A DISCRIMINATION STUDY OF HUMAN CORE-PROMOTERS

MICHAEL Q. ZHANG

Cold Spring Harbor Laboratory, 1 Bungtown Road, Cold Spring Harbor, NY 11724

A core-promoter, approximately from -60 bp upstream to +40 bp downstream of a RNA polymerase (RNAP) II transcription start site (TSS), binds to the preinitiation complex (PIC) and determine the position of TSS. Using position-specific k-tuple feature variables, a quadratic discriminant analysis (QDA) method is shown to be very effective in identifying human core-promoters.

1 Introduction

It is no secret that computational identification of eukaryotic RNAP II promoters is notoriously difficult. The field is still in its infancy and current algorithms are quite primitive (reviewed in Fickett & Hatzigeorgiou 1997). This is mainly due to our limited understanding about underlying molecular recognition mechanism of transcription initiation (*e.g*. Kornberg 1996; Nikolov & Burley 1997).

Recent advances in molecular genetics, biochemistry and structural biology have shown that (1) Promoter has a modular structure consisting of multiple short sequence elements, mostly transcription factor (TF) binding-sites. They can be dispersed or overlapped, largely populating in about 1 Kb region upstream and surrounding TSS. They can be either positive or negative and their functions are often context-dependent. Most of the distal elements are activational or regulatory, and their pattern of organization is often gene or pathway specific. (2) Core-promoters consist of minimal DNA elements that are necessary and sufficient for accurate transcription initiation in reconstituted cell-free systems. The fact, that it has most of the constitutive activity, it can drive heterologous gene transcription and its binding partner - PIC is made up basal TFs (most are universal), implies core-promoter may contain all the universal and positional elements (Bucher 1990). (3) Transcription initiation is hierarchical and dynamic. It starts from chromosomal derepression (through chromatin remodeling and nucleosome disruption) and TF (including PIC) binding. It results in activation of core-promoter via multitude interaction with network of TFs. Physically, chromosomal derepression is necessarily the first step during which CpG-island and chromatin structure should all play very important roles. It is entirely possible that a stable PIC-binding (i.e. core-promoter recognition) may require some other distally bound TFs in order to create a favorable environment (to lower the free energy barrier). Based on the biological information, I propose a 2-step approach to the computational promoter recognition and TSS mapping problem: given a large (human) genomic DNA sequence of size 100 Kb ~ 1 Mb, Step1 is to localize the general promoter approximately in about 1 ~ 1.5 Kb region; Step2 is to further "zoom-in" core-promoter/TSS down to about 100 bp region. This not only reduces the complexity for each individual step, but also can cope with the possibility that different signals may be used on different scales. Furthermore the result from each step can also have its own applications. For example, the result from Step1 will be sufficient for separating individual genes in a multi-gene sequence. Most often, bench scientists are able to subclone a promoter active region, the result of Step2 will therefore be sufficient for further fine TSS-

mapping. Here I shall introduce a new algorithm for Step2 analysis and describe my initial attempt to discriminating human core-promoters while mapping the TSSs.

2 Core-promoter classes and organization

Before going into computational aspects, I want to briefly summarize what's known about core-promoters (for more detailed reviews, see *e.g.* Kollmar & Farnham 1993; Orphanides *et al*. 1996; Tjian 1996; Roeder 1996).

As shown in the cartoon (Fig.1), TATA-box and Initiator (Inr) are the two key genetic elements in a core-promoter which play a central role in determining the TSS position (*e.g.* Novina & Roy 1996). TATA-box has the TATA(A/T)A(A/T) consensus and Inr has the YYAN(T/A)YY consensus (the underlined position indicates the TSS). They are functionally

