An Unusual Distribution of 6-nt Sequences Near The Transcription Start Site

Chanchal K. Mitra[1,2] and Luciano Milanesi[1]

[1]Institute of Biomedical Technologies, Segrate, Milan 20090, Italy

Summary

A new look at the transcription start is presented in which we can see transcription factors binding to both sides of the TSS as an essential requirement. Naturally the factor binding to the downstream region must be removed so that transcription process can continue. The presence of a number of distinct transcription factors also can be used to explain selective activation of various genes. The transcription start site by itself plays only a minor role in the whole process. We also suggest that mutations close to the TSS on the coding side can be fatal even if preserves the codon table.

1 Introduction

Eukaryotic transcription factors are structurally more complex and usually contain no characteristic conserved sequences as commonly found in prokaryotic systems [1]. This is attributed to the evolutionary pathway in which we need to organize functional and important genes for a ready access [2, 3]. However, common studies have targeted the transcription start site (TSS) as a signal to detect a functional gene. This approach works well for the prokaryotic sequences but does not work well for the eukaryotic sequences. Using the data from the SIB-EPD database for humans, we show that the problem is actually far more complex that it is currently believed.

In our earlier work on the similar lines, we did lump together all the eukaryotes together and studied their collective behaviour. This approach is useful only to locate gross characteristics but fails to show the details. In the present investigation, we have focused on the human sequences and have been able to see the results a bit more clearly.

2 Methodology

The promoter sequences were downloaded from the SIB-EPD (`http://www.epd.isb-sib.ch/`) site for Homo sapiens (human) selecting the option of representative set of not closely related sequences from –100 to +100 relative to the transcription start site. The nucleotide distributions were studied using small C programs written in gcc.

To study the distribution of hexanucletodies, we wrote a short C program to align pairwise all the sequences and locate all the common subsequences that are at least 6 nt in length. The program printed out all common subsequences with length greater than 5. All possible pairs (n-1)·(n-2)/2 were considered all the common sequences were sorted using gnu sort. For this

[2]To whom correspondence should be addressed. E-mail: `c_mitra@yahoo.com`

doi:10.2390/biecoll-jib-2008-103

experiment, we divided the downloaded sequences into two equal parts (corresponding to the upstream and downstream regions). We considered the most common 50 hexanucleotides on both promoter and coding sides. We also studied the actual positional distribution for the five most common 6-nt sequences on both sides of the TSS.

All programs were written in C in the gcc compiler environment and were run on a cluster running linux operating system. All the plots were made using the commercial software Sigmaplot.

3 Results

There are 1871 sequences in the downloaded file. We have studied only the region –100 to +100 (with respect to the transcription start site) and therefore each sequence has exactly 201 nucleotides. Out of the 1871·201=376071 total nucleotides, we had 605 bad characters (N: unknown base). This region is highly GC rich, as seen by the following base composition: A=64984 (17.30%), C=114744 (30.56%), G=125530 (33.43%) and T=70208 (18.70%). It is important to note that when we select "all promoters", we get a set of 4809 sequences and the GC bias in the complete set is far less compared to what we see above. This automatically suggests that the conclusions we derive cannot be extended to other families and we should study them in a similar fashion independently.

The distribution of the individual nucleotides in this interval is also non-uniform. However, we present the distribution of the four important dinucleotides (AT, TA, CG and GC) in the interval studies in Figure 1. There are 1871·(201-1)=374200 possible dinucleotides and 4·4=16 possible combinations and assuming all are equally likely we expect 374200/(16·200)=117 as a mean estimate for the frequency for any given nucleotide at any given position. We see in Fig. 1 that the AT and TA frequencies are uniformly below this notional value (except for a peak near –25) and GC and CG frequencies consistently above this value (except for a negative peak around –25 and a broad hump near +25). The sharp peak at TSS is due to the presence of the start codon. The observed "noise" in the distribution is consistent with the sample size studied. The graph suggests that there exists some "signal" in the broad region of TSS+/- 25 nucleotides. We also note the prominent peak in the AT/TA distribution ˜-25 and this corresponds to a negative peak in the in the GC/CG distribution. Around +25, the AT/TA distributions appear normal but the GC/CG distributions show a split. We see that the distribution curves for AT and TA (bottom pair) follow closely (practically identical) whereas the distribution curves for CG and GC show some interesting difference in the region of (0-50). This aspect needs further studies. A peak in the TA and AT pairs (near –25) is naturally balanced by a negative peak in the GC and CG pairs. We have not seen any significant visual features in the frequency plots of the remaining 12 nucleotide pairs (to improve the visual clarity, these plots have not been presented).

