RELATIVELY CONSERVED COMMON SHORT SEQUENCES IN TRANSCRIPTION FACTOR BINDING SITES AND miRNA

P. Putta[1, 3], Yu.L. Orlov[2], N.L. Podkolodnyy[2], C.K. Mitra[3]

[1] Medical and Molecular Genetics, School of Clinical and Experimental Medicine, University of Birmingham, Birmingham, UK;
[2] Institute of Cytology and Genetic, SB RAS, Novosibirsk, Russia;
[3] Department of Biochemistry, University of Hyderabad, Hyderabad 500046, India

Transcription factor binding sites (TFBS) are the specific DNA binding motifs that are recognized by a transcription factor, are typically short and are often degenerate. Sequence-specific binding of TFs to the DNA controls the gene expression regulation at transcription level. High throughput genome sequencing provides abundant data for TF binding profiles. We have developed computer program to study distribution of short motifs in the sequences. We are interested in studying the general features of transcription factor binding site sequences which can help in identifying conserved patterns at nucleotide level. We explored the role of common oligonucleotide patterns in TFBS and in miRNA based on the sequence specific similarity between these two sets.

Key words: genome, transcription factor binding sites, miRNA, oligonucleotides, statistics, sequencing.

Introduction

In eukaryotes, the protein coding genes are transcribed by RNA polymerase II. However, unlike in prokaryotes, only a set of genes are targeted prior to transcription. The selection of genes to be transcribed is carried out by various transcription factors (class of DNA binding proteins; TF) that recognize parts of the promoter region. This enables selective transcription of relevant genes but involves higher overheads. Transcription is a more complex process involving chromatin modifiers, transcription factors, co-factors and RNA polymerase (among others) that tightly regulate the basal transcription. Sequence-specific binding of TFs to short stretches of DNA, i.e., transcription factor binding sites (TFBS) within the vicinity of a gene (the promoter) is one of the critical components in the transcriptional regulation and control. Mutations within these TFBS sequences may result in diseases and are likely to change the phenotype variability within and across the species (Wray, 2007). Transcription factors are proteins and they usually signal transcription for a number of related proteins. When a cell needs a certain protein, the corresponding TF is activated, which produces the desired TF which in turn activates the transcription of a set of related proteins. The complex regulation mechanism enables the cell to produce desired proteins on demand. Recent progress of high throughput sequencing technologies allow study these processes in greater details in genome scale, but even today much of these processes have not been clearly understood.

Out of ~20,000 genes estimated for the human genome, ~100 TFs have been clearly identified and their binding sites mapped in genome by ChIP-seq technologies, in particular in frames of ENCODE project. It is possible that we have so far found less than 10 % of the TFs present in the whole genome. Based on experimental data available, it is known that the binding sites of the TFs are degenerate, i.e., they recognize several distinct but related binding sites. This is apparently puzzling as biological recognition process is usually highly specific and accurate.

However, we presume that degeneracy is perhaps helpful when one TF need to recognize several different genes (Bulyk *et al*., 2002; Man, Stormo, 2001). We have attempted to study this behaviour in the present study. We also analyse the miRNA sequences (mature miRNA sequences) for human

and look for common oligonucleotide sequences (of 6-nt length) between the TF binding sites and the miRNA. We have chosen to study the 6-nt sequences based on a few common observations: i) the most common representation of promoter elements (TATAAT and TTGACA in prokaryotes) are 6-nt in length; ii) most of the restriction enzymes recognize DNA sequence of 6-nts length with high accuracy; and finally iii) the minimum length of TFBS in JASPAR are about 6-nt sequences. We further assume that an oligonucleotide of 6-nt length can be recognized by standard protein motifs without errors. Longer the sequences, they exhibit degeneracy or redundancy or may be just error-prone (Yamamoto *et al*., 2007). Based on the presence of the common 6-nt sequences, we can perhaps classify the miRNA and the TFBS and reveal motifs specific only for transcription factors in general. We found common patterns for TFBS and miRNA and compared them to known binding motifs.