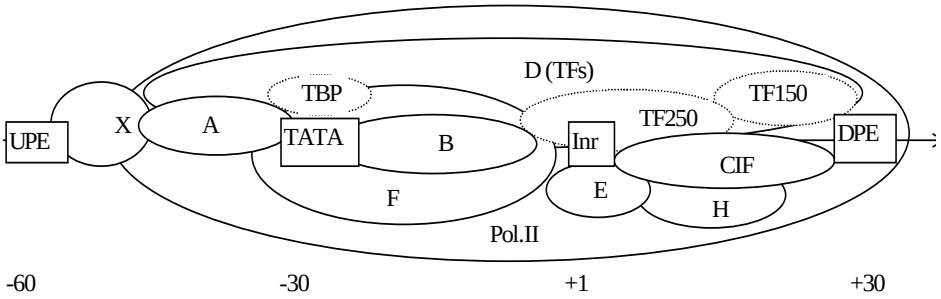


Figure 1 Core-promoter interaction with PIC, X (UPE-binding TF) and CIF (Co-initiator TF)

similar in two respects: both direct accurate transcription initiation by RNAP II in the absence of other control elements, and both direct a high level of accurately-initiated transcription when stimulated by an upstream activator (Smale 1997). Abundantly expressed genes (mostly cloned by 1980) often contain a strong TATA-box in their core-promoter. Housekeeping genes, several oncogenes, growth factors (GFs) and TFs are usually TATA-less. $TATA^{-}Inr^{+}$ promoters are mainly found in hematopoetic lineage-specific genes and homeotic genes; $TATA^{-}Inr^{-}$ promoters mainly found in housekeeping genes that have multiple TSS (often 40-80 downstream of a Sp1 site and some share a DPE (Downstream Promoter Element) called MED-1 with the GCTCC(G/C) consensus, Ince & Scotto 1995). Due to its overlap with other TF sites, Inr has much weaker consensus comparing to TATA-box. Tab.1 shows some mapped examples.

Gene	TATA-box	Inr	TF
AdML	TATAAAA	TCACTCT	II-I at (+7,+33)
gfa	CATAAAG	weak	II-D at (+10,+50)
hsp70	TATAAAT	weak	II-D at (+18,+30)
TdT	CTGCTGGTC	TCATTCT	II-I, YY1
dhfr	-	CAAACTT	E2F
PBGD	-	TCAGTGT	? at (+3,+12)
rpS16	-	TCCCTTT	YY1
P5	-	CCATTTT	YY1

Table 1
Examples of TATA and TATA-less promoters.

Transcription initiation involves assembly of PIC (Fig.1) on core-promoter. Although detailed structural information is still lacking, recent site-specific protein-DNA photocrosslinking done on a human TBP-IIB-IIF-RNAPII-core_promoter subcomplex (which is stable and fully competent for transcription under DNA melting condition) had revealed (Fig.2) that the interface between the largest and second-largest subunits of RNAPII (RPB1and RPB2) forms an extended (~240 Å) channel that interacts with core-promoter both upstream and downstream of TSS. It was also shown that RNAPII can compacts core-promoter DNA by the equivalent of ~50 bp (Kim *et al*. 1997; see also Forget *et al*. 1997).

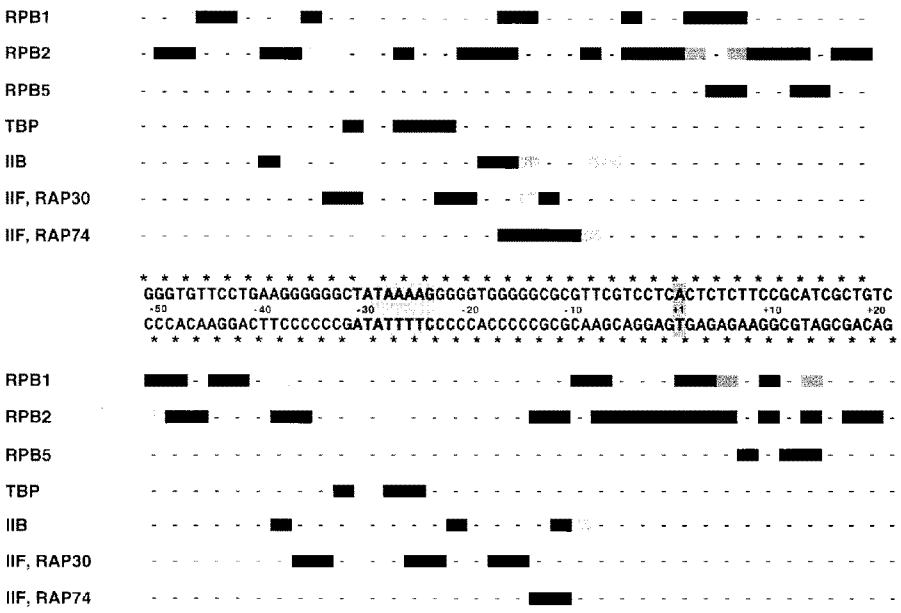


Figure 2 Summary of crosslinking data (Kim *et al*. 1997). Phosphates analyzed are indicated by asterisks. Sites of strong and weak crosslinking are indicated by solid and shaded bars, respectively.

