Automatic Extraction of Position Specific Cooccurrence of Transcription Factor Bindings on Promoters

Tatsuhiko TSUNODA[a] and Toshihisa TAKAGI
Genome DB, HGC, Institute of Medical Science, University of Tokyo, 4-6-1, Shirokanedai, Minatoku, Tokyo, 108, JAPAN

The difficulty of analyzing eukaryotic Pol II promoters originates in complex interaction between various transcription factors. Every precise investigation requires much labor. It is time consuming. However, promoters themselves conserve signs of such interaction. Each binding site will be position specific if we assume chemical interaction. Such traces can thus be stochastically extracted by aligning many promoters. Our new method, named *POSTSCRIPTER*, automatically identifies position specificity of each factor binding site, and also calculates their cooccurency for significance measuring. Applying these to 237 promoters, we extracted novel coincident patterns. They suggest unseen interaction, which we will discuss.

1 Introduction

Prokaryotic promoters simply regulate transcription. They need only several factors to align RNA polymerase II with transcription starting site(TSS) on DNA. A typical example suggests interaction between TFIID (basic transcript factor which contains TATA binding protein) and Sp1. TFIID binds to TATA box about 30bp upstream of the TSS. Sp1 binds to GC box about 100bp upstream of the TSS. They form a complex and stabilize the polymerase. We know such sequence patterns and inter-factor interaction well.

While, eukaryotic Pol II promoters use vairous regulation mechanisms. So many factors interact with each other that we can not examine all of them. Researchers only clarify inter-factor interaction they are interested in.

Moreover, in traditional method, we experiment on inter-factor interactions independent of promoter sequences. But such sequences will be good clue to catch them. Ideal method is to use sequences for prediction of factors boundable to each site and examine their interaction by experiments. However, investigation into every combination requires much labor and it will be time consuming. We must extract significant principles to decide the experimental reseach direction beforehand.

Here we first propose the method of automatical extraction of such pairs from many promoters. Each factor binding site selected is position specific relative to TSS on promoters. Our contribution is not only to collect information

[a]JSPS Research Associate of JSPS Research Project for the Future.
E-mail: **tatsu@ims.u-tokyo.ac.jp**

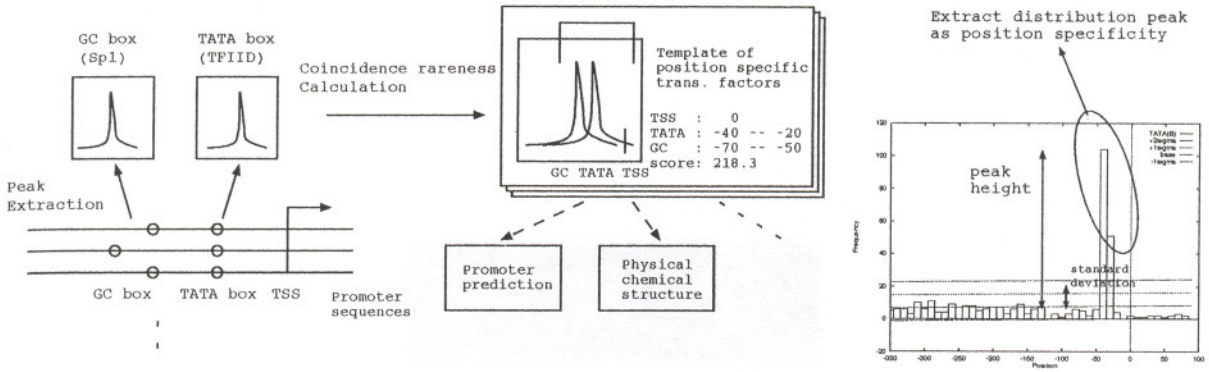

Figure 1: (a) Extraction of position specific transcription factor binding pairs by aligning many promoters on TSS. (b) Extraction of binding position peak of each transcription factor.

from biological experiments. Using many promoter sequences simultaneously, we can get statistical bias of the specificity which we can not see on single promoter. Such position specific binding pattern suggests a global structure dependency in transcription regulation mechanism. Moreover, by calculating, listing, and processing statistics of each factor binding site automatically, we can list up and give priority to all candidates without any artiticial mistakes. Such prediction will make biological experiments efficient and confirmable.

