Analysis of eukaryotic promoter sequences reveals a systematically occurring CT-signal

Niels I. Larsen, Jacob Engelbrecht* and Søren Brunak

Center for Biological Sequence Analysis, Department of Physical Chemistry, The Technical University of Denmark, Building 206, DK-2800 Lyngby, Denmark

Received November 8, 1994; Revised and Accepted February 14, 1995

ABSTRACT

A general data study of eukaryotic promoter sequences from widely different species is presented. Mammalian promoters with known transcription initiation sites represented the largest subclass of the data, and for this group neural network algorithms were trained to predict the location of the initiation site in a test set. The prediction accuracy of this local method was higher than what could be expected from the known non-local structure of eukaryotic promoters. Subsequent analysis revealed, besides the consensus of the two known important subregions: the TATA-box TATAAA and the Cap-signal CA, a CT-signal positioned on the average seven nucleotides downstream of the transcription initiation site. The consensus of the CT-signal is CTNCNG. The details of this core promoter element were disclosed using multiple alignment and have earlier only been described in a few isolated examples.

INTRODUCTION

The signals for initiation of transcription have a much more complex organization in eukaryotes than in prokaryotes (1,2). Eukaryotic cells genes fall into three classes, defined by their type of promoter corresponding to a specific RNA polymerase. Here we restrict ourselves to promoters interacting with RNA polymerase II only.

Eukaryotic RNA polymerases are unable to accurately initiate transcription by themselves and successful initiation is governed by a group of auxiliary proteins—the transcription factors (1–7). Different kinds of sequence elements are found at promoters: elements near the initiation sites and elements further away from the initiation site. Positions and distances below are given relative to the transcription initiation site at position 0.

Sequence elements near the transcription start site, regarded as necessary and sufficient for specific transcription initiation, are the TATA-box and the initiator (Inr) (1,5,6,9–11). Either the TATA-box or the Inr element or both may be used in making the transcription complex, and apparently all promoters have at least one of them. The TATA-box is located 20–30 nt upstream and has the consensus sequence TATAAA. It is usually surrounded by

G/C-rich regions (11), which may facilitate the opening of the DNA at the TATA-box location. The Inr element, which may vary in length from 2–20 nt, overlaps the transcription initiation site. Inr elements have been grouped into families based on sequence homology (1). Common to many of the Inr elements is the dinucleotide CA which often is located at position –1 or 0. An experimental study of a range of Inr sequences has revealed loose but consistent sequence requirements of Inr elements (12). The TATA-box is recognized by the TATA box-binding protein (TBP), which binds as a monomer to the minor groove of promoters (7). It has been shown that TBP is present in the transcription complex even when the TATA-box is absent, at the position where the TATA-box normally is located (13).

Other transcription signals have been investigated in detail, e.g., the GC-box and the CAAT-box (9,14). The distance between selected transcription signals has also been analyzed (15) and a general statistical study of nucleotide occurrences has been performed (16). The distances between the transcription signals and the initiation site have been thoroughly investigated (8,9,17,18). These signals have a very large variation in their observed distances to the transcription initiation site.

Several methods have been used to investigate DNA sequences with the purpose of analyzing binding sites, for example weight matrix descriptions of already well known promoter elements (9), multiple alignment of *E.coli* promoters using simulated annealing (19), small region frequency analysis (8,14,15,17,18,20), scanning of DNA sequences for known transcriptional elements (21,22), measurement of the Shannon information content of an aligned group of related binding site sequences (23,24) and neural networks, see (25) for a review.

Most methods have the common problem, when used for prediction, that they do not exploit all the available sequence information because they assume positional independence. Positional independence means that the choice of a given nucleotide at one position is treated independently of the choice at any other nucleotide position (26). For promoters there is extensive genetic evidence to the contrary in form of context-dependent and/or compensatory mutations (26). Neural network algorithms of the type used in this study do not treat the positions independently, but take also into account correlations of higher order. Neural networks have also been used on related problems (27), and for promoter prediction in *E.coli* (26,28–31).

* To whom correspondence should be addressed

In the earlier work on promoter prediction in *E.coli*, the approach has been either to use randomly generated sequences as negative examples (26,28,30,31) or sequences randomly chosen from coding regions (29). The investigation presented here use negative examples from the promoter regions themselves to get as close as possible to the biological discrimination problem. Previously reported results are not easily comparable due to differences in the data material (29).

In general the data used in DNA sequence analysis suffers from the uncertain knowledge about the quality of the data. Transcription initiation sites given in the database might be hypothetical, inferred from published consensus patterns. In the worst case this means that consensus patterns are rediscovered by complicated methods in a circular manner. However, the consistent finding of a new pattern reported below, make us believe that the reported analysis is in fact valid.

