Molecular Biology, Vol. 32, No. 2, 1998, pp. 220–232. Translated from Molekulyarnaya Biologiya, Vol. 32, No. 2, 1998, pp. 255–267.

GENE MOLECULAR BIOLOGY

UDC 577.113.3:57.087.1

Functional Sites in Pro- and Eukaryotic Genomes: Computer Models for Predicting Activity

N. A. Kolchanov[1], M. P. Ponomarenko[1], Yu. V. Ponomarenko[1], N. L. Podkolodnyi[2], and A. S. Frolov[1]

[1] *Institute of Cytology and Genetics, Siberian Division, Russian Academy of Sciences, Novosibirsk, 630090 Russia; E-mail: kol@bionet.nsc.ru*

[2] *Computing Center of Siberian Division, Russian Academy of Sciences, Novosibirsk, 630098 Russia*

Received August 4, 1997

Abstract—Here we propose an approach for predicting the activity of functional DNA and RNA sites. This approach includes (1) identification of context-dependent conformational, physicochemical, and statistical properties of sites significant for their functioning; (2) development of a model on their basis for predicting site activity from its sequence; and (3) automatic generation of programs for predicting site activity based on these models. This approach has been realized as a computer system ACTIVITY, which includes databases of site activity as well as conformational, physicochemical, and statistical properties of DNA and RNA. ACTIVITY is accessible via Internet (http://www.bionet.nsc.ru/SRCG/Activity/) and allows real-time analysis of experimental data on functional site activity. We analyzed 70 samples of sites involved in various molecular biological processes and revealed statistical, conformational, and physicochemical properties significant for activity of these sites. We also developed methods for predicting site activity from their nucleotide sequences.

Key words: activity, site, computer analysis, DNA, RNA

INTRODUCTION

Replication, transcription, splicing, translation, and other molecular genetic processes are controlled by specific activity of the sites functioning via interaction with the corresponding proteins or RNA-protein complexes [1]. At present we know thousands of such sites together with their nucleotide sequences and location in DNA or RNA [2–6]. Until recently, computer analysis was directed at development of methods for recognizing the sites in arbitrary sequences (reviewed in [7]). Progress in this field demonstrates that site recognition does not suffice for understanding the functional organization of nucleotide sequences, since any site is described by another important property—the level of its specific activity—in addition to sequence and position. Experimental data demonstrate that the sites of one type located in different DNA (RNA) regions can have activity differing by several orders of magnitude [8].

Studying the properties of the functional sites responsible for their activity becomes more and more important for molecular biology, primarily because the difference in the site activity forms the basis for differential activity of the genes and their coordinated functioning in pro- and eukaryotic organisms. Understanding these properties is also important for constructing molecular genetic systems with predefined levels of gene expression. In addition, possible mutational impairment of the sites that often leads to loss or, conversely, sharp increase in their activity [8], and thus to pathologies [9], also attracts considerable attention to this problem.

The problem of predicting the activity of functional sites from their nucleotide sequences was formulated by McClure *et al.* [10], who proposed a method for predicting *E. coli* promoter activity from the degree of their similarity to the consensus. Stormo *et al.* [11] introduced weight matrices for predicting site activity. Berg and von Hippel [12] proposed to use frequency matrices for predicting site activity on the basis of statistical mechanics of DNA–protein interaction. Weight matrices were also used for predicting *E. coli* ribosome affinity for synthetic RNA [13]. In 1993 Jonsson *et al.* [14] applied a neural network for predicting *E. coli* promoter activity. Later this approach was used for predicting INR site and TATA box activity in eukaryotes [15]. However, although hundreds of samples of the sites with known activities are available now, the possibility of site activity pre-

0026-8933/98/3202-0220$20.00 © 1998 МАИК Наука /Interperiodica Publishing

diction from their sequences was demonstrated only in a few concrete cases. Primarily, this is due to the lack of universal methods for predicting site activity from their nucleotide sequences.