2 Position specificity and binding cooccurrence of factor pairs on DNA

We assume several transcription factors bind on DNA to form a complex. Each must bind to fixed position. We want to automatically extract significant combination of factors in each position from sequence data(Figure 1 (a)).

Our method, named *POSTSCRIPTER* (POsition-Specific Tran SCRIption factor binding Pair extracTER), is original in the following points:

1. Extraction of narrow position specificity of each factor binding site (Figure 1 (b)). Each position approximates physical three-dimensional position needed to form the complex.

2. Calculation of coocurrency (probability of coincidence relative to random) of factors with position specific binding sites.

For example, many promoters have TATA box about 30bp upstream of TSS. Because this box often appears at the position, researchers can easily find it. We guess other factors have such position specificity. Here we originally considered they can be detected as distribution peaks if we align many

promoters at TSS. Contrast to our general opinion, quite a lot of factor revealed to have position specific binding sites.

Next, let's take two factors. If we assume no principle, i.e. by random coincidence, both factors rarely bind to each specific position. While, if we assume some biological significancy in their combination and their location, e.g. TFIID's TATA box and Sp1's GC box on prokaryotic promoters, the cooccurrence probability must be more intentional than random coincidence. Assume chemical structure regulates binding between two factors and the distance is significant. If one factor has position specificity to TSS, the other factor will also have position specificity.

Such pairs can be detected by counting coincident pairs on a set of many promoters. Each coincidence bias is calculated by dividing the number of coincident instances by its random coincidence probability on non-promoters, according to which these pairs are ordered.

3 Method

3.1 Database

1. Transcription factor database : To specify which factor may bind to each position on a DNA sequence, we used transcription factor database TRANSFAC[6]. Using the weight matrix data, each factor is assigned score in each position. Each matching score is calculated according to Bucher's method[1]. We set the threshold score as 0.7, which is rather low, because we want to take many candidates for the peak detection.

   We also used weight matrices and thresholds of TATA box, GC box, CCAAT box, and Cap site shown in Bucher's article[1] in addition to TRANSFAC. Because TRANSFAC sometimes includes old data, we judged the data by Bucher is currently more reliable.

2. Promoter data : We used EPD[2] for promoter sequence investigation. Non-redundant sequences were taken out from the mammalian promoters. Among them, we finally used 237 promoters in which region of -349-+100bp to TSS is determined.

3. Non-promoter data : Sequence included in the list of non-promoter which Dr.Prestridge of the Minnesota university made[5] was divided by 450 bases. 963 sequences were thus made.

3.2 Detection of position specificity of each factor

Counting frequency of each factor which binds in each position

First, transcription factor binding sites on DNAs are specified. Here we assume TSSs are correct on promoters in EPD. Then the binding frequency of each factor $N(x_i)$ at each position x_i (every 10bp) is counted on DNAs. We define whole range as $A \equiv \{x_i \in A | -349 \leq x_i \leq +100, 1 \leq i \leq m\}$.

Peak extraction

The position on DNAs where each factor tends to bind is extracted by the following procedure(Figure 1 (b)):

(a) Calculation of distribution baseline : background frequency is approximated by a linear equation $N_{base}(x_i) = ax_i + b$ for $x_i \in A$. The parameters a and b were set to minimize the mean square error along the distribution.

(b) Calculation of frequency standard deviation: Standard deviation of each bin from the baseline is calculated.

$$S = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(N(x_i) - N_{base}(x_i))^2}$$

(c) Extraction of peak range: peak range $P \equiv \{x_i \in P | N(x_i) > N_{base}(x_i) + 2S\}$.

(d) Inclusion of surrounding of each peak: updated peak range $P \equiv \{x_i \in P | N(x_i) > N_{base}(x_i) + S,\ x_{i-1} \in P \text{ or } x_{i+1} \in P\}$.

(e) Re-calculation of the baseline: Each parameter of (a) is calculated again excluding instances within the peak range($A - P \rightarrow A$).

(f) Re-extraction of peak: (b) - (d) is processed again based on the recalculated baseline.

(g) Normalize each peak height: Each peak height (frequency) is normalized by the standard deviation.

$$H(x_i) \equiv \frac{N(x_i)}{S}, x_i \in P. \quad (1)$$

Thus extracted range($x_i \in P$, e.g. -20 – -40bp relative to TSS) is defined as the final peak range. This process is applied with every factor. Each factor binding distribution can have several peaks.