MATERIALS AND METHODS

Nucleotide sequence data

Using the eukaryotic promoter database EPD (32), 1123 promoter sequences were extracted from the EMBL Data Library Nucleotide Sequence Database (33). Each of the extracted promoter sequences consisted of 250 nt (where data was available) before and after the transcription initiation site, giving a maximum of 501 nt in each extract. Entries were discarded: (i) if they contained more than one transcription initiation site, (ii) if the nucleotide at the transcription initiation site was not known, and (iii) if they belonged to the database divisions organelles or unclassified. This left 1006 sequences: 198 primate, 231 rodent, 51 other mammal, 104 other vertebrate, 195 invertebrate, 95 plant and 132 virus sequences. We will refer to nine different data sets containing the extracted sequences: the seven different divisions, a merged mammalian data set and the complete data set.

Homologous promoters were extracted (a total of 130). Two promoters were considered homologous if they exhibited a sequence identity >50% between positions -79 and +20 (32). The performance of prediction algorithms like neural networks will increase if the test set contains sequences similar to sequences present in the training set (25). Therefore each homology group was present in the test or the training set only.

Quantification of sequence information content

The Shannon information measure (34) used was calculated by the formula

$$I(i) = - \sum_x P_i^x \log_2 P_i^x, \quad (1)$$

where P_i^x is the probability of occurrence for a particular nucleotide X (A, C, G or T) at position i (23). The unit of information is bits/nucleotide. $I(i)$ falls in the range 0.0–2.0 bits/nucleotide, 0.0 meaning complete order and 2.0 random occurrence of nucleotides at a given position. The measure is useful to detect regions of 'non-randomness' and has been used before on DNA-sequences with success (see e.g. 23,27,35). A correction term for $I(i)$ compensating for the effect of finite data set sizes may be computed (36). The maximal finite set correction when computing the information content was of the order of 10^{-3} bits for

the complete data set, 10^{-2} bits for the mammalian part, and due to the low number sequences in 'other mammals' 10^{-1} bits.

Neural networks

Architecture. The neural networks used for locating transcription initiation sites were of the feedforward type and had three layers of processing units; see (27,37) for a detailed description. The input values were obtained by coding the DNA sequence into a binary string. The binary string was composed using a sparse coding, (A = 0001, C = 0010, G = 0100, T = 1000 and X = 0000), which has been shown (28) to give a substantially better performance than a more compact coding (A = 00, T = 01, G = 10, C = 11), due to the identical distance between the nucleotide encodings. The output layer consisted of one neuron only deciding whether the middle input nucleotide (when using symmetrical windows) was a transcription initiation site or not.

The network weights were initially randomly assigned values of 0.1 or -0.1. The learning rate was fixed at 0.05, higher learning rates caused oscillations. Weights and thresholds were adjusted using the backpropagation algorithm (37). During a training cycle all possible windows obtained from a training set were presented to the network in random order. In searching for a good network predictor the number of neurons in the intermediate layer was varied from 0 to 100. A large number of different window sizes (including asymmetrical windows) was assessed.

Evaluation measure. A good measure is needed when the performance of the different network configurations are to be compared. A network which correctly predicts all the true transcription initiation sites in a test set is of little use when the false positive rate is 50%. A measure (38) which includes both the correctly and falsely predicted positives and negatives is the correlation coefficient

$$C = \frac{PN - N^f P^f}{\sqrt{(N + N^f)(N + P^f)(P + N^f)(P + P^f)}} \quad (2)$$

P and N are the correctly predicted positives and negatives, and P^f and N^f are the incorrectly predicted positives and negatives. The correlation coefficient C is 1.0 for a perfect prediction, -1.0 when totally incorrect, and 0.0 when uncorrelated.

Training/test data set. Only the merged mammalian data set was used for prediction purposes. The data was split into a training and a test set by random selection, but it was ensured that the different homology groups were placed either in the training or the test set, thus preventing an artificial increase in the performance. The input windows were restricted to the following regions: [-250; 250], [-150; 150], [-100; 100] and [-60; 60] meaning that the sliding window was scanning the promoter sequences in the given region only. To obtain reliable results a fairly large test set was needed, (about one third of the sequences, and thus about two thirds remained for training). The exact number of sequences in each set depends upon the choice of region due to the lack of data in some of the entries. For the region [-60; 60] the training/test set consisted of 365 learning and 92 test examples giving a total of 18 615 and 4692 different input windows, respectively. The region [-100; 100] gave 332 and 81 sequences (43492/1061y1), [-150; 150] 286 and 74 (66066/17094) and with [-250; 250] only 221 and 55 sequences (95251/23705) were left in the training and test set, respectively.