Here we propose an approach to solving this problem. It includes (1) identification of context-dependent conformational, physicochemical, and statistical properties of the sites that are significant for their functioning; (2) development of a model on their basis for predicting site activity from their sequences; and (3) automatic generation of programs predicting site activity based on these models. This approach has been realized in a computer system ACTIVITY, which includes databases on the site activity as well as conformational, physicochemical, and statistical properties of DNA and RNA. ACTIVITY is accessible via Internet (http://www.bionet.nsc.ru/SRCG/Activity/) and allows real-time analysis of the experimental data on the functional site activity. Here we present a brief description of the proposed approach, while details can be found in our previous publications [16–20]. We used ACTIVITY to study 70 site samples and developed a method for predicting activity with good correspondence to the experimental data for each of the samples.

LINEAR ADDITIVE MODEL FOR PREDICTING SITE ACTIVITY

Diagram of the ACTIVITY system for predicting the activity of functional DNA and RNA sites is shown in Fig. 1. The main properties of the proposed approach are considered in the course of describing this system and discussing the results obtained.

Canonical equation. We assume that specific activity of site F is a function of context-dependent (statistical, physicochemical, and conformational) properties of its nucleotide sequence S. These properties are divided (see [20]) into (1) obligate ones, which are uniform for all sequences S_n of this site and determine the basal level of its activity, and (2) facultative ones, which are specific for each sequence S_n of a given site and provide its activity modulation relative to the basal level. Then activity of the sites with sequence S_n is described with the following equation within the frames of the linear additive model:

$$F(S_n) = F_0(S_n) + \sum_{k=1}^{K} F_k X_k(S_n), \qquad (1)$$

where $F_0(S_n)$ is the basal activity specific for sites of this type determined by their obligate properties; $\{X_k\}_{k=1,K}$ are facultative properties; F_k is contribution of facultative property X_k to activity F.

Within the framework of the model (1), development of the methods for predicting site activity F from its sequence is reduced to identification of the facultative properties $\{X_k\}_{k=1;K}$ significant for its functioning, calculation of weight coefficients $\{F_k\}_{k=0;K}$ that determine the contribution of each facultative property to the site activity, and determination of coefficient F_0 describing the contribution of the obligate properties to site activity. The linear additive model (1) was selected for approximation of the site activity as it required a minimal volume of the experimental data for its optimization.

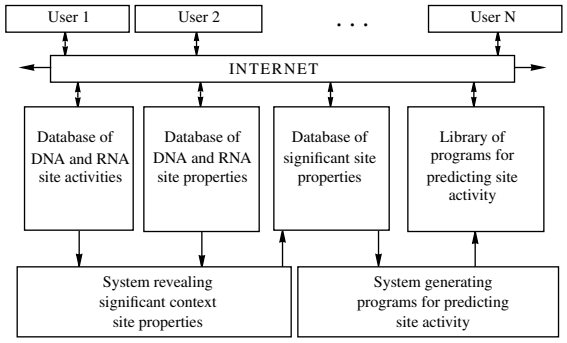

Fig. 1. Diagram of the ACTIVITY computer system.

Initial data for studying site activity are sets of nucleotide sequences with experimentally determined values of their specific activity. These data are stored in the site activity database (Fig. 1). In this case we use a specific format compatible with the SRS language [21] for automatic processing of "data search" demands. A typical example of such data is presented in Fig. 2. It describes *E. coli* promoters with determined "strength" values expressed as $-\log(P_{bla})$ [14] and includes the activity name (MN field), measure (AU field), data source (RA and RJ fields), name of each site variant (SC field), its sequence (SS field), and the activity value (SA field). For instance, the *E. coli* promoter LS1 has the 68 bp sequence "TCCGT … AGGAAT" and strength $-\log(P_{bla})$ = 2.143.

Context-dependent properties of sites. First, let us consider the statistical properties of the site nucleotide context [20] as potentially significant for their functioning. In the ACTIVITY system, such properties include weighted concentrations of mono-, di-, tri-, and tetranucleotides specific for a given site sequence S_n (Table 1). As shown earlier [16–20], weighted concentrations of oligonucleotides are important for both site recognition and predicting its activity.