To study the hexanucleotide distributions, we search the sequences using the 50 most common sequences. The raw frequencies are shown in Figure 2. We note that sequences in the upstream regions are usually more common but only slightly. We see that the two most common hexanucleotides are common on both sides of the TSS. The shape of the distribution is intriguing as it reminds one of a typical sigmoidal curve. It perhaps has origin in some internal cooperative process. The most interesting part of the graph is the relatively extended horizontal plateau. The shape of the graph is difficult to explain using conventional wisdom. We have studied the positional distributions of the five most common hexanucleotides (both upstream and downstream regions).

doi:10.2390/biecoll-jib-2008-103

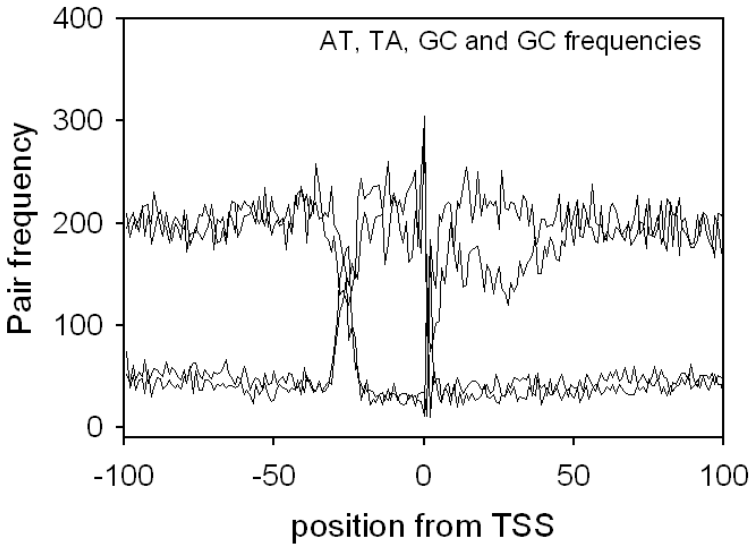

Figure 1: The four dinucleotides AT, TA, GC and CG frequencies have been plotted as a function of their position from the transcription start site. The 12 other dinucleotides have been omitted from this graph in order to improve the visual clarity. The lower graphs refer to the AT/TA and the upper graphs refer to GC/CG. We claim that the distributions in the region TSS-50 to TSS+50 is anomalous. The very sharp spike at the TSS is caused by the start codon.

If we consider a uniform distribution in which ACGTs are equally likely, the probability of any given hexanucleotide will be 4-6=2-12 or 1 in 4096. In the present sample, we expect any random hexanucleotide to occur 1871·(201-5)/4096 or about 90 times. The highest frequencies (>900) cannot be explained by statistical fluctuations. Even if we consider the fact that the given region under the present study is rich in GC and a large majority of the common hexanucleotides is composed of only GCs, we expect a frequency of less than 400 (this value approximately corresponds to the horizontal part of the graph). We have more than 30 sequences (on both upstream and downstream side) that are above this value. For a clear comparison, we have computed the expected frequency for the given sequence (using the actual proportions for A, C, G and T) and shown that in parenthesis in Table 1. The differences can be now seen as striking. The sample size is rather large (for statistical purposes) and the 99% confidence interval cannot be given exactly as it depends on the actual frequency (but is expected to be small: +/- 1 for the larger frequencies and increasing to +/- 3 for the smallest frequencies.

The rational conclusion that can be drawn from the above graph is that the sequences in the rising part of the curve (the first 10-15) are somehow important as they appear to be significant in some way. Although the present investigation cannot determine the exact nature of the significance, we can reasonably suggest that they are somehow involved in the binding of the transcription factors and this study shows clearly that the factors can bind well on both sides of the TSS. To study this in more detail, we looked into the positional distribution of the hexanucleotides in the original database.