3 Data and Methods

177 human non-redundant promoter sequences were extracted from EPD48 (Bucher and Trifonov 1986). Each sequence was then extended from the original range (-500,+100) to (-600,+600) by BLASTing GenBank (release 100). A few corrections were made after checking against both the original and recent publications.

The standard linear and quadratic discriminant analyses (LDA/QDA, see *e.g*. Zhang 1997 and the references therein) were used for core-promoter discrimination. All feature variables were 5-tuple scores averaged within a position-specific window. If one defines $f_w(s)$ to be the

signal frequency of a *k*-tuple *s* in the window *w* and $f_b(s)$ to be the background frequency calculated as the average of $f_L(s)$ and $f_R(s)$, where "*L*" and "*R*" indicate the left and the right nearest-neighbor non-overlapping windows, then the *k*-tuple score $x(s) = f_w(s)/(f_w(s)+f_b(s))$. All the f_w's were estimated from the aligned data and Bayesian priors were used to render all frequencies nonzero (Tanner & Wong 1987).

In the exploratory LDA studies, each sample was a sequence of length 120 bp which contained 4 non-overlapping windows of size 30 bp each (Fig.3). Samples were drawn from 177 EPD48 non-redundant human sequences (-500,+100) at a 30 bp interval. Each sequence would contain just one true sample (ignoring the few multi-TSSs) at (-70,+50).

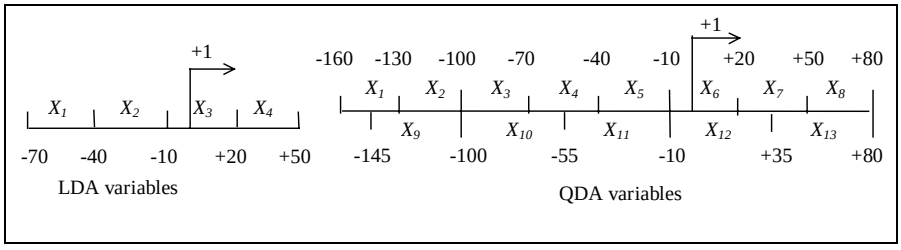


Figure 3 Feature variables in discriminant analyses.

In the QDA study, each sample was a sequence of length 240 bp, which contained two sets of windows of size 30 bp or 45 bp each. As shown in Fig.3, there were 13 5-tuple feature variables. Samples were drawn from 177 extended EPD48 sequences (-600,+600) at a 6 bp interval. Again each sequence was considered to contain just one true sample at (-160,+80).

4 Results

4.1 Statistical properties

I started investigation by examining statistical characteristics of human promoter sequences in EPD. I first looked for positional elements by ranking common 6-tuple frequencies. As shown in Fig4., in a (-100,100) window, TATAAA is the only recognizable top-ranking 6-tuple clustered around -30. Other high frequency 6-tuples resemble Sp1 site scattered all over in the region. The same was true for high-scoring putative TF sites. Fig.5 shows a representative plot of TF sites in 10 promoters, these sites were identified by using IMD (Chen *et al*. 1995) with a 0.96 score cut-off. This indicates that the current EPD may be biased by TATA promoters. Although it is known that the density of putative TF sites tends to be higher in general promoter region (this was actually the basis for many promoter prediction algorithms), it is certainly insufficient as a core-promoter indicator.