Table 1: (a) Variable description for calculation of cooccurrency. (b) Variable description.

(a)

E	Description of E	N_{1AB}	N_1	N_{2AB}	N_2
$CO_n(A, B)$	Bias of actual coincidence of factor A and B on promoters relative to their random coincidence on non-promoters	$N_p(A, B)$	N_p	$N_n(A)N_n(B)$	N_n^2
$CO_n(A_p, B)$	Bias of actual coincidence of factor A within its specific range and B on promoters relative to their random coincidence on non-promoters	$N_p(A_p, B)$	N_p	$N_n(A_p)N_n(B)$	N_n^2
$CO_n(A_p, B_p)$	Bias of actual coincidence of factor A and B within each specific range on promoters relative to their random coincidence on non-promoters	$N_p(A_p, B_p)$	N_p	$N_n(A_p)N_n(B_p)$	N_n^2

(b)

Variable	Description
N_p	The number of promoters
N_n	The number of non-promoters
$N_n(A)$	The average number of each factors which binds to 1 non-promoter
$N_p(A)$	The average number of each factors which binds to 1 promoter
$N_p(A, B)$	The number of factor pairs both appeared in coincidence on promoters
$N_p(A_p, B)$	The number of factor pairs both appeared in coincidence and one binds to its specific position (peak position extracted) on promoters
$N_p(A_p, B_p)$	The number of factor pairs both appeared in coincidence and both binds to each specific position (peak position extracted) on promoters
$P_n(A_p, B_p)$ $= N_n(A_p)N_n(B_p)/N_n^2$	Probability that a factor pair appeared in coincidence with relative position within each range

Calculation of cooccurrency

Next, *POSTSCRIPTER* counts the cases that a set of different factors binds simultaneously to each specific position (peak range extracted above) on every promoter. The frequency is divided by expected random coincidence rate on non-promoter sequences. We excluded cases when the two binding sites overlap with each other. Cooccurrency is calculated by the following equation:

$$E = \frac{N_{1AB}/N_1}{N_{2AB}/N_2}$$

Here, the content substituted for each variable depends on what we want to evaluate. We show the correspondance list on Table 1 (a). Each variable in this table is described in Table 1 (b).

Table 2: Information of each factor. Here, to distinguish from TRANSFAC TATA box, Bucher TATA box is written as TATA(B). CRE-BP1/c-Jun means hetero-dimer of CRE binding protein and c-Jun. Column '#peak binds' shows frequency of each factor which binds to site within each fixed range from TSS. Only top 20 are listed according to the peak frequency deviation H(max).

Factor name	#binds /pro.	#binds /non.	peak site from	peak site to	#peak binds	H (max)
TATA	0.25	0.16	-40	-20	26	6.25
MEF-2	0.22	0.34	-40	-30	18	5.92
TATA(B)	1.46	1.25	-40	-20	104	5.89
Oct-1	18.66	24.83	-40	-30	213	5.07
GC box	0.62	0.14	-70	-50	20	5.04
Barbie Box	0.14	0.12	-40	-20	5	4.27
CdxA	27.16	34.55	-40	-20	253	4.21
Pbx-1	1.92	2.87	-40	-30	24	3.96
Brn-2	0.16	0.20	-220	-210	5	3.84
Sox-5	0.48	0.73	-70	-60	10	3.74
Elk-1	0.32	0.27	-280	-270	7	3.74
ATF	0.14	0.04	-80	-20	5	3.71
CCAAT	0.41	0.19	-110	-60	10	3.49
NF-Y	0.30	0.12	-90	-60	8	3.31
c-Ets-1 p54	1.63	1.63	0	10	23	3.31
S8	0.15	0.21	-70	-60	4	3.30
CRE-BP1/c-Jun	0.18	0.12	-60	-40	5	3.20
Sp1	9.19	5.49	-100	-40	113	3.13
HFH-2	1.58	2.58	-140	-130	18	3.13
AP-4	0.25	0.25	-200	-190	6	3.12

4 Experimental Results

Tables from 2 to 4 shows the results. They are ordered according to E.