In the case of a sequence $S = s_1 \ldots s_i \ldots s_L$ of length L, the weighted concentration of the oligonucleotide $Z =$

```
MT  DATABASE "ACTIVITY"
//
MN  Escherichia coli promoter strength
AU  Digital logarithmic scale, -log[Pbla]
//
RA  Jonsson J, Norberg T, Carlsson L, Gustafsson C, Wold S
RJ  Nucleic Acids Res 21: 733 739 (1993)
//
SC  LS1
SS  TCCGTCTCGA CGGGTTGACA CAAAAGCCAC AAGGGGTTAT AATGAGCACA
SS  TAAACTTGAG AGAGGAAT
SA  2.143
//
SC  LS2
SS  TCCGTATAGA CAGTTTGACA CAAAAGCCAC AAGGTGTTAT AATGAGCACA
SS  TAAATTTGAG AGAGGAAT
SA  2.127
//
.....
SC  N25/aDSR
SS  CATAAAAAAT TTATTTGCTT TCAGGAAAAT TTTTCTGTAT AATAGATTCA
SS  TCCGGAATCC TCTTCCCG
SA  0.431
//
SC  con/anti
SS  ATTCACCGTC GTTGTTGACA TTTTTAAGCT TGGCGGTTAT AATGGATTCA
SS  TCCGGAATCC TCTTCCCG
SA  0.255
//
```

Fig. 2. Presentation of experimental data on DNA and RNA site activities (exemplified by *E. coli* promoter strength [14]): activity name (MN), measure (AU), data source (RA and RJ), name of each site variant (SC), its sequence (SS), and the activity value (SA).

$z_1 \ldots z_j \ldots z_m$ of length m is calculated from the equation

$$X_{Z,m,w}(S) = \sum_{i=1}^{L-m+1} w(i)\delta_Z(s_i s_{i+1} \ldots s_{i+m-1}), \quad (2)$$

where $1 \le m \le L$; $\delta_Z(s_i s_{i+1} \ldots s_{i+m-1})$ is a function describing the presence (1) or absence (0) of oligonucleotide Z in each position i of sequence S:

$$\delta_Z(s_i \ldots s_{i+m-1}) = \begin{cases} 1, & \text{if } s_i \in z_1, \ldots, s_{i+m-1} \in z_m; \\ 0, & \text{if } s_{i+h-1} \notin z_h (1 \le h \le m), \end{cases}$$

where $s_i \in \{A, T, G, C\}$; $z_j \in \{A, T, G, C, W = A/T, R = A/G, M = A/C, K = T/G, Y = T/C, S = G/C, B = T/G/C, V = A/G/C, H = A/T/C, D = A/T/G, N = A/T/G/C\}$; $w(i)$ is the function of significance of positions ($0 \le w(i) \le 1$), which takes into account the position dependence of oligonucleotides contributing to the site activity. Function $w(i)$ is governed by a simple rule: the greater is the significance of position i for the site functioning, the higher weight $w(i)$ is assigned to it. Figure 3 exemplifies such weight functions $w(i)$, which assign the highest significance to the site activity for its left (a), central (b), or right (c) regions. ACTIVITY uses 180 such weight functions $w(i)$, and each has one extremum (minimum or maximum) within the site and differs from the other functions by the location or halfwidth of this extremum. ACTIVITY analyzes all the possible combinations of oligonucleotides Z of length m and the weight functions $w(i)$ used for each site.

The efficiency of a site functioning to a large extent depends on its physicochemical and conformational properties. The conformational properties affect the stereochemical match between the sites and interacting proteins [1]. Melting temperature and DNA persistent length [22], its bending rigidity [23] and entropy [24] determine the conformational dynamics of the functioning site [25]. Multiple publications demonstrate local heterogeneity of the conforma-

Table 1. Statistical, conformational, and physicochemical properties of DNA and RNA used for studying functional site activity