Materials and Methods

Human transcription factor binding site sequences were downloaded from the JASPAR database (http://jaspar.cgb.ki.se/). A total of 6496 transcription factor binding site sequences that represent 65 human TFs (the database has reported 75 TFs but 10 of them had no sequence information) were extracted from the database. The 6496 sequences have an average length of 14 nucleotides (mean: 14,2; max: 28; min: 6) and the base composition (A: 22707, 24,54%; C: 20653, 22,32%; G: 21203, 22,91%; T: 21352, 23,08%; N: 6614; Total: 92529) is approximately uniform. Several sequences in the database had also residues that are not part of the binding site (indicated as lower case bases in the database) were ignored (i.e., were not reflected in the above computations). The TFBS sequences were next searched (using a custom-made C program) for all possible 6-nucleotide sequences (4^6 = 4096 possible sequences) and a count of each were maintained. These 4096 possible 6-nt sequences were assigned numerical values (lexically ordered) for computational convenience (by coding in degrees of 4; we set 0 for A, 1 for C, 2 for G and 3 for T or U). For example, AAAAAA = 1, AAAAAC = 2, …, TTTTTT = 4096 (Table 1). The number is coding number in degrees of 4 plus 1. In addition, for each 6-mer we have fixed number of

Table 1

Example of the 6-mer sequence enumeration scheme and frequencies in the datasets studied

6-mer	Numbering of 6-mer	Complementary 6-mer	Numbering of compl. 6-mer	# in TFBS	# in miRNA	Self-compl.?	# in TFBS with complement	# in miRNA with complement	Counted in list of 2016 6-mers
AAAAAA	1	TTTTTT	4096	4	12		5	20	1
AAAAAC	2	GTTTTT	3072	2	13		5	27	1
AAAAAG	3	CTTTTT	2048	10	8		14	17	1
AAAAAT	4	ATTTTT	1024	7	5		22	13	1
AAAACA	5	TGTTTT	3840	26	5		47	22	1
…	…	…	…	…	…	…	…	…	…
AGGAAA	641	TTTCCT	4056	226	14		426	25	1
AGGAAC	642	GTTCCT	3032	54	9		104	23	1
AGGAAG	643	CTTCCT	2008	294	23		554	44	1
…	…	…	…	…	…	…	…	…	…
GAGCTC	2206	GAGCTC	2206	5	15	yes	5	15	0
GAGCTG	2207	CAGCTC	1182	3	20		5	30	0
…	…	…	…	…	…	…	…	…	…
TTTTTG	4095	CAAAAA	1025	2	21		12	35	0
TTTTTT	4096	AAAAAA	1	1	8		5	20	0

complementary 6-mer. For example for AAAAAA it is TTTTTT with numbering 4096, for AAAAAC it is GTTTTT with numbering 3072, and so on (see Table 1 for examples).

Next we counted number of hits of each 6-mers in the TF dataset (Table 1, data not shown in full). Since orientation of 6-mer in regulatory regions is not known we used sum of numbers of 6-mer and its complement. In total we have 2080 non-redundant 6-mers. 64 of them are self-complementary, such as GAGCTC (Table 1).

Human mature miRNA sequences were downloaded from the microRNA database (http://microrna.sanger.ac.uk/) for humans. 1733 sequences were found and there were no unknown (unidentified) bases. This database is considerably smaller and has the following base composition (A: 8568, 22,95%; C: 8448, 22,63%; G: 10505, 28,14%; U: 9811, 26,28%; N = 0; Total = 37332) which is also approximately uniform (see above). The mean sequence length in this database is 21 nucleotides which is longer compared to the TFBS mean sequence length. This database was searched as before for the 4096 possible 6-nt sequences and its complements (Table 1). Last column in Table 1 shows selection of non-redundant set of 2080 sequences.

Then the data were sorted by 6-mers frequencies (highest frequency first). We next combine these two sets of results into a combined one: first set has 6-nt sequences that have high frequencies in both TFBS and miRNA; the second set has 6-nt sequences that have high frequencies in TFBS but not in the miRNA and the third set has 6-nt sequences that have high frequencies in the miRNA but not in the TFBS. The final set that has low frequencies in both sets was ignored. In this work, we consider low frequency to be less than 2σ (two times the standard deviation).

Results

We arranged number of 6-mer hits in the databases and presented it as histogram (Fig. 1).