Multiple alignment

Earlier work (9) has concluded that multiple alignment cannot be applied to eukaryotic promoters because the method implicitly assumes a rigid promoter architecture with the transcription signals occurring in the same order and orientation. When applied to subsequences multiple alignment can, however, reveal new information because it emphasizes small important signals like the Cap signal. Global multiple alignment of large data sets has not been possible until recently (19). Simulated annealing (39) is used to construct alignments with a high number of matches between the symbols in each column. The method builds on the standard Metropolis–Monte Carlo algorithm (40).

Starting from a set of unaligned sequences, gaps and shifts at random positions are inserted and removed at each step. As the alignment progresses only very favorable changes are accepted, the rate of ‘cooling’ determines the final quality of the alignment. The details have been reported earlier (19), where the method was used to align *E.coli* promoters.

A weight matrix method

A weight matrix method was used in further analysis of the interesting consensus sequences. The weight matrix is constructed by counting the occurrence of the nucleotides at each position of an alignment of the site in question. The matching score x of a given stretch of l nucleotides is

$$x = \sum_{i=1}^l \ln(f_i), \quad (3)$$

where f_i is the weight of the nucleotide found at the i 'th position of the stretch. The maximum matching score at position p of an oligonucleotide in a given freedom region of length r (where $r \geq l$) is denoted $x_{\max}(p, r)$. For example, the value of $x_{\max}(2, 10)$ for the pattern TATAAA in a sequence is the maximum of the five possible matching scores (starting at positions 2, 3, 4, 5 and 6). In this work it is desirable to investigate certain patterns without using the prior knowledge of the nucleotide frequency, therefore the nucleotide frequency of the consensus letters will be set to 100%. This means, for example, that given the TATA-box the nucleotide frequency will be set to T (100%) A (100%) T (100%) A (100%) A (100%) A (100%) on the respective positions. The average normalized maximum matching score $A(p)$ at position p for a data set S with N sequences is given by

$$A(p) = \frac{1}{N} \sum_{j=1}^N \ln \left(\frac{x_{\max}^j(p, r)}{x_{\max}} \right), \quad (4)$$

$$x_{\max}^j(p, r) = x_{\max}(p, r) \text{ on the } j\text{'th sequence in } S. \quad (5)$$

Note that $x_{\max}$ is the maximum matching score obtainable for a given oligonucleotide which is the length of the oligonucleotide in our case. The freedom region is introduced because a simple frequency plot ($r = l$) of a consensus can fluctuate drastically.

RESULTS

Below we report Shannon information content, nucleotide occurrences and nucleotide frequency distributions on seven

different eukaryotic data divisions: primates, rodents, other mammals, other vertebrates, invertebrates, plants and viruses. A data set of mammalian promoter sequences is further investigated using neural networks, multiple alignment and weight matrix methods.

Statistical description of the promoter sequences

Shannon information. The Shannon information for the nine data sets is shown in Figure 1. The plots for plants, other vertebrates and especially other mammals are more noisy due to the low number of sequences. All plots are somewhat smeared because the promoter signals do not have a fixed distance relative to each other. All plots except viruses show a strong non-random region at the TATA-box and a weaker signal at the transcription initiation site. Viruses show a dip in the TATA-box region only. An interesting observation is the dips between the two strong signals occurring in all mammalian data sets. This indicate the existence of important sequence elements besides the TATA-box and the Cap-signal. Note also the dip at position seven in mammals—the position of the novel CT-signal described below.

Nucleotide distributions upstream and downstream

Table 1 gives the average nucleotide distribution in different subregions. The nucleotide distribution in the complete EMBL sequences corresponding to the EPD promoter entries was calculated and used for comparison and denoted by $[-\infty; \infty]$. When the figures for A, C, G and T do not add up to 100% it is due to unknown nucleotides in a few sequences.

When comparing the intervals $[-\infty; \infty]$ and $[-50; 50]$ we observe that (excluding plants and viruses) the density of nucleotides A and T is smaller and G and C larger in the core promoter region. We also observe that the A and G density is larger while smaller for C and T in the interval $[-50; 0]$ compared to $[1; 50]$. The figures for primates, rodents, other mammals and other vertebrates have the same mutual relations. The most significant seems to be the low occurrence of T (~20%) in $[-50; 50]$ with a corresponding higher occurrence of C and G in vertebrates, in contrast to the $[-\infty; \infty]$ T-content which is close to 25%.

Invertebrates and plants show other interval relations. The general occurrence found of A and T is close to 60% in invertebrates, where the most significant change is the 3% higher occurrence of G in the interval $[-50; 0]$. Also note the difference of 4% in G between the intervals $[-50; 0]$ and $[1; 50]$. The joint occurrence of A and T is over 60% in plants in general and we observe a change in the nucleotide distribution from interval $[-\infty; \infty]$ to $[-50; 0]$ quite different from the other data divisions: nearly 3% higher occurrence of A and C and nearly 4% lower occurrence of G.

