SEARCH FOR DNA CONFORMATIONAL FEATURES FOR FUNCTIONAL SITES. INVESTIGATION OF THE TATA BOX

M.P. PONOMARENKO, J.V. PONOMARENKO, A.E. KEL, N.A. KOLCHANOV
Institute of Cytology and Genetics, 10 Lavretyev Ave, 630090 Novosibirsk, Russia.

A method for search of DNA conformational features significant for functional sites is developed. The method uses helical angles averaged for known X-ray structures. Nucleotide sequences are assigned mean angles in a given region. Choice of the significant angles is based on their capabilities to discriminate functional sites from random sequences. The yeast, invertebrate and vertebrate TATA boxes are analyzed using this method. Regions neighboring the TATA boxes are found to have smaller helical twist and roll angles. The results agree with the experimental data on Dickerson-Drew dodecamers. There is a significant decrease in the length of a small roll angle region with increasing complexity of taxon organization.

1. INTRODUCTION

The problem of transcription regulation has become pivotal in structural molecular biology. There is a growing body of experimental information on the molecular mechanisms of transcription [1]. Attention has been focused on the TATA box because it is utilized to anchor transcription complexes [2]. X-ray structures for Dickerson-Drew dodecamers [3] and the TATA binding protein (TBP) complex with the TATA box [4] have been determined; the equilibrium dissociation constant of the complex has been measured [5]; nonspecific TBP-binding to DNA and its stable sliding along the DNA is now known to precede the formation of the TBP-TATA complex [2]. All this strongly suggests that a region in the neighborhood of the TATA box may be involved in the TBP-TATA recognition [2].

The theoretical approaches used to analyze the TATA box sequences have included consensus [6], long consensus pattern [7], information content [8], weight matrix [9, 10], statistical mechanics [11], neural network [12], recursive systems [13], and mathematical statistics [14]. These approaches have provided new insights into the structure and function of the TATA box. However, based on the structural results the TATA box has not been, as yet, accurately predicted. The question is raised: Are there conformational features in the neighborhood of the TATA-box that can be used to increase prediction accuracy ?.

Interest in the effect of nucleotide context on the helical configuration of B-DNA has been stirred by Dickerson and Drew [3, 15] and Calladine [16]. Subsequent analysis of sequence-dependent DNA structure relied on the Calladine rules [17]. The number of unraveled X-ray structures of B-DNA and DNA-protein complexes is increasing [1, 2, 4, 18, 19, 20, 21, 22, 23].

A method using a library of low energy conformation theoretically calculated for B-DNA hexanucleotides [24] has been developed [25]. It is improved here by introducing experimental data on averaged helical angles for dinucleotides [18, 21, 22, 23]. The yeast, invertebrate and vertebrate TATA boxes were analyzed using the combined method. It was determined that the mean values for the roll and helical twist angles of a TATA box are smaller than those for a random sequence. Closer examination revealed that the length of the region with a small roll angle decreases with increasing complexity of evolutionary organization.

2. MATERIALS AND METHODS

The conformational angles X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8, X_9 for dinucleotides are given in Table 1. Angles X_1, X_2, X_3 were taken from [18, 21, 22]; their values were averaged from known X-ray structures of DNA. Angles X_4, X_5, X_6, X_7, X_8, X_9 were

Table 1. The helical angles were averaged for known X-ray structures of DNA (degrees)

Dimer	Direction	Wedge	Helical twist			Roll		Tilt	
	Free	**Free**	**Free**	**Both**	**Comp**	**Both**	**Comp**	**Both**	**Comp**
s_1s_2	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9
AA	-154	7.2	35.6	35.5	35.5	0.5	1.0	1.0	1.5
AT	0	2.6	31.5	29.5	29.0	0	0	0	0
AG	2	8.4	27.7	35.5	35.0	5.0	5.0	1.0	1.0
AC	143	1.1	34.4	31.0	30.5	1.5	0	0.5	0
TA	0	0.9	36.0	37.5	39.5	4.0	2.5	0	0
TT	154	7.2	35.6	35.5	35.5	0.5	1.0	1.0	1.5
TG	64	3.5	34.5	36.0	36.0	5.0	6.5	-0.5	0.5
TC	-120	5.3	36.9	36.5	35.5	2.5	2.5	1.0	1.5
GA	120	5.3	36.9	36.5	35.5	2.5	2.5	1.0	1.5
GT	-143	1.1	34.4	31.0	30.5	1.5	0	0.5	0
GG	57	2.1	33.7	35.0	36.0	5.0	3.0	0	1.0
GC	180	5.0	40.0	35.0	34.5	-3.0	-2.5	0	0
CA	-64	3.5	34.5	36.0	36.0	5.0	6.5	-0.5	0.5
CT	-2	8.4	27.7	35.5	35.0	5.0	5.0	1.0	1.0
CG	0	6.7	29.8	34.5	34.5	5.0	6.5	0	0
CC	-57	2.1	33.7	35.0	36.0	5.0	3.0	0	1.0
Ref.	[18, 21, 22]		[23]						