The analysis of the TFBS resulted in 3668 sequences (out of 4096 possible 6-nt sequences). The highest frequency was 393 (for the sequence numbered 1185 corresponding to the sequence CAGGAA (in list of 4096). The same sequence was most frequent in non-redundant list of 2080 6-mers too. Total number of 6-mers is 53463. The frequency distribution of the sequences is shown in Fig 1 a. We note that many sequences were ignored as they contained the unidentified base (N). We consider all 6-nt frequencies greater than 2σ to be significant in this study (i.e. frequencies greater than 140 in TFBS set).

The analysis of the miRNA database was performed in an identical fashion. We found 3893 sequences (out of 4096; slightly more than the TFBS sequences). The highest frequency was 67 (corresponding to the 6-nt sequence AAGTGC) and the sum of the frequencies is 28667 (this is about half of the value for the TFBS). The frequency distribution sequences are seen in Fig. 1 b. We consider all 6-nt frequencies greater than 2σ to be signifi-

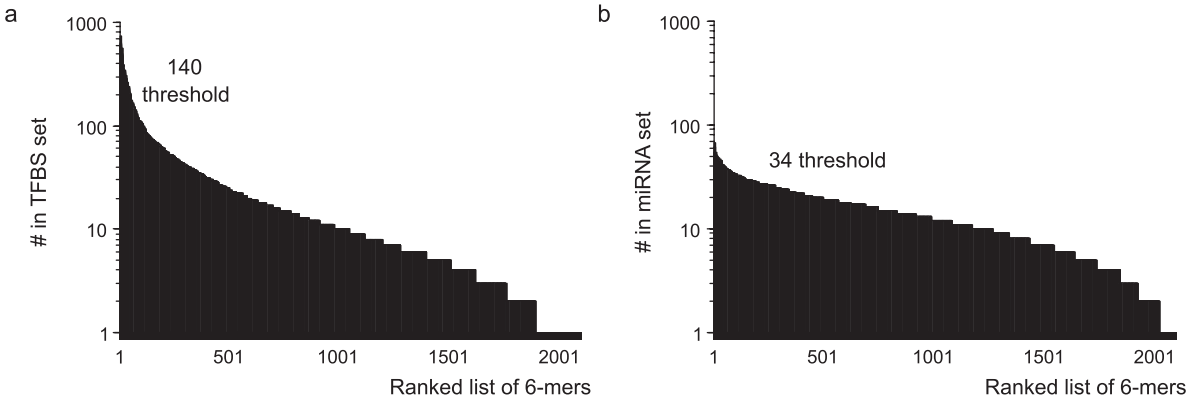

Fig. 1. Histogram for distribution of 6-mers (a) for the TFBS (b) for the miRNA database.

The Y-axis (number of 6-mers found) has been plotted on a log scale for ease of comparison. Also the 6-nt sequence order for the two plots are not necessary same (ranked by the 6-mer occurrence number). The distribution of 6-mers in TFBS is clearly more uniform compared to 6-mer distribution in miRNA (see text for details of the statistical parameters for the two distributions).

cant in this study (i.e frequencies greater than 12). The coefficients of variation (σ/μ) for these two distributions (TFBS and miRNA) are 2,18 and 0,72 respectively. This suggests that the 6-nt sequences are relatively more uniformly distributed for the 6-nt sequences in miRNA. We have also calculated skewness of 6-mer distribution in TFBS and in miRNA sets. Skewness characterizes the degree of asymmetry of a distribution around its mean. Higher value of skewness indicates a distribution with an asymmetric tail extending toward larger values. It is ~6,1 for TFBS and only ~1,2 for miRNA set (Fig. 1).

Base Composition

Base composition of the 6496 TFBS were reported below. It was clear that the four bases appear nearly uniform in distribution in the binding sequences. Following are the base frequencies observed in the binding site sequences. A = 22707 (24,54%), C = 20653 (22,32%), G = 21203 (22,91%) and T = 21352 (23,08%). We therefore consider them to be uniformly distributed at the single nucleotide level. For the miRNA database, there were 1733 sequences and the following base composition was observed: A = 8568 (22,95%); C = 8448 (22,63%); G = 10505 (28,14%); T(U) = 9811 (26,28%). Both these distributions appear to be reasonably uniform. At the same time, distributions of oligonucleotides for TF and miRNA are different.