First, we show information of each factor in Table 2. Bias of the maximum frequency of the peak from the distribution ('H(max)', i.e., peak height normalized by standard deviation; calculated by eq.(1))are shown, which we use as ordering index of this table. The larger the value, the more significant peak we considered. Among 84 peaks(one factor may has several peaks) in all, we listed only top 20. Typical examples inculde TATA boxes(TRANSFAC and Bucher), Bucher GC box which unites Sp1, and Sp1 using TRANSFAC weight matrix. Their binding sites are consistent with the current experimental results. Column '#binds/pro' in this table illustrates average binding frequency of each factor on 1 promoter(450bp). While, '#binds/non' means average binding frequency of each factor on 1 non-promoter(450bp). These suggest factors which bind well on promoters tend to bind also on non-promoters well.

Next, Table 3 (a) and (b) shows examples of coincidence probability of two factor sites on promoters and non-promoters. GC box, ATF(CREB), CCAAT etc. cooccur with TATA box. This is in line with our general opinion. Column $P_n(A, B)$ shows random coincidence probability of factor A and B on one

Table 3: (a) Coincidence probability of factor pairs binding on promoters and non-promoters(top 10 according to $CO_n(A,B)$). '#of pairs' shows the total number of pairs which appeared to the set of promoters. '#of pro.' stands for the number of promoters which contain both factor A and B. (b) Coincidence probability of TATA box and other factor binding sites on promoters and non-promoters (top 10 according to $CO_n(A,B)$).

	Factor A	Factor B	$P_n(A,B)$	# of pairs	# of pro.	$CO_n(A,B)$
	NF-Y	GC box	0.0169	61	26	15.3
	GC box	ATF	0.0063	22	14	14.7
	Sp1	GC box	0.7753	1959	92	10.7
	CCAAT	GC box	0.0274	66	33	10.2
(a)	GC box	CREB	0.0785	129	36	6.9
	GC box	Elk-1	0.0387	59	29	6.4
	CCAAT	NF-Y	0.0232	35	21	6.4
	N-Myc	GC box	0.0700	99	39	6.0
	ATF	CREB	0.0248	33	11	5.6
	ATF	NF-Y	0.0053	7	5	5.5
	TATA(B)	GC box	0.1761	117	58	2.8
	TATA(B)	NF-Y	0.1489	96	42	2.7
	TATA(B)	ATF	0.0557	33	16	2.5
	TATA(B)	CCAAT	0.2422	136	55	2.4
(b)	TATA(B)	NF-kappaB	0.2642	128	32	2.0
	TATA(B)	CRE-BP1/c-Jun	0.1489	61	26	1.7
	TATA(B)	CREB	0.6929	285	65	1.7
	TATA(B)	Brn-2	0.2461	88	24	1.5
	TATA(B)	v-Myb	0.2810	91	38	1.4
	TATA(B)	C/EBPbeta	3.1638	1027	130	1.4

non-promoter. $CO_n(A,B)$ means coincidence bias of the pair which actually appears to one promoter relative to the random coincidence probability on one non-promoter. Because random cooccurrence of factor pairs is not rare on non-promoters, each $P_n(A,B)$ is still considerably high and thus $CO_n(A,B)$ is low. We think they are not sufficient for promoter region prediction, which we discuss later.

We show information of coincident cases of position specific factors on Table 4 (a) and (b). The column $P_n(A_p,B_p)$ means random coincidence probability of factor A and B within each specific range on one non-promoter. While, $CO_n(A_p,B_p)$ means bias of actual coincidence of factor A and B within each specific range on promoters relative to their random coincidence on non-promoters. We show only top 10 pairs by this evaluation which appeared on two promoters or more. Because such coincidences are quite rare on non-promoters, each $P_n(A_p,B_p)$ is quite low, and thus each $CO_n(A_p,B_p)$ is quite high(more than hundreds). The rightmost column ("Exp int") means interaction type (D: direct, I: indirect, ?: unseen(unknown)), which we will discuss next.

Table 4: (a) Coincident cases of position specific factors(top 10 only). (b) Factor pairs which cooccurr with Bucher TATA box within range of each determined position(top 10 only).