Type	k	Property	Designation	Units	Min	Max	Ref.
Statistical (DNA and RNA)	1	Concentration of nucleotide $Z = z_1$	$X_{Z,1,w}$	bp	0.00	L	[16]
	2	Concentration of dinucleotide $Z = z_1z_2$	$X_{Z,2,w}$	bp	0.00	L-1	[16]
	3	Concentration of trinucleotide $Z = z_1z_2z_3$	$X_{Z,3,w}$	bp	0.00	L-2	[16]
	4	Concentration of tetranucleotide $Z = z_1z_2z_3z_4$	$X_{Z,4,w}$	bp	0.00	L-3	[16]
Conformational (B-DNA)	5	Angle	twist	deg	31.1	41.4	[26]
	6	Angle	propeller	deg	–17.3	–6.7	[29]
	7	Angle	tip	deg	–1.64	6.7	[26]
	8	Angle	inclination	deg	–1.43	1.43	[26]
	9	Angle	tilt	deg	–0.7	2.8	[27]
	10	Angle	bend	deg	2.16	6.74	[26]
	11	Angle	wedge	deg	1.1	8.4	[28]
	12	Angle	direction	deg	–154	180	[28]
	13	Angle	roll	deg	–2.0	6.5	[27]
	14	Shift	rise	Å	3.16	4.08	[26]
	15	Shift	slide	Å	–0.37	1.46	[29]
	16	Minor groove dimension	width	Å	4.62	6.40	[26]
	17	Minor groove dimension	depth	Å	8.79	9.11	[26]
	18	Minor groove dimension	size	Å	2.7	4.7	[29]
	19	Minor groove dimension	dist	Å	2.79	4.24	[29]
	20	Major groove dimension	WIDTH	Å	12.1	15.5	[26]
	21	Major groove dimension	DEPTH	Å	8.45	9.60	[26]
	22	Major groove dimension	SIZE	Å	3.26	4.70	[29]
	23	Major groove dimension	DIST	Å	3.02	3.81	[29]
	24	Grooves difference	clash	ϕ	0.00	2.53	[29]
Physicochemical (B-DNA)	25	Frequency of contact with nucleosome	P_{nucl}	%	1	18	[22]
	26	Flexibility at minor groove	m	μ	1.02	1.27	[23]
	27	Flexibility at major groove	M	μ	0.99	1.18	[23]
	28	Persistent length	λ	bp	20	130	[22]
	29	Melting temperature	T_{m}	°C	36.7	136.1	[22]
	30	Enthalpy change	ΔH	kcal/mol	–11.8	–5.6	[24]
	31	Entropy change	ΔS	cal/mol/K	–28.4	–15.2	[24]
	32	Free energy change	ΔF	kcal/mol	–2.8	–0.9	[24]

Note: L, site length; $z \in \{A, T/U, G, C, W = A/T/U, R = A/G, M = A/C, K = T/U/G, Y = T/U/C, S = G/C, B = T/U/G/C, V = A/G/C, H = A/T/U/C, D = A/T/U/G, N = A/T/U/G/C\}$; w is the function $w(i)$ of significance of a site position i for its activity [see equation (2)].

tional and physicochemical properties of DNA and their dependence on the nucleotide context [22–24, 26–30]. X-Ray analysis of B-DNA duplexes as well as their complexes with proteins provided the mean conformational coefficients for each dinucleotide in B-DNA [26–29]. Several mean physicochemical coefficients were also determined for each dinucleotide [22–24]. Examples of conformational and physicochemical properties of this kind are shown in Table 1. Such data are stored in a special database of the ACTIVITY system (Fig. 1). For instance, Fig. 4 demonstrates storage of a conformation property **Direction**, the angle between the long axis of a base pair and the axis between current and preceding pair planes [28]. One can see that AA and GC dinucleotides have the lowest and the highest **Direction**

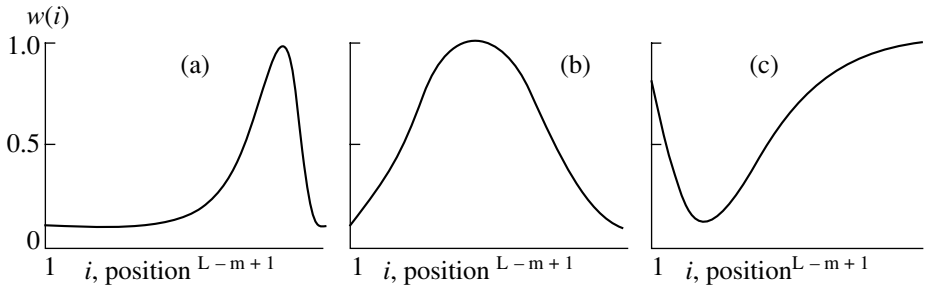

Fig. 3. Examples of weight functions $w(i)$ of site position i significance for equation (2).

```
MT DATABASE "ACTIVITY"
//
MN  Conformational
MD  B-DNA
//
RA  Shpigelman ES, Trifonov EN, Bolshoy A
RJ  Comput Appl Biosci  9: 435-440 (1993)
//
PN  Direction, Eulerian angle
PM  Averaged for X-ray structures known
PU  degree
//
AA  -154.0
AT     0.0
AG     2.0
AC   143.0
TA     0.0
TT   154.0
TG    64.0
TC  -120.0
GA   120.0
GT  -143.0
GG    57.0
GC   180.0
CA   -64.0
CT    -2.0
CG     0.0
CC   -57.0
//
```

Fig. 4. Presentation of conformational and physicochemical properties of DNA and RNA within the ACTIVITY database (exemplified by "**Direction**" angle of B-DNA helix [28]).