Notes: **Free** is for the angles X_1, X_2, X_3 averaged from free DNA; **Comp** is for the angles X_5, X_7, X_9 from DNA-protein complexes; **Both** is for the angles X_4, X_6, X_8 from the complexes and from free DNA

taken from [22]; X_5, X_7, X_9 were averaged from DNA-protein complexes; and X_4, X_6, X_8 from the complexes and from free DNA.

The definitions of the direction and wedge angles, X_1 and X_2, and their geometrical implications have been given in [22]. The helical twist (X_3, X_4, X_5), roll (X_6, X_7) and tilt (X_8, X_9) angles have been recommended as nomenclature for description of DNA conformations by the EMBO Workshop [20]. As shown in Table 1, the helical twist angles averaged for different data sets (X_3, X_4, X_5) differ from one another. It was not clear which angle set would provide significant

Table 2. Sets of nucleotide sequences

no	Set		Sequences			TATA-pattern	
	promoters	type	number	length	margins	start	aligned by
1	Yeast	S^+	75	70 bp	[-30; +39]	0	subsequence TATAT
2	Invertebrate	S^+	158	70 bp	[-30; +39]	0	weight matrix [10]
3	Vertebrate	S^+	486	70 bp	[-30; +39]	0	weight matrix [10]
4	*E.coli*	S^+	135	70 bp	[-39; +30]	0	consensus, TATAAT [6]
5	Random	S^-	500	70 bp	by $\{S^+\}$	none	none

Note: sets 1 and 4 are compiled on the basis of EMBL Data Library, sets 2 and 3 on the basis of EPD

conformational features of the TATA box. Therefore, each single set was analyzed one by one. Two data sets for the roll angle (X_6, X_7) and two data sets for the tilt angle (X_8, X_9) were also analyzed for the same reason.

Table 2 presents the nucleotide sequences used in analysis. There were 5 sequence sets. Set 1, 75 yeast promoters aligned by their subsequence TATAT. Sets 2 and 3, 158 invertebrate and 486 vertebrate promoters. The TATA box weight matrix [10] was used to align these promoters. Set 4, 135 *E.coli* promoters. The TATA pattern of these promoters, Pribnow box, was identified by the consensus TATAAT [6]. Sets 1 and 4 were compiled on the basis of EMBL Data Library (release 42) with keywords «promoter» and «primary transcript»; and sets 2 and 3 were taken from the database EPD (release 45). Set 5, 500 random sequences with the same nucleotide frequencies. This set served to search for significant differences between the promoters and the random sequences. All the sequences were 70 b.p. long. The start of the TATA pattern was taken as 0. Sequence position was numbered in this way to pinpoint the TATA pattern whose position varied with respect to transcription start.

Sequence regions $\{s_a...s_i...s_b\}$ located from a to b and of length (b-a+1) were examined. All the possible sequence regions not smaller than one dinucleotide were

taken into account. Their number for a sequence L long was n(L)=L×(L-1)/2. The total number of sequence regions was n(70)=70×(70-1)/2=2415.

Each region a, b with the sequence S={s_a...s_i...s_b} was assigned a mean value of a conformational angle X_k:

$$X_{k,a,b}(S) = \frac{1}{b-a}\sum_{i=a}^{b-1} X_k(s_i s_{i+1}), \qquad (1)$$

where X_k is any one of the angles given in Table 1.

Let formula (1) be applied to the sequence S={TATA}. The mean value $X_{7,1,4}$(TATA) for the roll angle X_7 was calculated as: $X_{7,1,4}$(TATA) = [X_7(TA) + X_7(AT) + X_7(TA)] / (4 - 1) = [2.5° + 0.0° + 2.5°] / 3 = 5° / 3 = 1.67°.