Joint Frequency distribution of sequences

A casual examination of the two sets of 6-nt sequences revealed some similarities. However, there are also some dissimilarities. We plotted the frequencies for all 6-nt sequences on a graph for both the TFBS and miRNA to observe correlation in terms of frequencies (Fig. 2). Initially we plotted with all 6-nt frequencies in both the sets (Fig. 2). Then we applied the 2σ cutoff to the frequencies. We are interested in the 6-mers with high frequencies in both the sets. 6-mers with high frequencies were labeled. Among them, we can note 3 distinct groups that are common between the TFBS and miRNA datasets.

The purpose of this exercise is to locate 6-nt sequences that are common to both TFBS and miRNA databases. For this exercise, we first removed all 6-nt sequences from the TFBS results that have frequencies less than 60 (this was arbitrarily chosen as the 2σ cutoff value). A large number of 6-nt sequences were therefore removed and we were left with 166 sequences. We also removed from the miRNA results all 6-nt sequences that have frequencies less than 12. As noted earlier,

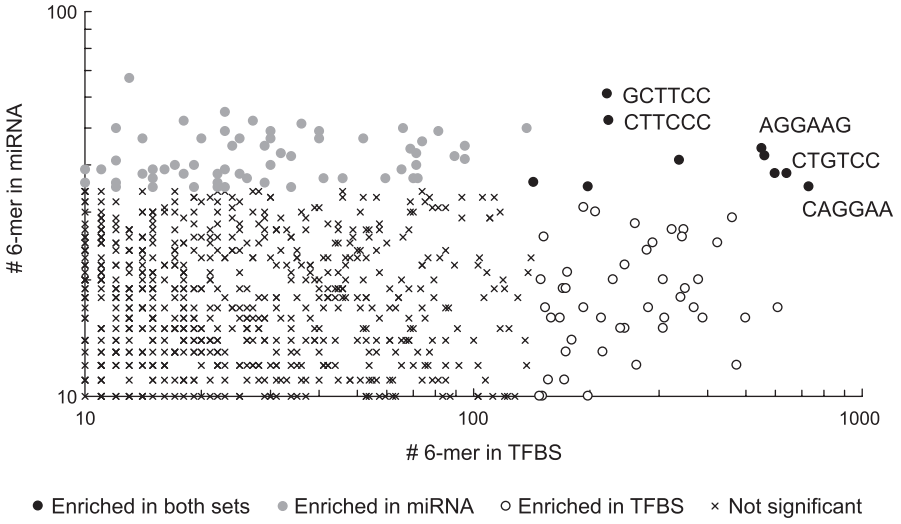

Fig. 2. Correlation between frequencies of 6-nt sequences in TFBS and miRNA.

After applying the 2σ cutoff in both sets i.e. frequencies of 6-mers (in non-redundant list of 2080) above 140 and 34 for TFBS and miRNA respectively all the 6-mers were separated in enriched in both sets, in TFBS only, in miRNA only, and not significantly enriched. 4 groups of 6-mers are clearly visible that are specific for TFBS and miRNA. The most frequent 6-nts are labelled.

the distribution for the miRNA is relatively more uniform and therefore we were left with a larger number of 6-nt sequences (724 sequences). For all the 6-nt sequences the frequencies for the TFBS and the miRNA are multiplied and the sequences were next sorted as per the resulting product. The result was 57 sequences (that are common to both sets with high frequencies). The final results were summarized in Table 2 (partial list).

We have started with 65 human TFs from the JASPAR database but these binding sites are considerably degenerate and we therefore end up with 6496 putative binding sites. We compared these binding sites with the 6-nt sequences from the miRNA database and located 57 sequences that can be considered as mapping these 6496 binding sites back into the TFs. The correlation is significant but we also need to explain the presence of other 6-nt sequences with high frequencies (Table 2).

Finally TFs bind to the same sequence (or very close to it) and this process may be helped by the presence of other 6-nt sequences with high frequencies present in the TFBS database. It may be useful to map the various 6-nt sequences to individual TFBS but this need to be done manually as there is no clear way to relate the 6496 binding sites to the 65 human TFs. It is indeed interesting to find such a strong correlation between the TFBS and the miRNA.