Frequency distributions

Figure 2 shows the nine raw (non-multiply aligned) consensus sequences. As already mentioned the region from -30 to -24 (for plants and invertebrates $[-32; -21]$) represents the TATA-box region and $[-1; 0]$ the Cap signal. The consensus sequences for primates, rodents, other mammals, mammals, other vertebrates and viruses show the interesting result that nearly all consensus letters are C and G, except at the TATA-box region and the transcription initiation site. Combined with the low general

Figure 1. Shannon information content as function of the position in the promoter region [-250; 250]. The unit of information is bits/nucleotide. Note the two clear dips in the regions [-30; -24] (TATA-box region) and [-1; 0] (Cap signal).

occurrence of T it emphasizes the importance of the high occurrence in the TATA-box region. Invertebrates have an interesting consensus besides the TATA-box region and the transcription initiation site because the dominant nucleotides are T and A after the transcription initiation site. This is even more significant for plants where nearly all consensus nucleotides are T and A. A potential TATA-box may be found at position 36.

It is interesting to note that the frequency of G in the mammalian data in general is high between position -150 and position 0, peaking at 36% G at position -7.

The data observations for the data divisions primates, rodents and other mammals correspond well to the observations outlined in earlier works (8,9). However, Penotti's statistical analysis (8) for eukaryotic TATA-boxes and transcription initiation sites

Table 1. Nucleotide distribution

Division	Interval	#	nt	A	C	G	T
Primates	$-\infty; \infty$	195	551462	25.1	24.4	25.2	25.3
	-50;50	198	19734	22.7	28.5	28.4	20.4
	-50;0		9921	24.1	26.3	29.8	19.9
	1;50		9813	21.3	30.9	26.9	21.0
Rodents	$-\infty; \infty$	228	543332	24.9	24.8	25.0	25.3
	-50;50	231	22886	24.0	27.3	26.9	21.8
	-50;0		11564	25.6	25.1	27.8	21.5
	1;50		11322	22.4	29.5	26.0	22.1
Other mammals	$-\infty; \infty$	50	152291	26.9	22.5	23.0	27.5
	-50;50	51	5039	25.1	29.5	25.6	19.9
	-50;0		2545	25.9	27.0	28.5	18.7
	1;50		2494	24.4	32.0	22.6	21.0
Other vertebrates	$-\infty; \infty$	102	304026	25.7	24.0	24.7	25.5
	-50;50	104	10340	21.7	29.1	29.3	19.9
	-50;0		5182	22.8	27.5	30.8	18.9
	1;50		5158	20.6	30.8	27.8	20.8
Invertebrates	$-\infty; \infty$	159	369246	29.3	21.9	21.1	27.1
	-50;50	195	19381	28.0	22.8	22.2	26.9
	-50;0		9777	28.4	21.4	24.3	25.9
	1;50		9604	27.7	24.1	20.1	28.0
Plants	$-\infty; \infty$	94	216694	29.9	19.9	19.6	30.6
	-50;50	95	9439	34.2	23.0	15.5	27.3
	-50;0		4759	32.4	22.2	15.8	29.6
	1;50		4680	35.9	23.9	15.2	25.0
Viruses	$-\infty; \infty$	57	812060	21.9	28.5	28.8	20.8
	-50;50	132	12941	22.0	29.1	26.2	22.6
	-50;0		6494	23.7	27.8	26.1	22.4
	1;50		6447	20.3	30.4	26.4	22.8
All mammals	$-\infty; \infty$	473	1247085	25.2	24.3	24.8	25.6
	-50;50	481	47659	23.6	28.0	27.4	21.0
	-50;0		24030	25.0	25.8	28.7	20.5
	1;50		23629	22.2	30.3	26.0	21.5
All divisions	$-\infty; \infty$	885	2949111	25.2	24.8	25.1	24.8
	-50;50	1006	99760	25.1	26.8	25.3	22.8
	-50;0		50242	26.0	25.0	26.5	22.5
	1;50		49518	24.1	28.6	24.1	23.2

(based on 22 sequences from several species) differs from ours, which is based on a much larger dataset.

Prediction by neural network algorithms

The first issue to investigate was the importance and characteristics of local correlations by using different regions and observe the changes in the correlation coefficient. The input layer window in the neural networks was kept fixed at 71 nt. The region was set to [-250; 250], [-150; 150], [-100; 100] and [-60; 60] and the number of units in the intermediate layer was varied from 0 to 100. To obtain significant positive values of the maximum correlation coefficient C_{max} , at least two intermediate neurons

Figure 2. Raw (non-multiply aligned) promoter consensus sequences for primates (top), rodents, other mammals, other vertebrates, invertebrates, plants, viruses, (below the line) all mammals and the complete data set in the region [-50; 50]. Upper case letter means >50% occurrence and lower case letter means >33%. In case two consensus letters are needed they were placed above each other with the most frequently occurring at the top.

were needed for all regions showing that the transcription initiation site prediction and discrimination problem is indeed non-linear. For the test sets C_{max} was always positive and varied between 0.1 and 0.24 meaning that non-trivial local correlations are present in the nucleotide composition. To get as many samples as possible both in the training and test sets the [-60; 60] region was selected for further investigation.