Formula (1) was applied to each of the 9 angles given in Table 1. A number of angles $X_{k,a,b}$ can be calculated N(L)=9×n(L) for a sequence L long. Thus, a total number of N(70)=9×2415=21735 angles was tested.

The SITEVIDEO system [26, 27] was used to analyze the conformational angles $X_{k,a,b}$. The system allows to test for significance various physicochemical, statistical, contextual and conformational properties defined theoretically or experimentally. In a previous analysis [25], we have modified SITEVIDEO by using Sklenar's theoretical parameters for B-DNA [24]. In the present study, the SITEVIDEO was further modified by introducing experimental parameters for B-DNA [18, 21, 22, 23].

Fig.1 illustrates how the significant angles for the TATA patterns are identified by the current version of SITEVIDEO. Two sequence sets S^+ and S^- were initial, with S^+ containing the TATA patterns and S^- random sequences. SITEVIDEO treats S^+ and S^- in four steps.

Step 1, calculation of all the 21,735 angle values $X_{k,a,b}(S)$ for all the sequences from S^+ and S^- by formula (1). This yields the sets $X_{k,a,b}(S^+)$ and $X_{k,a,b}(S^-)$ for the TATA and random sequences, respectively.

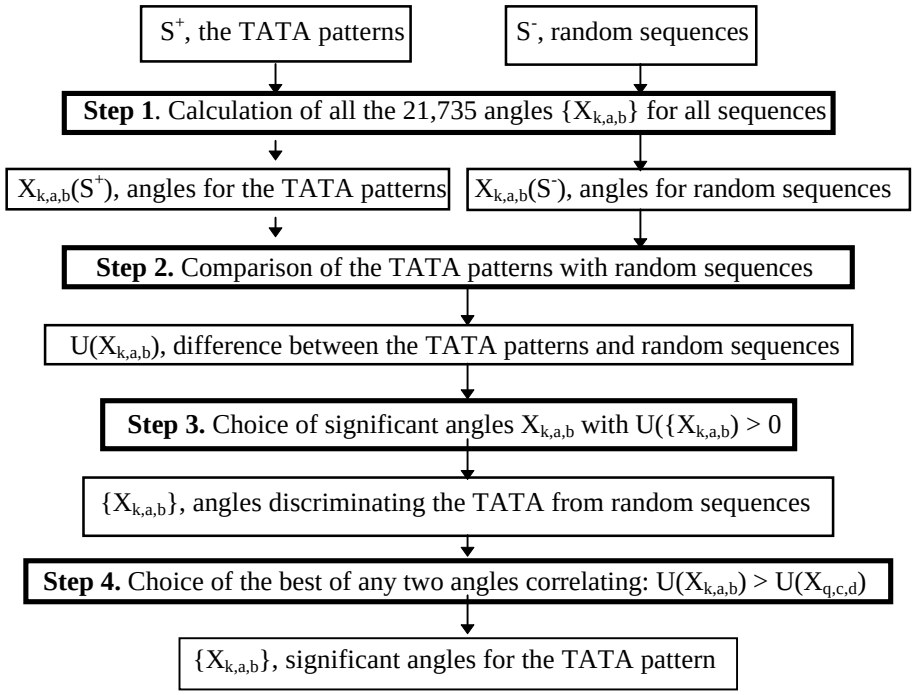


Fig. 1. A search for significant conformational angles in the TATA pattern.

Step 2, pairs $X_{k,a,b}(S^+)$ and $X_{k,a,b}(S^-)$ are analyzed. This yields an estimate $U(X_{k,a,b})$ of the difference between $X_{k,a,b}(S^+)$ and $X_{k,a,b}(S^-)$. The algorithm [27] was used to test for significance as mean, variance, range and other statistics. A positive weight from 0 to 1 was assigned to each significant difference, and a negative weight from -1 to 0 to each insignificant difference, and the mean value of the weights is the estimate $U(X_{k,a,b})$ of the differences between $X_{k,a,b}(S^+)$ and $X_{k,a,b}(S^-)$. According to the decision making theory [28], the U value varies from -1 to 1; insignificant differences predominate at $U<0$, and significant at $U>0$; and the greater the U, the larger are the number of significant differences.

Step 3, all the conformational angles $X_{k,a,b}$ with $U(X_{k,a,b})>0$ are taken for further analysis so that only angles in which the TATA patterns differ from random sequences are chosen.