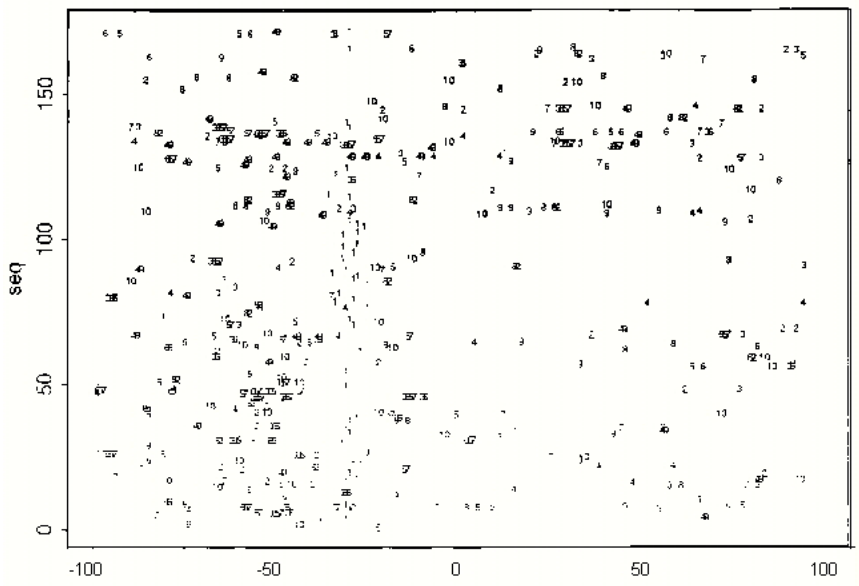


Figure 4 Scatter plot of top 10 most frequent commom 6-tuples in each promoter. The ranking is : 1-TATAAA, 2-GGAGGG, 3-GGGGCG, 4-CCCGCC, 5-GGCGGG, 6-GGGCGG, 7-GCGGGG, 8-AGGAGG, 9-CCGCCC and 10-GGGCAG.

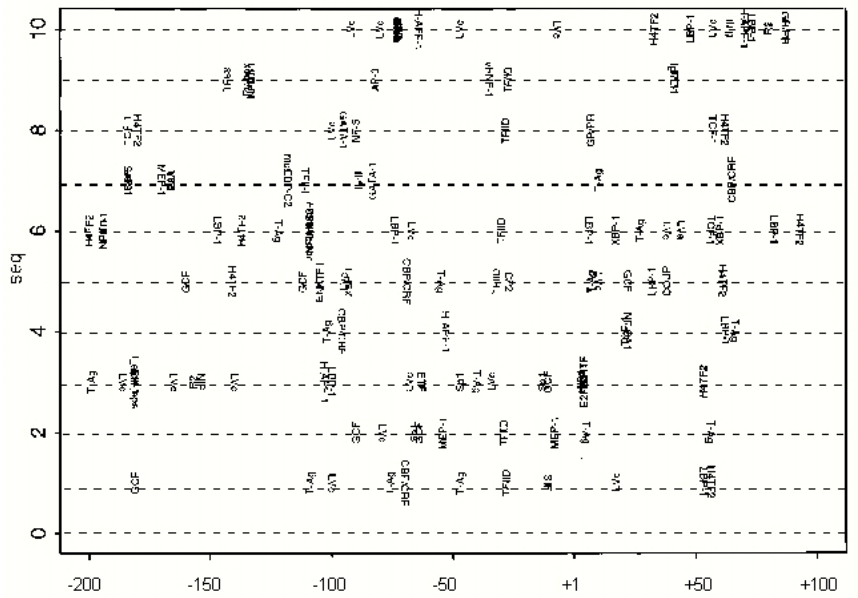


Figure 5 Putative TF-sites (predicted by IMD with cutoff=0.96) in 10 promoters.

Figure 6 Scatter plot of TATA-score, Inr-score and Maximum CpG-density per 10 bp.

Next, I examined possible correlation among TATA, Inr and the maximum CpG-density segment in the core region. From Fig.6, only TATA and CpG-density scores appeared to be correlated, which may be simply related to more GC-boxes in TATA-less promoters. But the correlation is too weak to be useful as a core-promoter discriminant.