In Figures 3(a) and 3(b), we have plotted histograms for the frequencies of the 10 most common subsequences (5 upstream + 5 downstream) by looking into the actual position of the hexanucleotide in the downloaded sequences. The individual sequences and their actual frequencies are shown in the respective graphs. For a random distribution, we would have seen a curve approximating a normal distributions. However, none of the plots appear to represent a normal distributions and in fact some of the distributions appear to be bimodal.

doi:10.2390/biecoll-jib-2008-103

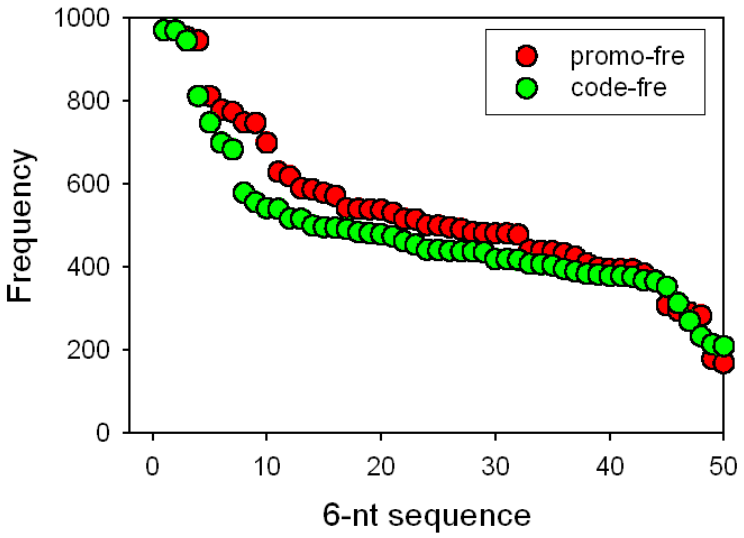

Figure 2: The frequencies of the 50 most common 6-nt sequences plotted after arranging the frequencies on a decreasing scale. The 6-nt sequences for the promoter side and the coding side are not necessarily identical and the x-axis represents simple ordinality. The actual sequences corresponding to these frequencies are given in the "additional material" section. The graph shows typical "sigmoidal" shape suggesting some internal cooperative process.

For computational simplicity, we have chosen a "bucket size" of 8 in plotting these histograms (there are 25 vertical steps in all the graphs) and therefore the actual positions can be off by as much as 8 nucleotides. Nevertheless, the general idea that the most common hexanucleotides on the promoter side are found around –50 (with respect to the TSS) and they provide a very likely site for the binding of the transcription factors. For the most common sequences on the coding side, we note that they are also common on the promoter side and we see a very clear bimodal distribution in all the cases. It would be interesting to study whether the given subsequence occurred multiple times in the same sequence. The general conclusion one can make from this graph is the sequences tend to cluster at or near –50 but the distributions are quite broad. In fact, another smaller peak can be seen at or near –25 which is quite interesting (in Fig. 3(a)). In Fig. 3(b), we notice that the sequences (one some of them) found on the coding side are also seen on the other side, although with a smaller frequency. The last sequence seen in Fig. 3(b) (GCGGCC) appears to be more or less uniformly distributed in position.

4 Discussions

It is usually considered that the eukayotic promoter region is quite complex because of the lack of consensus sequences [4]. The TSS by itself cannot act as a signal as it is too weak to provide a reliable signal. On the other hand, sequences as long as 5-8 can be easily recognized by various factors and the longer sequences can be relatively unique. If we consider 20K genes are an estimate, we shall expect a similar number of TSS and based on the estimates we have given above, we shall expect the total genome size to be close to 80Mbase, which is approximately 1/40th the current estimate. This suggests that other factors may be playing additional and important roles.