Step 4, when any angles $X_{k,a,b}$ and $X_{q,c,d}$ correlate, the angle $X_{q,c,d}$ with the smaller $U(X_{q,c,d})$ value is discarded, the other $X_{k,a,b}$ with the greater $U(X_{k,a,b})$ is fixed. A list of significant angles $X_{k,a,b}$ for the TATA pattern is thus generated.

3. RESULTS AND DISCUSSION

The yeast, invertebrate, vertebrate TATA boxes (Table 1) were analyzed using the developed method. Analysis was based on a set of 500 random sequences. Table 3 summarizes the obtained results. Two independent conformational angles were found for the yeast TATA box. One was a mean helical twist angle in the region from -2 to 9 with U=0.95. The mean helical twist angle was $33.1^o \pm 0.3^o$ for the TATA box and $34.7^o \pm 0.4^o$ for the random sequences. The occurrence probability

Table 3. Significant conformational angles for the TATA box

Promoters	Angle				Value, (degrees)		Sig-
	name	X_k	region, [a; b]	$U(X_{kab})$	promoters mean ± s.d.	random mean ± s.d.	nifi-cance
Yeast	twist	X_4	[-2; 9]	0.95	33.1±0.3	34.7±0.4	10^{-7}
	twist	previous[#]	[-10; 9]	0.81	32.2±0.9	35.6±0.8	10^{-9}
	roll	X_7	[-13; 7]	0.93	1.97±0.26	2.53±0.28	10^{-7}
Invertebrate	twist	X_4	[0; 6]	0.91	33.8±0.4	34.7±0.6	10^{-13}
	twist	previous[#]	[-3; 5]	0.78	33.7±0.7	36.3±1.3	10^{-8}
	roll	X_7	[-7; 7]	0.90	1.93±0.23	2.67±0.27	10^{-40}
Vertebrate	roll	X_7	[0; 6]	0.99	0.88±0.25	2.22±0.57	10^{-40}
E.coli	roll	X_7	[-30; 10]	0.72	2.35±0.22	2.60±0.16	10^{-7}

[#]) Asterisks, our previous results [25] based on Sklenar's parameters [24]

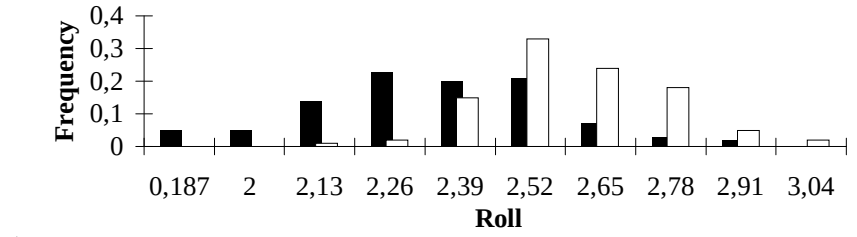
for the difference between the mean values was less than 10^{-7}. The mean helical twist angle for the invertebrate TATA box in the region from 0 to 6 is $33.8^{o}\pm0.4^{o}$, significantly smaller than for the random sequences ($34.7^{o}\pm0.6^{o}$).

Thus, a small helical twist angle in the neighborhood of the TATA box was found. Based on Sklenar's parameters [24], we have obtained [25] a small helical twist angle in the neighborhood of the yeast and invertebrate TATA boxes (Table 3). The previous results are in agreement with our present and also with the X-ray structure of Dickerson-Drew dodecamers [3].

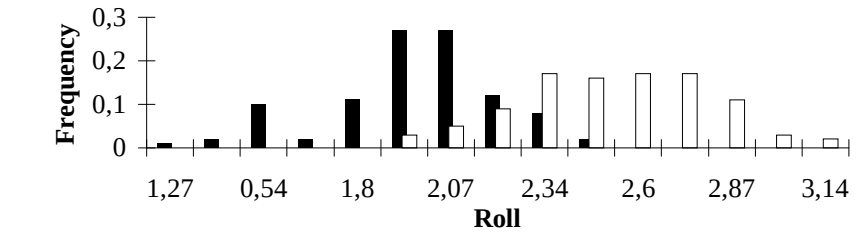
A mean roll angle in the region from -13 to 7 was the other significant angle found in the neighborhood of the yeast TATA box (Table 3). Angle value was $1.97^{o}\pm0.26^{o}$ for the TATA box and $2.53^{o}\pm0.28^{o}$ for the random sequences, the difference between the two mean values being significant ($\alpha<10^{-7}$). The roll angle is also smaller for the invertebrate and vertebrate TATA boxes than for the random sequences (Table 3).