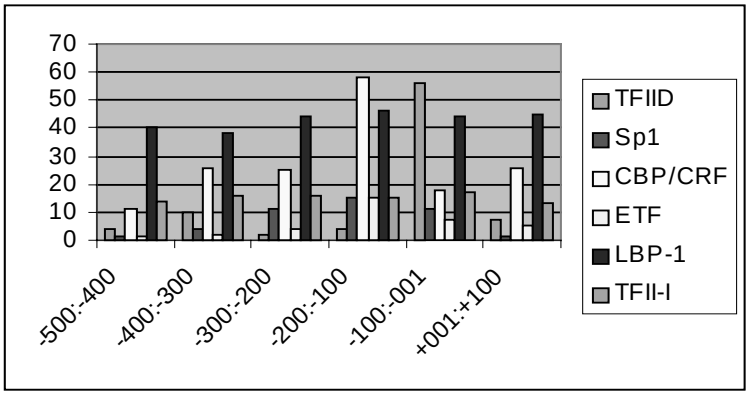


Figure 7 Histograms of putative TF-sites in EPD48 promoters (as predicted by IMD at 0.96 cut-off.)

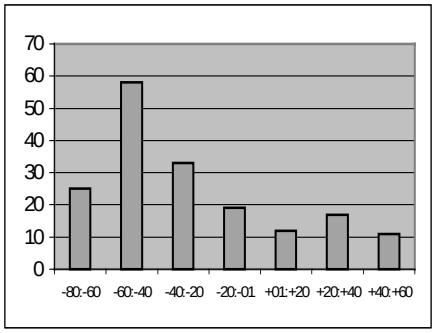
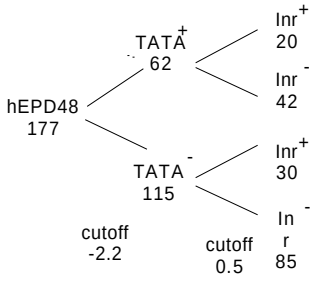


Figure 8 Histogram of putative Sp1 core-sites in $TATA^{-}Inr^{-}$ promoters.



Quantile distribution of TATA and Inr scores

	Min	Q1	Med	Mean	Q3	Max
TATA	-22.19	-8.18	-3.49	-5.43	-1.5	0.0
Inr	-2.76	-0.72	-0.03	-0.09	-0.56	2.0

Are there other position-specific TF sites than TATA and Inr? I have done the histogram for every major TF. Fig.7 shows some typical examples. Most TF-sites are position-nonspecific, such as LBP-1 and TFII-I. Some (Sp1, CBP/CRF, ETF) have a position preference, but are still quite variable. Interestingly, I found the positional bias could be enhanced if the promoters were subclassified. For example, if I used Bucher's TATA score (Bucher 1990) with a -2.2 cut-off, I could divide 177 promoters into $TATA^{+}$ (62) and $TATA^{-}$ (115) classes; if then used Inr-scoring scheme of Kraus *et al*. (1996) with a 0.5 cut-off, these classes could be further subclassified as $TATA^{+}Inr^{+}$ (20), $TATA^{+}Inr^{-}$ (42), $TATA^{-}Inr^{+}$ (30) and $TATA^{-}Inr^{-}$ (85). Sp1 sites could be better clustered around -50 in the $TATA^{-}Inr^{-}$ subclass (Fig.8), which is consistent with experimental findings (Smale 1997). Even though, this classification was scoring and cut-off dependent, but the result was not sensitive to either (data not shown). The point is that a more sensible position-specific TF-site study would require biological relevant classification of the promoters. Unfortunately, there was not enough data for more quantitative analysis.

4.2 Exploratory LDA studies

As methods based on putative TF-sites have severe limitations (such as: important context effect may be overlooked; majority of TF-sites are false positives; they are scoring function and cut-off dependent; they may be biased by the limited mapped TFs; *etc*.), I preferred to start pursuing an objective statistical approach without having to use any putative TF-site information. 6-tuple frequency method had been used for promoter prediction as a "content" measure in the sense of Staten (Hutchinson 1996). But this "content" approach would loss all the positional information which is crucial for fine-mapping of core-promoter/TSS. On the other hand, a pure "signal" approach would be powerless because of the large variation in signal positions. This suggests a "mixed" approach by using position-specific windows. Here I chose the average *k*-tuple frequency score *x* in a window of size *w* as the feature variable (see **Data and Methods**). Then there was a choice of *k* and *w*. Reliable