doi:10.2390/biecoll-jib-2008-103

No	Sequence (promoter)	Frequency	Sequence (code)	Frequency
1	GGCGGG	970 (468)	GGCGGG	970 (468)
2	GGCGGC	969 (428)	GGCGGC	969 (428)
3	GGGCGG	953 (468)	GCGGCG	945 (428)
4	GCGGCG	945 (428)	CGGCGG	812 (428)
5	CGGCGG	812 (428)	CCCGCC	748 (327)
6	GCGGGG	779 (468)	GCCGCC	699 (357)
7	CCGCCC	774 (327)	CGCCGC	683 (357)
8	CCCGCC	748 (327)	GCGGCC	579 (391)
9	GGGGCG	747 (468)	CCGCCG	556 (357)
10	GCCGCC	699 (357)	GGGAGG	542 (265)
11	CGGGGC	629 (428)	GGCCGG	540 (428)
12	CCCCGC	618 (327)	CGCGGC	517 (391)
13	GCGCGG	590 (428)	GCCGGG	516 (428)
14	CGCCCC	587 (327)	GGCCGC	500 (391)
15	GCGGCC	579 (391)	GCCGCG	496 (391)
16	GCGCGC	571 (391)	GCGGAG	495 (242)
17	GGGAGG	542 (265)	CCCGGC	491 (357)
18	GGCCGG	540 (428)	GCCCGG	484 (391)
19	GCCCCG	539 (357)	CCGGGC	482 (391)
20	CGGCGC	538 (391)	GCTGGG	479 (262)
21	GGCGCG	531 (428)	GGGCCG	474 (428)
22	GCCGGG	516 (428)	GGCTGC	462 (239)
23	GCGGGC	515 (428)	CGGCCG	454 (391)
24	GGCCGC	500 (391)	CCGCGG	442 (391)
25	GCCGCG	496 (391)	GGGGCC	441 (428)
26	CCCGGC	491 (357)	GGAGGC	439 (242)
27	GCCCGG	484 (391)	GGCTGG	438 (262)
28	CCGGGC	482 (391)	GCTGCG	438 (239)
29	GGAGGG	481 (265)	GCTGCT	436 (134)
30	CGCGCC	481 (357)	TGGCGG	419 (262)
31	GCGCCG	478 (391)	GAGGAG	419 (137)
32	GGGGCC	441 (428)	CTGCTG	418 (134)
33	GGAGGC	439 (242)	CCCGGG	408 (391)
34	GGGGGC	438 (468)	GGAGCC	407 (221)
35	AGGCGG	433 (265)	GGCAGC	396 (221)
36	CCTCCC	425 (183)	CGGAGC	390 (221)
37	GGCGCC	409 (391)	GGCTCC	384 (219)
38	GAGGGG	399 (265)	GCAGCC	382 (202)
39	GGCGGA	395 (265)	GGTGAG	379 (148)
40	GGCCCC	395 (357)	GCTGCC	379 (219)
41	GCGCAG	394 (221)	CCGGGG	377 (428)
42	GCCCCC	384 (327)	CGCTGC	368 (219)
43	CGGAAG	367 (125)	TGCTGC	366 (134)
44	AGGGGC	308 (242)	GCGGCT	353 (239)
45	TTCCGG	295 (134)	CCTGGG	313 (239)
46	CCGGAA	291 (115)	ATGGCG	270 (148)
47	GGAAGT	283 (77)	GTGAGT	234 (83)
48	GATTGG	179 (83)	CCATGG	214 (124)
49	TATAAA	169 (11)	CATGGC	210 (124)

Table 1: The distribution of the most common 49 hexanucleotide sequences around the transcription start site (the reported frequencies cover the entire range –100 to +100). The numbers in parenthesis are the expected frequencies computed statistically.

In addition to the TSS regions, they are perhaps other regions in the genome that are GC rich. We believe it to be an inefficient process to scan the whole genome to locate the desired location. Just like a computer does not scan the whole disk to locate a given file (it maintains a directory) the nature need not look all over to locate the gene [5]. Biologists have proposed some mechanisms that are apparently ad-hoc but worth investigating from the biostatistics angle. A simple suggestion, not experimentally confirmed, is that two factors must straddle the TSS and this should cause the other factors to be bound sequentially. This idea has the roots in the Fig. 2, otherwise we cannot explain the presence of these sequences on the coding side. This essentially means that the assembly of the transcriptosome is more complex that is presently believed. However, somehow the factors on the coding side must be removed prior to transcription initiation and that is perhaps the reason that experiments have failed to see these sites [6].