It is of interest that the roll angles were equally small for the TATA and Pribnow boxes (Table 3), and also for the Dickerson-Drew dodecamers [3]. Moreover, the conformational invariant in the neighborhood of the TATA box is consistent with the experimental data indicating that TBP stably sliding along the DNA precedes the formation of the TBP-TATA complex [2].

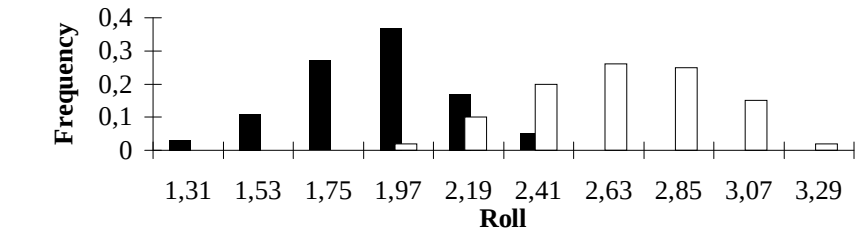
The roll angle for the TATA patterns (black) and for the random sequences (white) are given in Fig.2. The increasing difference between the TATA and random sequences may be ordered as *E.coli* $\rightarrow$ yeast $\rightarrow$ invertebrate $\rightarrow$ vertebrate. The significance of the order was tested by assigning rank values to each of the taxon according to the increasing complexity of organization (Fig.3, the vertical scale). The length of the region with a small roll angle was also scaled. The linear correlation coefficient r between the rank and the length was -0.96 ($\alpha<0.05$).



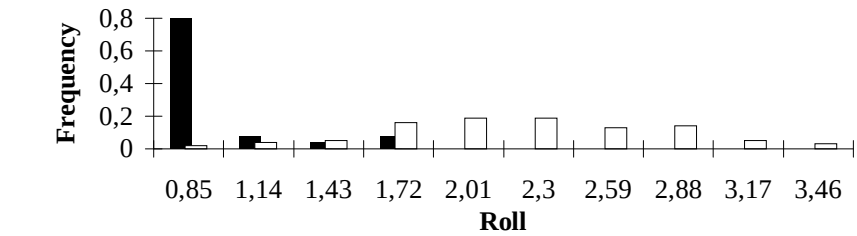
a)



b)



c)



d)

Fig. 2. Frequency histograms of the roll angle for the TATA (black) and random (white) sequences. Taxa: a) *E.coli*; b) yeast; c) invertebrate; d) vertebrate.

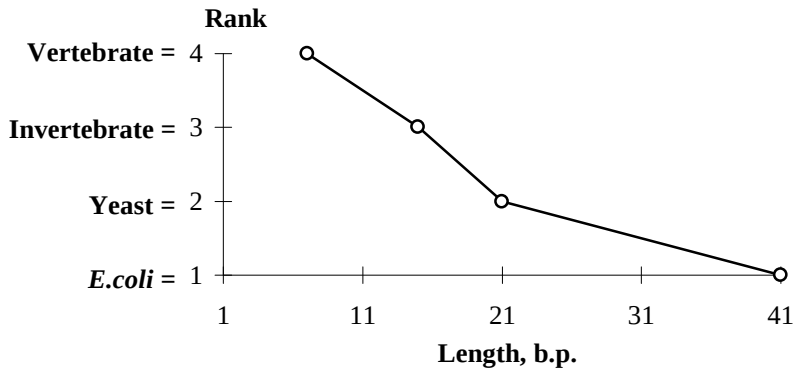


Fig. 3. Reduction of the promoter region with a small roll angle with increasing complexity of taxa organization. The linear correlation coefficient r=-0.96 (α<0.05).

Thus, the promoter region with a small roll angle reduces with increasing complexity of taxon organization. The reduction might have resulted from an increase in the number of transcription factors associated with increasing complexity of taxon organization. The *E.coli* transcription machinery is simplest because RNA-polymerase recognizes the Pribnow box and performs all the other functions. Consequently, an error admitted by RNA-polymerase in the false recognition cannot be corrected. For this reason, the Pribnow box must be safeguarded by other signals necessary for the *E.coli* transcription machinery. Thus, the *E.coli* region with a small roll angle were expected to be longest. This was, indeed, the case.