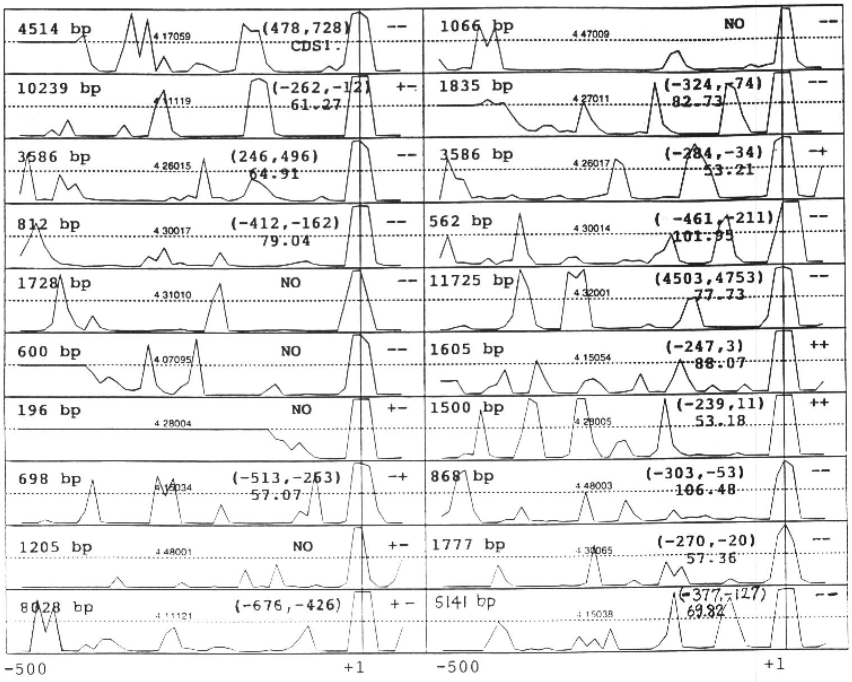


Figure 9 LDA profiles for 20 EPD48 human promoter sequences. The original GenBank sequence size, EPD entry-ID, PROSCAN V1.7 (cutoff=53) prediction (promoter region and its score) and core-promoter class are indicated in order.

statistics required $N*(w-k+1)$ be larger than 4^k, where $N=177$ was the number of independent sequences. Since w would determine the resolution and it needed to be large enough to contain major TF elements but small enough to capture the positional variation of the TF-sites. I found $w=30$ bp worked well. Although $k=6$ might be barely workable, I chose $k=5$ to be on the safe side and it represents a half-turn distance of a DNA double helix which often corresponds to the core-binding site of a typical TF (or a half-core in the case of a dyad). For an exploratory test, I chose *4* non-overlapping windows (hence *4* feature variables) and did various LDA studies by varying different parameters (w, k, sample interval or adding other feature variables). Fig.9 shows a typical LDA discriminant score profile for 20 promoters. A vertical line indicates the true TSS position. The EPD entry-IDs are printed in the middle. "+-" stands for $TATA^{+}Inr^{-}$ and so forth (according to the core-promoter classification mentioned above). Although, there were still quite a bit of noices, true signals tended to have a distinct shape as well as their height. For comparison, I also extracted the original GenBank sequences (sequence sizes are indicated on the far left), ran PROSCAN V1.7 (Prestridge 1995) with the default cut-off score 53 and printed the predicted promoters with the scores after the EPD entry-ID (the coordinates were defined relative to the TSS). It is clear that PROSCAN missed more than 50%.