Part of the information presented to the network is probably unimportant as already indicated by the Shannon plots. Trying to narrow down window positions with significant features, a series of runs was made. The number of intermediate units was set to 15, while the size of the non-symmetrical input windows was varied in one direction (starting at an anchor point) with a maximum of [-35; 35]. We started by a 71 nt window and in steps of one made the window smaller, eventually with only 1 nt in the window. The results are shown in Figure 3.

The curve with anchor point 35 represents runs where the window was varied in the direction right to left giving windows of [-35; 35], [-34; 35], [-33; 35], ..., [35; 35]. We observe a peak $C_{max} = 0.085$ with a window of [10; 35] indicating that important information may be present around +10 (small peak on the Shannon plot). No more information seems to be added until the TATA-box region (-30 to -24), where the curve raises to 0.25.

With anchor point 0 the runs had windows of [-35; 0], [-34; 0], ..., [1; 0], [0; 0]. We observe that the networks are totally unable to predict the transcription initiation site using just the first 5 nt [-5; 0] so the Cap signal information is not sufficient. With a window of [-19; 0] we get $C_{max} = 0.14$ indicating that some important information, besides the Cap signal, is present especially around [-19; -17] (two small peaks on the Shannon plot), but this conclusion is somewhat uncertain because a similar effect could not be observed on the curve with anchor point 35. Finally the C_{max} again increases when the TATA-box region is included.

With anchor point -35 (windows [-35; -35], [-35; -34], ..., [-35; 35]) we change the window in the direction left to right. Including only the TATA-box region [-35; -26] the network produced $C_{max} = 0.20$. The next increase happens, not surprisingly, when the transcription initiation site is included $C_{max} = 0.26$ and culminating with the window [-35; 8] ($C_{max} = 0.29$). With a larger

Figure 3. The maximum correlation coefficient C_{max} for the test set as function of the position in the reduced window. The number of neurons in the intermediate layer was 15. The curve with anchor point 35 represents results obtained by varying the window in the order [-35; 35], [-34; 35], ..., [35; 35]. The maximum correlation coefficient for the network trained with window [10; 35] is found at position 10 on the x-axis. The curves with anchor point 0 and -35 were obtained in the same manner.

window the generalization is getting worse indicating that the extra information is blurring the picture.

Summarizing the results from Figure 3, we conclude that the three most important regions are the TATA-box [-29; -24], the Cap signal [-1; 0] and a region around [8; 10]. These observations were also confirmed by the Shannon plot except that the [8; 10] region appears to be much more important than indicated by the Shannon plot. Furthermore some information also seems to be present around [-19; -17].

33;50	-----g-----c-----
33;50	*AG-Ca*TG-C-AG-ca-TGg-C--cAG-Ca-Tga-tg-CA--Gatg-CAG-Gca-tG-Ct-AGcaG-c-Ca ² G*
40;60	-aG-c--TG-C-Ag-ca-TG--C--cAG-c--Tga-tg-Ca--G----Cag-Gca-tG-C--ag-ag-c-C-G-
50;70	-aG-c--tG-c-ag----tG--c--aG-c--T---g-Ca--G--Ca-g--G-C--ag--g--C-G-
60;80	-g----tg-c-a----tg--c--ag---t-----c--G--c--g--g-c-----c-G-
	-90 -80 -70 -60 -50 -40 -30
33;50	---c-----c--g-----tt ² AAAA--gg-g--c-g--g--gg--c-CA-C--c-c-----
33;50	tCaG*ag-CtG-g*cTATAaAAGg-CaG-g-C- ² aG-CaG-Ct-aGcaGCT-CAGag-CT-gCt-C ² G*Ca-CAG*C
40;60	tCaG-a--ctG--*cTATAaAAG--CaG-g-C--aG-c-G-Ct-aG--gCt-CA-ag-Ct--Ct-CaG-ca-caG*c
50;70	-c-G-----G---TATA-aAG--c-G--C--G-c-g-C--G--gct-CA---Ct--C--c-G-c--cag-c
60;80	-c-g-----g---taTA-aAG--c-g--c--g--g-c--G--c--Ca---c--C--c-g-----g--
	-50 -40 -30 -20 -10 0 10 20
33;50	--g--gg--c-cA---c-c-----c-----c-----c-----c-----
33;50	C ² *AGct*C--CAG ² C ² *CTgC--AG-C-gCAGctC*-tGAGCT-a-CtG-CAGct*CatgCaG-;CA ² *
40;60	C**aGct*C--CAG-Ct--CT-c--aG-C--cAG-t*C--Gagct-a-Ctg-cAGct-C-tgcaG--Cag*
50;70	C-*aG--C--CAG-c--CT-c--aG-C--cAG-t-C--Gagct--C-g-cag-t-C--c-G--ca--*
60;80	c--g--C--cA--c--Ct-----g-c--ag--C--g--t--c--ag--C--g--c--
	-10 0 10 20 30 40 50 60