We analysed most common 6-mers in both TFBS and miRNA set by similarity to known PWM (position weight matrices) in JASPAR and TRANSFAC databases using STAMP software (Mahony, Benos, 2007). Common pattern (of top 5 common oligonucleotides) is close to AGGACAG / CTGTCCT (Table 2). Fig. 3 a contains this motif as logo.

Table 2

The sequences that have high frequencies in both TFBS and miRNA datasets

6-mers common for TFs and miRNA	No. in numerated list (1-4096)	Number of occurrences in TF set (together with complement)	Number of occurrences in miRNA set (together with complement)
GCTTCC	2550	221	61
CTTCCC	2006	223	52
AGGAAG	643	554	44
CCAGGA	1321	564	42
CTGGGA	1961	338	41
CTGTCC	1974	643	38
CCTGGA	1513	599	38
CCAGGG	1323	143	36
CAGGAA	1185	731	35
CTTCCA	2005	198	35

Next we compared the consensus sequences found in the present study to known PWM (positional weight matrices). Majority of similar matrices were related to ETS family of transcription factors. Members of the large ETS family of transcription factors (TFs) have highly similar DNA-binding domains and have diverse functions and activities in physiology and oncogenesis. Some differences in DNA-binding preferences within this family have been described, but differences in sequence are minor (Wei *et al*., 2010).

Fig. 3 b shows most similar motif for this common pattern (core region TCCT), as estimated by STAMP program.

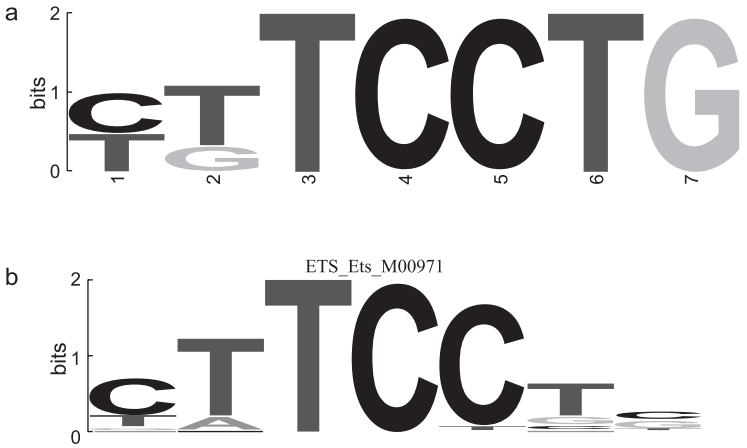

Fig. 3. a – the logo of common motif for TF and miRNA oligonucleotides; b – ETS binding motif, most close binding matrix from TRANSFAC database for common 6-mer found in both TF and miRNA sets.

Contrary, most common pattern of TF, which is not present in miRNA, is AGGTCA (See Table 2). Comparison to the database of binding motifs by STAMP shows that it is similar to RORα1motif. Nuclear receptor retinoid-related orphan receptor alpha (RORα1) is a member of ROR-family receptors. It is broadly expressed in various tissues and organs during embryonic development (Benderdour *et al*., 2011).

Overall, the DNA base composition for whole genome, mRNA and regulatory regions are different. It is known that genetic code imposes statistical constraints on protein-coding DNA sequences. Content of miRNA is limited by presence or avoidance hairpin loops in RNA structure. In general, DNA text contains different codes, or information messages that could be superimposed: classical triplet code, DNA shape code, chromatin code, gene splicing code, nucleosome positioning code and other, including those that have not yet been formally described (Trifonov, 2011). Each could be associated to the constraints imposed by information content and reflecting in oligonucleotide frequencies, number of poly-A tracts and text complexity (Orlov *et al*., 2006). Identifying regions of DNA with extreme statistical characteristics is an important aspect of the structural analysis of genomes and sequenced genome fragments.

Acknowledgement

Authors are grateful to Drs. D. Afonnikov, V. Levitsky and M. Ponomarenko for critical comments. The software was tested on high-throughput computer cluster at ICG. The work is supported in part by RFBR 11-04-01888, 11-04-92712-IND, the Russian Ministry of science and education (projects N 07.514.11.4011, 07.514.11.4023, 857).