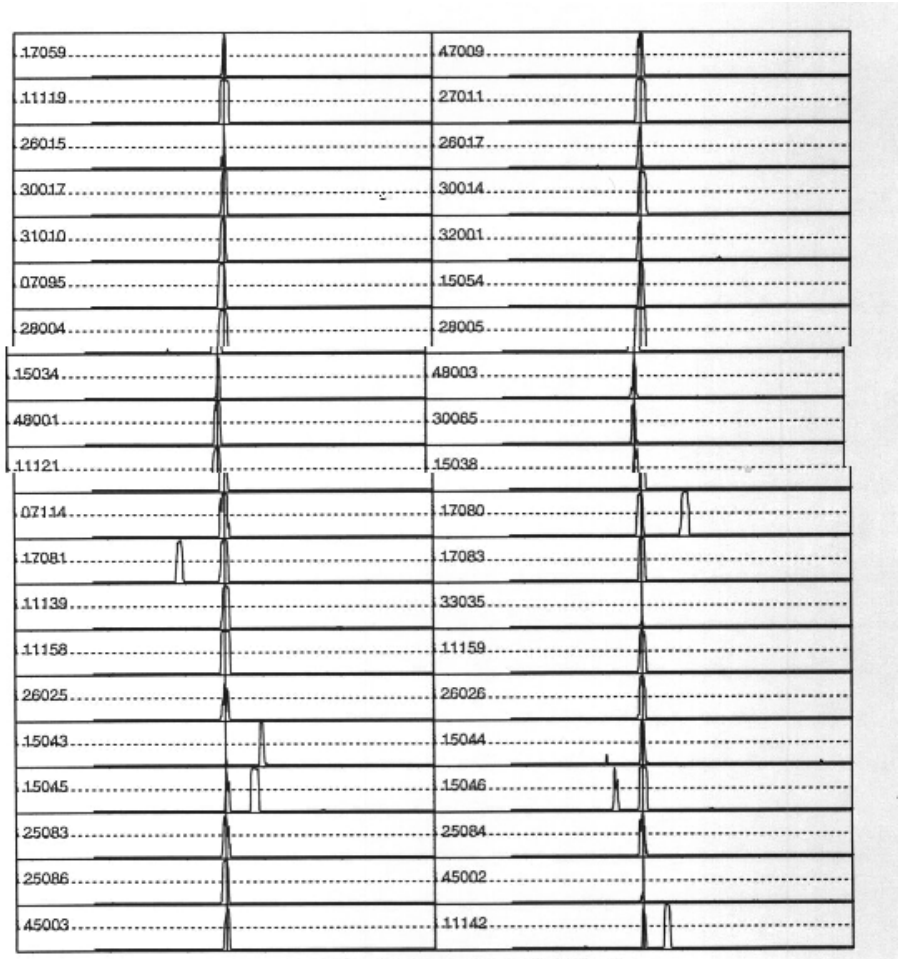


Figure 10. QDA profiles for 40 extended EPD48 human promoter sequences. EPD entry-ID is indicated for each sequence. The 20 sequences on the top correspond to the 20 sequences in Fig.9.

4.3 A QDA study

How could the noises be reduced? I found the false positives were much more sensitive to change of parameters. Especially, when varying the window size or the sampling interval, the true signals tended to remain at the same position while the noises tended to displace randomly (data not shown). This immediately suggested to apply the "principle of resonance": if two profiles corresponding to different parameters were combined, the true signals would tend to enhance each other while the noises would tend to cancel each other. Further more, as the sample size was increased, the height of the noises tended to be suppressed. In order to maintain a high resolution and to limit the dimension of the multivariate space, by trial-and-error, I found the 13-window system (8 windows of size 30 bp and 5 windows of size 45 bp) with sample size of 240 bp seemed to be optimal. Because the overlapping windows were used (see Fig.3), a more covariance-sensitive method — QDA had to be used (see *e.g.* Zhang 1997). Fig.10 shows the new discriminant profiles obtained by QDA of 13 feature variables (the sampling interval = 6 bp) on the extended EPD48 human promoters. The remarkable reduction of signal-to-noise ratio indicates the "interference amplification" at work.

	1	2	3	4	5	6	7	8	9	10
sn	0.857	0.771	0.629	0.571	0.657	0.771	0.657	0.686	0.629	0.857
sp	0.857	0.871	0.815	0.800	0.852	0.771	0.885	0.828	0.880	0.769

Table 2 Cross-validation statistics (sn - sensitivity; sp - specificity) (see *e.g.* Zhang 1997).

To further analyze the stability (reliability), standard 10 cross-validation tests were performed on the 177 true-samples and 42480 pseudo-samples (20% test-set and 80% training-set was drawn randomly in each test). The statistical variation may be seen from Tab.2. The average sensitivity and specificity were 0.71 and 0.83, respectively.