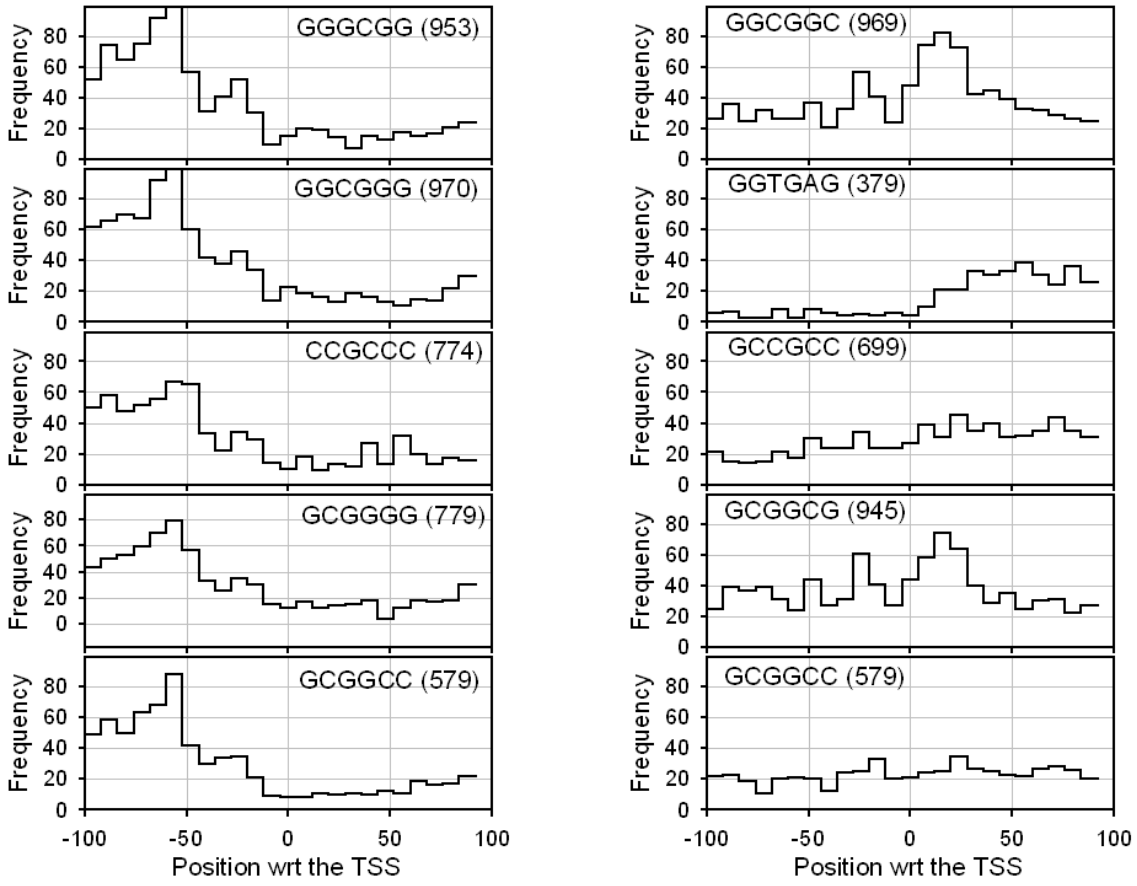

Figure 3: (a) The distribution of the most five common hexanucleotides on the promoter side of the TSS are seen above. It is clear that the most of them show a peak at/around –50 upstream from the TSS. The actual sequences and their total counts are seen in the respective graphs. This distribution has been plotted as a histogram with a bin size of 8. All the graphs have been plotted with an identical scale for ease in visual comparison. (b) The positional distribution of the five most common hexanucleotides on the coding side of the TSS. We note that some of them show peaks on both sides of the TSS. The actual sequences are indicated along with the frequency in parenthesis. We see clear peaks very close the TSS, which is expected (see text for details).

References

[1] S. L. Baldauf. The deep roots of eukaryotes. *Science*, 300(5626):1703–1706, 2003.

[2] D. Ashok Reddy and Chanchal K. Mitra. Comparative analysis of transcription start sites using information content. *Genomics, Proteomics and Bioinformatics*, 4(3):189–195, 2006.

[3] T. S. Rekha and C. K. Mitra. Frequency analysis of the splice site regions of different organisms. *Journal of Integrative Bioinformatics*, 4(2), 2007.

[4] Chuhu Yang, Eugene Bolotin, Tao Jiang, Frances M. Sladek, and Ernest Martinez. Prevalence of the initiator over the TATA box in human and yeast genes and identification of DNA motifs enriched in human TATA-less core promoters. *Gene*, 389(1):52–65, 2007.

[5] Vladimir B. Bajic et al. Mice and men: Their promoter properties. *PLoS Genetics*, 2(4):e54, 2006.

[6] J. Ponjavic et al. Transcriptional and structural impact of TATA-initiation site spacing in mammalian core promoters. *Genome Biology*, 7(8):R78, 2006.

doi:10.2390/biecoll-jib-2008-103