(–154° and 180°, respectively). Such a strong dependence of physicochemical and conformational properties of DNA and RNA on the nucleotide context allows us to consider them as potentially significant for the site activity.

The sequence of site $S = s_1 \ldots s_a \ldots s_i \ldots s_b \ldots s_L$ of length L can be described by the mean conformational or physicochemical property of DNA or RNA in the region $[a; b]$:

$$X_{q,a,b}(S) = \frac{\sum_{i=a}^{b-1} P_q(s_i s_{i+1})}{b-a}, \qquad (3)$$

where P_q is a qth property from the database of DNA and RNA properties; $1 \le a \le (b-1) \le (L-1)$.

Hence, the ACTIVITY system calculates and analyzes two types of context-dependent properties of the site sequences: statistical properties (weighted concentrations of oligonucleotides) and conformational and physicochemical properties (averaged for a site) using equations (2) and (3), respectively.

Exhaustive property search. Usually we have no information on the position within the sites where concrete oligonucleotides affect their activity. Hence, while studying the statistical properties of the sites, we consider 180 functions $w(i)$ with different positions and shapes of the maxima for each oligonucleotide Z of length m (Fig. 3). Provided a fixed combination $\langle Z, m, w\rangle$ for each sequence S_n, the weighted concentration of oligonucleotide $X_{Zmw}(S_n)$ is calculated from equation (2). The total number of $\langle Z, m, w\rangle$ combinations equals $\approx 15^m \times 180$; for $m = 4$ it is $\approx 10^7$. In a similar way, the region $[a, b]$ within the site with the highest effect of qth DNA (RNA) property on the site activity is not known in advance for conformational and physicochemical properties. In this case we also carry out exhaustive search for all possible combinations $\langle q, a, b\rangle$ and value $X_{qab}(S_n)$ is calculated for each of them from equation (3). Their number for a site of length $L = 20$ with the number of DNA (RNA) properties $Q = 30$ is $\approx QL^2/2$ or $\approx 10^5$.

Utility of a property for predicting site activity. Let us consider the operation of the ACTIVITY system for equation (2) with fixed $\langle Z, m, w\rangle$. For each sequence S_n with known activity F_n, the concentration $X_{Zmw}(S_n)$ of oligonucleotide Z of length m weighted by function $w(i)$ is calculated. As a result, we have pairs $\{X_{Zmw}(S_n), F_n\}$. When they meet the requirements of regression analysis [31], X_{Zmw} can be used to predict F_n. A simple regression is generated to test these requirements:

$$F_{Z,m,w}(S_n) = f_0 + f_1 X_{Z,m,w}(S_n), \qquad (4)$$

where f_0 and f_1 are regression coefficients calculated for pairs $\{X_{Zmw}(S_n), F_n\}$.

Equation (4) is used to predict site activity $F_{Zmw}(S_n)$ for each sequence of site S_n from the property $X_{Zmw}(S_n)$. The deviation $\Delta_n = F_{Zmw}(S_n) - F_n$ of the predicted from the experimental activity is also calculated. The obtained values $\{F_{Zmw}(S_n), F_n, \Delta_n\}$ are tested for meeting the 11 requirements of regression analysis [31]: four (linear, sign, and two range) correlations between $F_{Zmw}(S_n)$ and F_n; four tests for even distribution of $F_{Zmw}(S_n)$ and F_n; and three tests for Δ_n deviation values. To minimize the influence of heterogeneity in the set of tested values $\{F_{Zmw}(S_n), F_n, \Delta_n\}$, this set was divided into two disjoint samples of equal volume containing lower $\{F_{Zmw}(S_n), F_n \text{ and } \Delta_n\}_1$ and higher $\{F_{Zmw}(S_n), F_n \text{ and } \Delta_n\}_2$ predicted activities $F_{Zmw}(S_n)$. Each of these samples is independently tested for meeting the 11 requirements. At the same time, the corresponding statistical test [32] is used to estimate significance α_{rt}, when the rth requirement $(1 \le r \le 11)$ is met for tth subsample $(1 \le t \le 2)$. On the basis of α_{rt} for the tested property X_{Zmw}, the value $u_{rt}(X_{Zmw}, F)$ is calculated, which is called "utility of property X_{Zmw} for predicting site F activity" in terms of fuzzy sets of Zadeh [33] and the utility theory for decision making [34]:

$$u_{rt}(X_{Zmw}, F) = \begin{cases} 1, & \text{if } \alpha_{rt} < 0.01; \\ 1.3 - 28.3\alpha_{rt} + 55.6\alpha_{rt}^2, & \\ \text{if } 0.01 \le \alpha_{rt} \le 0.1; & \\ -1, & \text{if } \alpha_{rt} > 0.1. \end{cases} \qquad (5)$$

Equation (5) assigns the highest value $u_{rt}(X_{Zmw}, F) = 1$ to property X_{Zmw} if it conforms to rth requirement in tth sample with $\alpha_{rt} < 0.01$. The lowest value $u_{rt}(X_{Zmw}, F) = -1$ is assigned to property X_{Zmw} if it does not conform to rth requirement in tth sample with $\alpha_{rt} > 0$. Intermediate value of $u_{rt}(X_{Zmw}, F)$ between –1 and 1 is assigned to property X_{Zmw} if $0.01 \le \alpha_{rt} \le 0.1$. After testing all 11 requirements in the two subsamples, property X_{Zmw} is assigned $11 \times 2 = 22$ utility values $u_{rt}(X_{Zmw}, F)$. Their mean value is the "integral utility of property X_{Zmw} for predicting F activity" [34]:

$$U(X_{Z,m,w}, F) = \frac{\sum_{t=1}^{2}\sum_{r=1}^{11} u_{rt}(X_{Z,m,w}, F)}{22}. \qquad (6)$$

If the majority of the regression analysis requirements [31] are met, equation (6) assigns positive utility $U(X_{Zmw}, F) > 0$ for predicting activity F to the tested property X_{Zmw} calculated from equation (2). The value $U(X_{Zmw}, F)$ increases with the number of satisfied requirements.

Note that at least three site sequences with known activity are required to assess the significance of correlations in each of the two tested subsamples [31]. Hence equations (4)–(6) can be used for analysis of sites with at least six sequences with known activity. In case when no less than 12 such sequences are analyzed for site, in order to decrease the dependence of the result on the initial training sample, ACTIVITY additionally generates three 50%-volume training samples, calculates additional utilities $\{U_{q,50\%}(X_{Zmw}, F)\}_{1 \le q \le 3}$ for each property X_{Zmw} for the additional samples using equation (6), and, by analogy to equation (6), additionally averages all the utility values obtained, to produce its final utility $U(X_{Zmw}, F)$.

Revealing significant context site properties. We assessed utility $U(X_{Zmw}, F)$ accounting for all possible

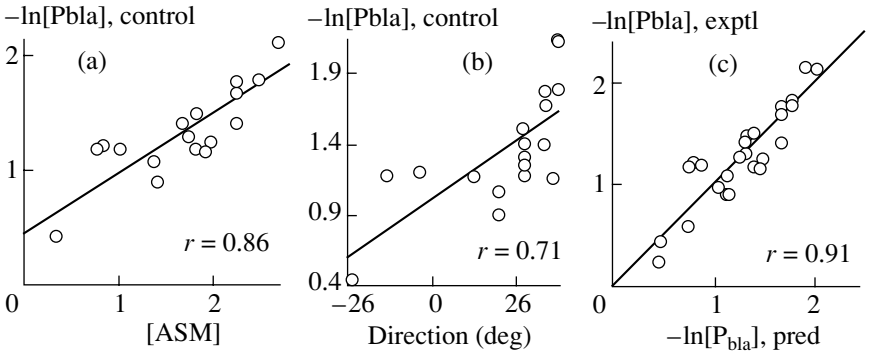

Fig. 5. The results of studying *E. coli* promoter strengths: control test of ASM trinucleotide weighted concentration (a) and **Direction** angle (b); comparison of predicted and experimental strengths of all 27 promoters (c).