	Factor A	peak site (A)	Factor B	peak site (B)	$CO_n(A_p, B)$	$CO_n(A, B_p)$	$P_n(A_p, B_p)$	# of pairs	# of pro.	$CO_n(A_p, B_p)$	Exp int
	Barbie Box	-240 -230	TATA	-40 -20	20.1	35.2	0.0000	2	2	452.4	?
	Elk-1	-280 -270	ATF	-80 -20	31.0	18.1	0.0000	2	2	232.7	I
	Elk-1	-280 -270	GC box	-70 -50	58.9	17.2	0.0000	2	2	220.7	?
	TATA	-40 -20	MyoD	10 20	19.4	12.1	0.0001	4	4	218.5	?
(a)	TATA(B)	-40 -20	GC box	-70 -50	39.9	11.9	0.0003	18	13	218.3	D
	TATA	-40 -20	GC box	-70 -50	66.8	8.4	0.0000	2	2	187.9	D
	TATA	-40 -20	ER	-10 0	26.2	7.0	0.0001	2	2	157.3	I
	MEF-2	-40 -30	c-Rel	-250 -240	13.0	2.9	0.0002	6	3	130.5	?
	TATA(B)	-40 -20	CRE-BP1 /c-Jun	-60 -40	17.2	7.2	0.0003	9	6	129.1	D
	ATF	-80 -20	TATA(B)	-40 -20	9.7	34.1	0.0003	9	6	115.1	I
	TATA(B)	-40 -20	GC box	-70 -50	39.9	11.9	0.0003	18	13	218.3	D
	TATA(B)	-40 -20	CRE-BP1/c-Jun	-60 -40	17.2	7.0	0.0003	9	6	129.1	D
	TATA(B)	-40 -20	ATF	-80 -20	34.1	9.7	0.0003	9	6	115.1	I
	TATA(B)	-40 -20	NF-Y	-90 -60	26.1	13.2	0.0004	11	7	105.2	I
(b)	TATA(B)	-40 -20	NF-kappaB	-100 -90	17.6	12.9	0.0003	6	3	97.0	D
	TATA(B)	-40 -20	AP-4	-200 -190	14.7	4.9	0.0003	6	2	82.5	?
	TATA(B)	-40 -20	AP-4	-70 -60	14.7	3.7	0.0003	6	2	82.5	?
	TATA(B)	-40 -20	Elk-1	-280 -270	10.8	3.9	0.0003	6	3	75.0	I
	TATA(B)	-40 -20	CCAAT	-110 -60	23.5	6.7	0.0012	21	13	74.1	D
	TATA(B)	-40 -20	Evi-1	-200 -180	12.3	4.2	0.0008	12	2	60.2	?

5 Discussion

5.1 Peak spectrum and coincidence

We clarified many transcription factors have binding site specificity. Some of them have not been biologically examined in detail yet. We could also extract significant factor pairs without biological knowledge and confirmed many instances already known.

Sign "D" in rightmost column("Exp int") of Table 4 indicates direct interaction experimentally confirmed (binding, complex formation, (co)activation, association, competance, transactivation). We assumed inter-subunit interaction also binds their complexes. Their references are summarized in Table 5 (a).

While, sign "I" in Table 4 indicates indirect interaction assumed by combination of direct interaction. Although each direct interaction was experimentally confirmed, simultaneous interaction has not been confirmed yet. Table 5 (b) shows information of each hypothesized mediator. Typical example CBP was experimentally suggested to interact with Elk-1[9]. It is also coactivator which binds CREB and TFIID. While, high coincidence rate of Elk-1 and ATF/CREBP shown in this table hence suggests the probability that CBP

Table 5:
(a) Direct interaction confirmed by experiments.

Factor A	Factor B	Reference examples(A–B)
TATA box	Sp1(GC box)	*Proc Natl Acad Sci USA*(1996)**93**, 13611–6.
(TBP, TFIID,	c-Jun	*J Biol Chem*(1995)**270**, 10754–63.
TFIIA, TFIIE,	NF-kappaB	*Nature*(1993)**365**, 412–9.
TAFs)	AGP/EBP(CCAAT)	*Mol Cell Biol*(1997)**17**, 230–9.

(b) Indirect interaction assumed by combination of direct interaction. Hypothesized mediators are shown in the second column.