Figure 4. The consensus sequences of the best alignments in three regions shown with different cutoffs. For example '33; 50' means that the upper case letter has >50% occurrence, lower case >33% occurrence. The '*' represents a gap sign and '-' represents positions where no symbol occur more than the lower limit. In each section the top row corresponds to the unaligned sequences.

To estimate the limit of the prediction accuracy, four network configurations using windows of [-35; 35], [-35; 15], [-29; -24 and -1; 0] and [-29; -24] (intermediate units from 0 to 100) were analyzed further. The four configurations represent regions with all information, all important information, information presented at the TATA-box region and Cap signal, and the TATA-box region only. These networks had with a few intermediate units performances in the range 0.18–0.22, 0.20–0.25, 0.25–0.32 and 0.00–0.15, respectively. The best network had a [-29; -24 and -1; 0] window, 60 intermediate units, and gave $C_{max} = 0.32$ (false positive rate 0.1%). By changing the threshold of the output neuron the ratio between true and false positives can be varied, e.g. with a threshold of 0.31 we obtained a correct recognition of 27.2% true sites (1.0% false positives), while a cutoff of 0.18 gave 44.6% with a false positive rate of 4.9%.

The result emphasizes the concerted importance and interplay of the Cap signal and the TATA-box in the recognition of the transcription initiation site. Important information is present in small regions around +10 and possibly also at -18, but the network is not able to benefit from it, because most of the extra information is uncorrelated with the transcription initiation site location.

Multiple alignment

Several runs were made on each of the three data sets covering the positions [-100; -30], [-50; 20], [-10; 50]. The best alignment was obtained after the longest computational time (>14 CPU-days on an hp-735 workstation). The consensus sequences from the best alignments of the merged mammalian data set are shown in Figure 4.

Comparing with the unaligned consensus sequences given in the top row for each alignment the resulting consensus is clearly improved. As expected, the results obtained from the data set [-100; -30] did not show strong similarities. The most interesting seems to be the repeated occurrence of dinucleotides TG and AG,

and a very strong G at position -47 (33.4% in the unaligned data set). The alignment of the data sets [-50; 20] and [-10; 50] however gave interesting results. Clearly visible are the known consensus sequences for the TATA-box TATAaAAG and the Cap signal CA. The GC-rich areas surrounding the TATA-box can also be identified. Furthermore a strong consensus CTNCNNAGNC (the new CT-signal) occurs at position +10. Note that in the alignment of the data set [-50; 20] the CT-signal is split into two signals, while the [-10; 50] alignment discover that it is in fact the same signal. Another interesting consensus is the frequently occurring triplet CAG.

Further analysis of the promoter signals

Distance distributions. The aligned data sets were used to obtain the distance distribution between important features. It was made by counting the number of non-gap symbols between the regions in the aligned data set. To find the distance distribution between the TATA-box and the Cap signal the [-50; 20] alignment was used, while the [-10; 50] alignment was used for distances between the Cap and CT-signal. The distance distribution between the TATA-box and the CT-signal was found by concatenating each sequence in the two alignments. This was done because the TATA-box is so strong a signal that the CT-signal is nearly destroyed in the [-50; 20] alignment. The second T in the TATA-box (strongest consensus letter), the A at the transcription initiation site and the first C in the CT-signal was used as markers. Results are shown in Figure 5. The distance distribution between the TATA-box and the Cap signal is more narrow than between the TATA-box and the CT-signal. This is in agreement with the CT and Cap-signal distance distribution, showing that even though the signals are close to each other, the distribution is quite broad.

Complementarity of signals. Trying to uncover the interdependence of the CT-signal and the TATA-box, weight matrix

Figure 5. Distribution of distances (in nucleotides) between the three core promoter elements, the TATA-box, the Cap site and the CT-signal found in the promoter region.

Figure 6. The normalized weight matrix matching score for the TATA-box and the CT-signal in 481 promoter sequences plotted against each other. Note that there are no instances where a low TATA-signal is coupled to a low CT-signal.

calculations for the two signals were done using the concatenated aligned data set. The weak *a* in the TATA-box consensus was not included in the calculations. The matching score for the two signals was calculated in each of the 481 aligned sequences by summing the nucleotide frequency for the nucleotides occurring at the consensus positions (see equation 3). The matching score was normalized by the maximum obtainable matching score. In Figure 6 the matching score for the CT-signal and TATA-box is plotted against each other. Notice that a low matching score of one signal always gives a significant score of the other signal. This means that at least one of the signals is present. The bands seen in the plot correspond to the number of symbol deviations from the consensus. In the EPD database the mammalian promoters are categorized in eight different main groups. To check whether the strength of the CT-signal is related to promoter type, the matching scores for each of the eight categories were investigated. No bias towards one particular category was found (data not shown).

