A. and Karp, R.M. (2003) '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*, pp.283–291.

Singh, A. and Stojanovic, N. (2006) 'An efficient algorithm for the identification of repetitive variable motifs in the regulatory sequences of co-expressed genes', *Proceedings of the 21st International Symposium on Computer and Information Sciences*, Springer-Verlag LNCS, Vol. 4263, pp.182–191.

Stojanovic, N., Florea, L., Riemer, C., Gumucio, D., Slightom, J., Goodman, M., Miller, W. and Hardison, R. (1999) 'Comparison of five methods for finding conserved sequences in multiple alignments of gene regulatory regions', *Nucleic Acids Res.*, Vol. 27, pp.3899–3910.

The ENCODE Project Consortium (2004) 'The ENCODE (ENCyclopedia Of DNA elements) Project', *Science*, Vol. 306, pp.636–640.

Tompa, M., Li, N., Bailey, T.L. *et al.* (2005) 'Assessing computational tools for the discovery of transcription factor binding sites', *Nature Biotechnology*, Vol. 23, pp.137–144. AUTHOR PLEASE PROVIDE REMAINING AUTHORS NAME.

van Helden, J., Andre, B. and Collado-Vides, J. (1998) 'Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies', *J. Mol. Biol.*, Vol. 281, pp.827–842.

van Helden, J. (2004) 'Metrics for comparing regulatory sequences on the basis of pattern counts', *Bioinformatics*, Vol. 20, pp.399–406.

Weiner, P. (1973) 'Linear pattern matching algorithms', *Proceedings of the 14th IEEE Symposium on Switching and Automata Theory*, pp.1–11.

Note

[1]However, 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.