Factor A	Mediator M	Factor B	Reference examples(A–M)	Reference examples(M–B)
TATA box	NF-kappaB	ATF	*Nature*(1993)**365**,412-9.	*J Biol Chem*(1994)**269**,1159-65.
(TBP,		ER		*Mol Cell Biol*(1995)**15**,4971-9.
TFIID,	Oct-1	ATF	*J Biol Chem*(1995)**270**,19613-23.	*J Virol*(1996)**70**,332-40.
TFIIA,		NF-Y		*J Clin Inve*(1995)**95**,1684-9.
TFIIE,	CBP	ATF	*Nature*(1994)**370**,223-6.	*Nature*(1994)**370**,223-6.
TAFs)		Elk-1		*Bioc Biop Res*(1996)**228**,831-7.
Elk-1	CBP	ATF	*Bioc Biop Res*(1996)**228**,831-7.	*Nature*(1994)**370**,223-6.

binds also Elk-1 and ATF/CREBP. But we must prove them by biological experiments.

Other pairs(signed "?" in Table 4, e.g. Barbie – TATA, Elk-1 – GCbox(Sp1), TATA – MyoD, MEF-2 – c-Rel, TATA – AP-4, TATA – Evi-1) have currently no support of any interaction. For example, MyoD is known as a transcription factor specific to muscle organization. The relation between TFIID and MyoD is not experimentally elucidated in detail yet. However, this table shows the possibility of such strong interaction. Besides the combinations with TFIID, we could list up combinations of position specific factors. For example, they are a pair of Elk-1 and ATF family, a pair of Elk-1 and Sp1(GC box), a pair of MEF-2 and c-Rel, etc. Position specific coincidence of MEF-2 and c-Rel is rather mysterious. Because c-Rel is known to be specific to the immunity system, while MEF-2 is known to be related to the frame muscle. We must assume some unseen mechanism to understand such a coincidence.

In addition, Table. 4 suggests that direct(indirect) interaction tends to be short(long) distance. Unseen interaction includes both cases.

POSTSCRIPTER could thus extract significant patterns of binding sites. We expect that we could catch a part of an effective and general solid structure of promotors.

During our analysis, we found narrow peaks in Oct-1 and CdxA binding sites (see Table 2). But they disappeared from the result of the coincident pair extraction. This is because most of their binding sites overlapped the TATA box. Thus *POSTSCRIPTER* automatically excluded such overlapping binding pairs. We consider these bindings are just mistakes by misrecognizing TATA box as their own binding sites; their weight matrix is similar to TATA box. When the promoter area is distinguished from other areas by coincident factor

pairs, these trivial coincidence may be harmful. Our algorithm can exclude such unexpected cases.

5.2 Related methods

Kondrakhin et al. devised Smirnov's statistic ω^2 to calculate distribution bias of each factor binding site [3,4]. It can evaluate stochastically the global difference of upstream binding site relative to TSS. They also proposed to catch coincidence of factor pairs with χ^2 statistics. However, their method lacks algorithm to specify location peaks and failed to extract position specific factor coincidences.

Wingender compiled results of biological experiments[6]. They stored the information of binding site and the organization specificity of the transcription factor pairs to a transcription regulatory region database(TRRD). Moreover, they also brought together the knowledge of touching transcription factor pairs and made data base COMPEL. They will be significant clue to clarify the regulating mechanism on promoters. However, their database largely depends on the experiments already done. Their aim is not to propose predicting method. In addition, binding sites of each pair sometimes overlaps. It does not contribute to find global structures.

We first proposed the algorithm of automatic extraction of each factor binding site specificity on promoters, and the calculation method of their coincidence bias. We applied them to a set of promoter sequences and found that many factors have position specificity on them. Some of them proved the experimentally known results, and the others are new.

5.3 Promoter region prediction

Using the proposed method, we can extract many sets of transcription factors with specific position necessary to start transcription. This information can be used to discriminate promoter region from non-promoter region.

Let's look Table 4 again. For example, we can identify 13 promoters by catching pairs of TATA box and GC box on each specified binding site. This hits 10 percent or more of TATA-box including promoters. That is, we can prove that the cooccurrency of this pair is very high as well as the case of prokaryotic promoters. While, its random coincidence probability on non-promoter($P_n(A_p, B_p)$) is extremely low ($CO_n(A_p, B_p)$ is thus high). The number of promoters which include such pairs are so many that we may manage to separate promoters from non-promoters with high recall.