Consensus of the CT-signal. The above analysis which is based on the multiply aligned sequences, should be complemented by

Figure 7. Average normalized maximum matching score for the TATA-box, Cap and CT-signal as function of the position relative to the Cap site. Note that the score for the three curves are not directly comparable since different freedom regions were used.

analysis of the original mammalian promoter sequences. To find the real consensus of the interesting signals the aligned data set was used. The new consensus of the signals gave (with upper and lower case thresholds of 50 and 33%) for the TATA-box TATAAA(g/a)G [78, 69, 67, 77, 65, 71, (41, 39) and 51%], the Cap-signal CAG (73, 66 and 42%) and the CT-signal CTgc-ggc (82, 65, 37, 41, 38, 36 and 34%). The most significant consensus and corresponding optimal freedom region was found as follows: for each signal we started with the consensus above and step by step removed the letter with the lowest frequency. Plots of the average normalized maximum matching score of the different signals with variation of the freedom region were inspected, and they showed the following new consensus sequences, with indication of the freedom region and the preferred region, see Figure 7. For the TATA-box (TATAAA, 11, [-32, -22]) with maximum occurrence at position -29; for the Cap-signal (CA, 2, [-1, 0]) and the CT-signal (CTNCNG, 7, [6, 12]) with maximum occurrence at position 7. Investigation of the other strong consensus signals obtained from the alignments was also done, but added nothing new.

The analysis shows that the CT-signal is significant. The preferred region is close to the transcription initiation site and indicates that the signal could be a domain in the initiator binding element. Searching the Transcription Factors Database (TFD) (20,41) for CT-signals did not reveal any known signals downstream of the transcription initiation site. It has been reported that the HIV-1 promoter has a CTCTGG signal starting at position six (our numbering), which when mutated at position 10 and 11 (the two G's) weakens the promoter activity (42).

DISCUSSION

The EPD data based study revealed new significant details about the structure of eukaryotic promoters. Statistics of nucleotide occurrences for the vertebrate data set divisions around the TATA-box, the Cap signal and the C and G rich regions showed differences in relation to invertebrates and plants, where the C and G rich regions were not present—instead the promoter region consensus of plants was very A and T rich. In the virus promoter region the Cap signal was undetectable, indicating that the viral

initiator elements are much more diverse than what is found in the other data divisions.

Mammalian promoter regions were investigated further by neural networks and multiple alignment. Especially multiple alignment proved to be a very strong method in uncovering the consensus of 'small' regions in the general transcription machinery.

Neural networks were used to predict the exact location of the transcription initiation site in a data set consisting of 481 promoter regions from mammals. The best result was 27.2 and 44.6% recognition of true sites with 1.0 and 4.9% false positive rate, respectively. Papers dealing with prediction of eukaryotic promoters or transcription initiation sites have not been found making a comparison of results hard.

Investigation of the performance of different network configurations and results obtained by the Shannon information measure showed three important regions: the TATA-box region [-30; -24], the Cap signal [-1; 0] and a new region [+5; +10]. The method of multiple alignment uncovered the consensus of the new region termed the CT-signal CTNCNG.

Only few papers dealing with sequence analysis of eukaryotic promoters have been published, but this is due to the fact that the data quantity is low compared with the complexity of the biological mechanism. With the exponential growth of available DNA sequences including eukaryotic promoters regions, sequence analysis will have a high potential for providing a better understanding of the complex mechanism that controls the transcription process.

ACKNOWLEDGEMENTS

We thank Anders Gorm Pedersen and Olle Karlström for comments on the manuscript. This work was supported by the Danish Natural Science Research Council and the Danish National Research Foundation.