5. Final comments

I would like to make a few comments in order:

1. It is interesting that the double peaks in Fig.10 were actually corresponding to the alternative TSS as annotated in GenBank. Since, for simplicity, there was only one sample at position (-160,+80) per sequence considered as the true sample, alternative TSS and the high scoring samples in the neighborhood of the true sample were considered as false positives. There is still more room for further reducing the false positives in future improvement.
2. This QDA study involved only the position-specific 5-tuple frequency bias. Since the local background (as characterized by f_b) was used, chromosomal GC-content variation was therefore taken care of automatically. It was some what surprise that adding TATA, Inr or CpG-density scores as discriminant feature variables did not make any noticeable improvement (data not shown). Apparently they were also automatically "built-in" by the specific choice of windows.
3. Currently, this QDA algorithm of human core-promoter prediction has only been implemented in S-PLUS (StatSci. 1993) and is available upon request to the author.

Acknowledgement

This work was supported by NIH grant K01 HG00010-05.

References

Bucher, P. 1990. Weight matrix descriptions of four eukaryotic RNA polymerase II promoter elements derived from 502 unrelated promoter sequences. *J. Mol. Biol*., 212:563-578.

Bucher, P. and Trifonov, E. N. 1986. Compilation and analysis of eukaryotic POL II promoter sequences. *Nucl. Acid. Res.,* 14:10009-10026.

Chen, Q., Hertz G. Z. and Stormo, G. D. 1995 MATRIX SEARCH 1.0: a computer program that scans DNA sequences for transcriptional elements using a database of weight matrices. *CABIOS*, 11:563-566.

Fickett, J. W. and Hatzigeorgiou, A. G. 1997. Eukaryotic promoter recognition. Submitted.

Forget, D., Robert, F., Grondin, G., Burton, Z. F. and Greenblatt, J. 1997. *Proc. Natl. Acad. Sci. USA*, 94:7150-7155.

Hutchinson, G. B. 1996. The prediction of vertebrate promoter regions using differential hexamer frequency analysis. *CABIOS*, 12:391-398.

Ince, T. A. and Scotto, K. W. 1995. A conserved downstream element defines a new class of RNA polymerase II promoters. *J. Biol. Chem*., 270:30249-30252.

Kim, T.-K., Lagrange, T., Wang, Y.-H., Griffith, J. D., Ebright, R. and Reinberg, D. 1997. *Proc. Natl. Acad. Sci. USA*, submitted.

Kollmar, R. and Farnham, P. J. 1993. Site-specific initiation of transcription by RNA polymerase II. *Proc. Exp. Biol. Med*., 203:127-139.

Kornberg, R. D. 1996. RNA polymerase II transcription control. *TIBS*, 21:325-326.

Kraus, R. J., Murray, E. E., Wiley, S. R., Zink, N. M., Loritz, K., Celembiuk, G. W. and Mertz, J. E. 1996. Experimentally determined weight matrix definitions of the initiator and TBP binging site elements of promoters. Nucl. Acid. Res., 24:1531-1539.

Nikolov, D. B. and Burley, S. K. 1997. RNA polymerase II transcription initiation: A structural view. *Proc. Natl. Acad. Sci. USA*, 94:15-22.

Novina, C. D. and Roy, A. L. 1996. Core promoters and transcriptional control. *TIG*, 12:351-355.

Orphanides G., Thierry, L. and Reinberg, D. 1996. The general transcription factors of RNA polymerase II. *Gene & Dev.*, 10:2657-2683.

Prestridge, D. S. 1995. Prediction Pol II promoter sequences using transcription factor binding sites. *J. Mol. Biol*., 249:923-932.

Roeder R. G. 1996. The role of general initiation factors in transcription by RNA polymerase II. *TIBS*, 21:327-334.

Smale, S. T. 1997. Transcription initiation from TATA-less promoters within eukaryotic protein-coding genes, *Bioch. Biophy. Acta* 1351:73-88.

StatSci. 1993. S-PLUS User's Manual, V3.2, Seattle: StatSci, a division of MathSoft, Inc..

Tjian, R. 1996. The biochemistry of transcription in eukaryotes: a paradigm for multisubunit regulatory complexes. *Philos. Trans. R. Soc. Lond. B Biol. Sci*., 351:491-499.

Zhang, Q. M. 1997. Identification of protein coding regions in the human genome by quadratic discriminant analysis. *Proc. Natl. Acad. Sci*. USA, 94:565-568.