The currently proposed promoter region predicting tools commonly have the problem of high false positive rate[5]. This occurrs because they calculate

plausibility using only information of single factor occurrence. Because factors which binds well to promoters have tendencies that they binds also to non-promoter well, they are not effective if we use each of them independently. To identify promoters more effectively, we must set more severe limitations. One idea is to use our results. Each of the listed combination includes factor pairs with position restrictions from TSS. If we apply them to promoters, some templates will matches to each promoter. While, if we apply them to non-promoters, we expect that no template will match with any sequence.

If we can predict TSS candidates, it will be good clue to identify promoters because we can judge whether each factor binds to sites which have each fixed distance from the TSS respectively. If either factor does not bind to the specific sites, we can exclude the sequence. However, in general, we can not detect TSS well. We must only use gap between two factors. This loosens the restriction and rises random coincidence probability on non-promoters. To solve this problem, it will be good to expand our method to more general one in which templates consists of n factors. An easier method is to combine two or more templates of factor pairs obtained by this technique. That is, when some promoters include more than one templates, they are taken out as one group. Actually, most promoters include more than one templates.

5.4 Limitation of our method

Even if suggestive, our method can not catch full information. First, it does not handle adequately with the case that only relative distance between two factors are significant. Partialy they can be detected by our method because it processes special cases when they are located in each specific position. So we can grasp the tendency but not accurate.

Secondly, because it only detects discrete positional peaks, it can not take into account continuous mechanism. Typical example is DNA bending[7]. The chemical physical structure depends on global situation of DNA itself, and whether DNA-binding materials exist around it. If the bending range is not so wide, we can detect peaks with rather broad distribution. But we can not suggest strict rule behind them. There may be dropping information in one dimensional sequence of DNA.

Finally, we must additionary consider about inter-protein interaction and other signal transduction pathways in organization specific cells.

We took the approach that we have to catch special but simple cases. After that, we apply extracted rules to more generalized cases. It will be the first step to solve them.

6 Conclusion

We proposed an algorithm to detect position specificity of transcription factor binding sites on promoters and to calculate their coincidence novelty. Using these, we managed to extract many significant factor pairs. Some of them proved the interaction already found by biological experiments. The others also suggest unseen interaction.

In the traditional biological method, if we do not know much about proteins, we must examine the interaction with brute-force. However, many DNA sequences themselves suggest statistical information on the appearing patterns. With our prediction, we can take experiments more efficiently.

Our method *POSTSCRIPTER* can be generally used for the problem concerning position specificity of gene sequence besides promoters. One example is presumption of disease caused by two or more point mutations in coding regions. When the position specificity of mutations appeared as spectrum, their combination can be extracted if they are not random coincidence. We have the plan to apply our method as an effective means of *gene finding*.

Acknowledgements

This work is partially supported by Grant-in-Aid for Scientific Research on Priority Areas, "Genome Science" from the Ministry of Education, Science, Sports, and Culture, Japan.

References

1. P.Bucher. *J. Mol. Biol.*(1990)**212**, 563–578.
2. P.Bucher and E.N.Trifonov. *Nucl.Acids.Res.*(1986)**14**, 10009–10026.
3. A.E.Kel, Y.V.Kondrakhin, Ph.A.Kolpakov, O.V.Kel, A.G.Romashenko, E.Wingender, L.Milanesi and N.A.Kolchanov. ISMB-95197-205.
4. Y.V.Kondrakhin, A.E.Kel, N.A.Kolchanov, A.G.Romashenko and L.Milanesi. CABIOS(1995), Vol.11, No.5, 477–488.
5. D.S.Prestridge. *J. Mol. Biol.*(1995)**249**, 923–932.
6. E.Wingender, A.E.Kel, O.V.Kel, H.Karas, T.Heinemeyer, P.Dietze, R.Knuppel, A.G.Romaschenko and N.A. Kolchanov. *Nucleic Acids Research*(1997), Vol.25, No.1, 265–268.
7. E.Sjottem, C.Andersen and T.Johansen. *J.Mol.Biol.*(1997)**267**, 490–504.
8. S.Ou, L.F.Garcia-Martinez, E.J.Paulssen, R.B.Gaynor. *J.Virol.*(1994)**68(11)**, 7188–7199.
9. R.Janknecht and A.Nordheim. *Biochem.Biophys.Res.Commun.*(1996)**228(3)**, 831–837.