REFERENCES

- Weis, L. and Reinberg, D. (1992) *FASEB* **6**, 3300-3309.
- Travers, A. (1993) *DNA-Protein Interactions*, Chapman & Hall.
- Green, M. R. (1992) *Nature* **357**, 364-365.
- Johnson, P. F. and McKnight, S. L. (1989) *Ann. Rev. Biochem.* **58**, 799-839.
- Killeen, M., Coulombe, B., and Greenblatt, J. (1992) *J. Biol. Chem.* **267**, 9463-9466.
- Zenzie-Gregory, B., O'shea-Greenfield, A., and Smale, S. T. (1992) *J. Biol. Chem.* **267**, 2823-2830.
- Conaway, R. C. and Conaway, J. W. (1993) *Ann. Rev. Biochem.* **62**, 161-190.
- Penotti, F. E. (1990) *J. Mol. Biol.* **213**, 37-52.
- Bucher, P. (1990) *J. Mol. Biol.* **212**, 563-578.
- O'shea-Greenfield, A. and Smale, S. T. (1992) *J. Biol. Chem.* **267**, 1391-1402.
- Horikoshi, M., C., B., Takada, R., Wang, J., Yamamoto, T., and Roeder, R. G. (1992) *Proc. Natl. Acad. Sci. USA* **89**, 1060-1064.
- Javahery, R., Khachi, A., Lo, K., Zenzie-Gregory, B., and Smale, S. T. (1994) *Mol. Cell. Biol.* **14**, 116-27.
- Zenzie-Gregory, B., Khachi, A., Garraway, I. P., and Smale, S. T. (1993) *Mol. Cell. Biol.* **13**, 3841-3849.
- Nussinov, R. (1992) *J. Biomol. Struct. Dyn.* **9**, 1213-1237.
- Christensen, N. C. S. and Larsen, N. I. (1992) In *Proceedings of the first Danish Conference on Pattern Recognition and Image Analysis*: Department of Computer Science, University of Copenhagen.
- Trifonov, E. N. and Brendel, V. (1987) *GNOMIC: A Dictionary of Genetic Codes*, Volker Brendel, Weinheim, Deerfield Beach, FL.
- Nussinov, R. (1987) *DNA* **6**, 13-22.
- Nussinov, R. (1992) *J. Theor. Biol.* **155**, 243-270.
- Lukashin, A. V., Engelbrecht, J., and Brunak, S. (1992) *Nucleic Acids Res.* **20**, 2511-2516.
- Pesole, G., Prunella, N., Luini, S., Atimonelli, M., and Saccone, C. (1992) *Nucleic Acids Res.* **20**, 2871-2875.
- Prestridge, D. S. (1991) *Comput. Appl. Biosci.* **7**, 203-206.
- Prestridge, D. S. and Burks, C. (1993) *Hum. Mol. Genet.* **2**, 1449-1453.
- Schneider, T. D., Stormo, G. D., Gold, L., and Ehrenfeucht, A. (1986) *J. Mol. Biol.* **188**, 415-431.
- Schneider, T. D. (1988) In *Maximum-Entropy and Bayesian Methods in Science and Engineering* Vol 2, Dordrecht, The Netherlands: Kluwer Academic Publishers. pp. 147-154.
- Hirst, J. D. and Sternberg, J. E. (1992) *Biochemistry* **31**, 7211-7218.
- O'Neill, M. C. (1991) *Nucleic Acids Res.* **19**, 313-318.
- Brunak, S., Engelbrecht, J., and Knudsen, S. (1991) *J. Mol. Biol.* **220**, 49-65.
- Demeler, B. and Zhou, G. (1991) *Nucleic Acids Res.* **19**, 1593-1599.
- Horton, P. B. and Kanehisa, M. (1992) *Nucleic Acids Res.* **20**, 4331-4338.
- Lukashin, A. V., Anshelevich, V. V., Amirikyan, B. R., Gragerov, A. I., and Frank-Kamenetskii, M. D. (1989) *J. Biomol. Struct. Dyn.* **6**, 1123-1133.
- O'Neill, M. C. (1992) *Nucleic Acids Res.* **20**, 3471-3477.
- Bucher, P. *The Eukaryotic Promoter Database* EPD EMBL Nucleotide Sequence Data Library Release 33, Postfach 10.2209, D-6900 Heidelberg (1992).
- EMBL Data Library Nucleotide Sequence Database*, Release 33 European Molecular Biology Laboratory, Postfach 10.2209, D-6900 Heidelberg (1992).
- Shannon, C. E. (1948) *Bell System Tech. J.* **27**, 379-423, 623-656.
- Engelbrecht, J., Knudsen, S., and Brunak, S. (1992) *J. Mol. Biol.* **227**, 108-113.
- Grassberger, P. (1988) *Phys. Lett. A* **128**, 369-373.
- Hertz, J., Krogh, A., and Palmer, R. G. (1991) *Introduction to the Theory of Neural Computation, Studies in the Sciences of Complexity*, Addison-Wesley Publishing Company.
- Mathews, B. W. (1975) *Biochim. Biophys. Acta* **405**, 442-451.
- Kirkpatrick, S., Gelatt, C., and Vecchi, M. (1983) *Science* **220**(4589), 671-680.
- Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A., and Teller, E. (1953) *J. Chem. Phys.* **21**, 1087-1092.
- Ghosh, D. (1993) *Nucleic Acids Res.* **21**, 3117-3118.
- Sheldon, M., Ratnasabapathy, R., and Hernandez, N. (1993) *Mol. Cell. Biol.* **13**, 1251-1263.