The formation of an eukaryotic transcription complex is multistep. For the majority of genes, the TBP-TATA recognition initiates the complex formation. In a cascade, other transcription factors bind to the DNA sites proximal to or distant from the TATA box. Hence, a false TBP-TATA recognition would not lead to the formation of the transcription complex, because they have no transcription factor

binding sites providing recognition around the false TATA box. Therefore, the role of the TATA box in the transcription machinery is restricted compared to the Pribnow box. The regression may possibly result in a reduction of the promoter region with a small roll angle in the neighborhood of the TATA box. For this reason, the reductions were expected in the passages from unicellular (yeast) to multicellular (invertebrate, vertebrate) and from the invertebrate to the vertebrate. These expectations were met (Fig.3).

Taken together, the obtained data evidence that analysis of "sequence-conformation" relationships allows to reveal conformational features maintained unaltered during evolution. There is a good agreement between our previous results [25] based on Sklenar's theoretical parameters [24] and our present using experimental parameters [18, 21, 22, 23]. Thus, the Sklenar's parameters for the B-DNA [24] offers promise in structural molecular biology.

ACKNOWLEDGMENTS

The authors are grateful to A.Fadeeva for translating this paper into English. This research was supported by the Russian Foundation of Fundamental Investigation.

REFERENCES

1. S. Neidle, DNA structure and recognition, IRL Press, New York (1994)
2. R.A. Coleman and B.F. Pugh, J. Biol. Chem., 270, 13850 (1995)
3. R.E. Dickerson and H.R. Drew, J. Mol. Biol., 149, 761 (1981)
4. Y. Kim et al, Nature, 365, 512 (1993)
5. S. Hahn et al, Proc. Nath. Acad. Sci. U.S.A., 86, 5718 (1989)
6. R. Staden, Comput. Appl. Biosci., 5, 89 (1989)
7. P. Taylor et al, Comput. Appl. Biosci., 7, 495 (1991)
8. T.D. Schneider et al, J. Mol. Biol., 188, 415 (1986)

9. D.K. Hawley and W.R. McClure, Nucleic Acids Res., 11, 2237 (1983)
10. P. Bucher, J. Mol. Biol., 212, 563 (1990)
11. O.G. Berg and P.H. von Hipple, J.Mol.Biol., 193, 723 (1987)
12. M.C. O'Neill, Nucleic Acids Res., 19, 313 (1991)
13. L. Milanezi et al, in Guide to human genome computing, Ed. M. Bishop, Academic Press, London, 249 (1994)
14. Y.V. Kondrakhin et al, Comput. Applic. Biosci., 11, 477 (1995)
15. R.M. Wing et al, Nature, 287, 755 (1980)
16. C.R. Calladine, J. Mol. Biol., 161, 343 (1982)
17. C.R. Calladine and H.R. Drew, J. Mol. Biol., 178, 773 (1984)
18. W. Kabsh et al, Nucleic Acids Res., 10, 1097 (1982)
19. R.E. Dickerson, J. Mol. Biol., 166, 419 (1983)
20. R.E. Dickerson et al, EMBO J., 8, 1 (1989)
21. A. Bolshoy et al, Proc. Nath., Acad. Sci. U.S.A, 88, 2312 (1991)
22. E.S. Shpigelman et al, Comput. Applic. Biosci., 9, 435 (1993)
23. M. Suzuki and N. Yagi, Nucleic Acids Res., 23, 2083 (1995)
24. H. Sklenar, in Proceedings of the International Workshop on Computational analysis of eukaryotic transcriptional regulatory elements, Heidelberg, 44 (1996)
25. N.A. Kolchanov et al, in Proceedings of the International Workshop on Computational analysis of eukaryotic transcriptional regulatory elements, Heidelberg, 40 (1996)
26. A.E. Kell et al, Comput. Applic. Biosci., 9, 617 (1993)
27. M.P. Ponomarenko et al, in Computer analysis of genetic macromolecules. Eds. N.A. Kolchanov and H. Lim, World Sci. Pub., Singapore, 35 (1993)
28. P.C. Fishburn, Utility theory for decision making, Wiley, New York (1970)