References

Benderdour M., Fahmi H., Beaudet F. *et al*. Nuclear receptor retinoid-related orphan receptor α1 modulates the metabolic activity of human osteoblasts // J. of Cell. Biochem. 2011. V. 112. N 8. P. 2160–2169.

Bulyk M.L., Johnson P.L.F., Church G.M. Nucleotides of transcription factor binding sites exert interdependent effects on the binding affinities of transcription factors // Nucl. Acids Res. 2002. V. 30. P. 1255–1261.

Fickett J.W., Hatzigeorgiou A.C. Eukaryotic promoter recognition // Genome Res. 1997. V. 7. P. 861–878.

Mahony S., Benos P.V. STAMP: a web tool for exploring DNA-binding motif similarities // Nucl. Acids Res. 2007. V. 35 (Web Server issue). P. W253–W258.

Man T.K., Stormo G.D. Non-independence of Mnt repressor-operator interaction determined by a new quantitative multiple fluorescence relative affinity (QuMFRA) assay // Nucl. Acids Res. 2001. V. 29. P. 2471–2478.

Orlov Y.L., Te Boekhorst R., Abnizova I.I. Statistical measures of the structure of genomic sequences: entropy, complexity, and position information // J. Bioinform. Comput. Biol. 2006. V. 4. N 2. P. 523–536.

Putta P., Mitra C.K. Conserved short sequences in promoter regions of human genome // J. Biomol. Struct. Dyn. 2010. V. 27. N 5. P. 599–610.

Trifonov E.N. Thirty years of multiple sequence codes // Genomics Proteomics Bioinformatics. 2011. V. 9. N 1/2. P. 1–6.

Wei G.H., Badis G., Berger M.F. *et al*. Genome-wide analysis of ETS-family DNA-binding *in vitro* and *in vivo* // The EMBO J. 2010. V. 29. N 13. P. 2147–2160.

Wray G.A. The evolutionary significance of cis-regulatory mutations // Nat. Rev. Genetics. 2007. V. 8. P. 206–216.

Yamamoto Y., Ichida H., Matsui M. *et al.* Identification of plant promoter constituents by analysis of local distribution of short sequences // BMC Genomics. 2007. V. 8. N 67. P. 1–23.

ОТНОСИТЕЛЬНО КОНСЕРВАТИВНЫЕ ОБЩИЕ КОРОТКИЕ ПОСЛЕДОВАТЕЛЬНОСТИ В САЙТАХ СВЯЗЫВАНИЯ ТРАНСКРИПЦИОННЫХ ФАКТОРОВ И миРНК

П. Путта[1, 3], Ю.Л. Орлов[2], Н.Л. Подколодный[2], Ч.К. Митра[3]

[1] Медицинская и молекулярная генетика, Школа клинической и экспериментальной медицины, Университет Бирмингема, Бирмингем, Великобритания;
[2] Учреждение Российской академии наук Институт цитологии и генетики Сибирского отделения РАН, Новосибирск, Россия;
[3] Университет Хайдарадаба, Хайдарабад, Индия

Сайты связывания транскрипционных факторов (ССТФ), специфичные ДНК-связывающие мотивы, которые распознаются факторами транскрипции, – это короткие и часто вырожденные последовательности. Специфичное к последовательности связывание транскрипционного фактора к ДНК контролирует регуляцию экспрессии генов на уровне транскрипции. Высокопроизводительное геномное секвенирование дает растущие большие объемы данных по профилям связывания транскрипционных факторов в геноме, требующие разработки новых средств анализа. Мы разработали компьютерную программу для изучения коротких мотивов в последовательностях ДНК. Нас интересовало изучение общих характеристик последовательностей сайтов связывания транскрипционных факторов, которое может помочь в определении консервативных паттернов на нуклеотидном уровне. Была исследована роль общих олигонуклеотидных паттернов в ССТФ и миРНК, основанных на близости последовательностей между этими двумя наборами.

Ключевые слова: геном, сайты связывания транскрипционных факторов, миРНК, олигонуклеотиды, статистика, секвенирование.