properties X_{Zmw} using equation (6), and selected those conforming to condition

$$U(X_{Zmw}, F) > 0. \quad (7)$$

Selection of property X_{Zmw} by condition (7) requires that it should meet at least 11 out of 22 tested requirements at $\alpha < 0.01$. The highest probability of arbitrary selection can be assessed by the binomial distribution

$$p(f, g, v) = \sum_{h=g}^{v} C_v^h f^h [1-f]^{v-h}, \quad (8)$$

where f is frequency of arbitrary appearance of property X_{Zkw} conforming to one requirement, v is the number of tested requirements, g is the number of met requirements.

Equation (8) for $v = 22$, $g = 11$, and $f = 0.01$ results in $p < 10^{-16}$. Since ACTIVITY analyzes 10^7 properties X_{Zmw}, the probability of arbitrary selection of property X_{Zmw} with utility $U(X_{Zmw}, F) > 0$ is below $10^7 \times 10^{-16} = 10^{-9}$. Stated differently, each property X_{Zmw} with utility $U(X_{Zmw}, F) > 0$ is significant for site activity at $p < 10^{-9}$.

After selecting all the significant properties X_{Zmw} with utility $U(X_{Zmw}, F) > 0$, those relating with more utile properties are excluded. This results in a limited set of linear-independent site properties X_{Zmw} reliably correlating with this site activity and not correlating with other selected properties. Significant conformational and physicochemical properties X_{qab} are selected in a similar way [equation (3)]. All selected properties $\{X_1, \ldots X_k, \ldots, X_K\}$ are ordered $\{U(X_1, P) > \ldots > U(X_k, P) > \ldots > U(X_K, P) >\}$ and stored in the knowledge base (Fig. 1).

Figure 5 demonstrates the ACTIVITY-generated analysis of experimental data on the strength of 27 *E. coli* promoters described in Fig. 2. A training sample containing nine out of 27 promoters (L/N25DSR, D/E20, L, N25, G25, J5, N25/lac, con, and con/anti) was analyzed. The weighted concentration of ASM trinucleotides proved to be most significant for promoter strength. Weighted function $w(i)$ of this property is shown in Fig. 3a. This function assigns high significance $w(i) > 0.5$ to positions i from –1 to 11 relative to the transcription start. This means that this region is the most important for ASM trinucleotide contribution to the promoter strength. The utility of this property is $U = 0.59$. Figure 5a demonstrates the correspondence between weighted concentration of ASM and strength $\ln[P_{bla}]$ for the other 18 control promoters not included in the training sample. One can see that the weighted concentration of ASM reliably correlates with the promoter strength for the control data ($r = 0.86$, $\alpha < 10^{-3}$).

The same training sample including nine promoters allowed us to reveal a significant conformational property, **Direction** angle (Fig. 2) averaged for the [–4, 16] region around the transcription start (U = 0.50). Reliable correlation between **Direction** and the promoter strength ($r = 0.71$, $\alpha < 10^{-2}$) for the control data is demonstrated in Fig. 5b.

Significant site properties are stored in the special knowledge base (Fig. 6). In the case of weighted concentration of ASM trinucleotides, it includes the property type (CT field), description (CD field), ASM trinucleotide (CO field), utility (CU field), and values of significance function $w(i)$ at position i of the promoter (CW and CI fields, respectively). In the case of the mean **Direction** angle at the [–4; 16] promoter region it also includes the most significant information.

Building the method for activity prediction. The canonical equation (1) is optimized by multiple linear regression on the basis of the selected properties for predicting site activity. Such optimization for predicting *E. coli* promoter strength from its sequence using

```
MT  KNOWLEDGE BASE "ACTIVITY"
//
MN  Escherichia coli promoter strength
//------------------------------------------------------
CF  SEQUENCE-DEPENDENT FEATURE
CT  Statistical
CD  Weight(ASM)
CU  0.59
CO  ASM
CW  0.10, 0.10, 0.10, 0.10, 0.10, 0.10, 0.10, 0.10, 0.10, 0.10,
CI  -48   -47   -46   -45   -44   -43   -42   -41   -40   -39
CW  0.10, 0.10, 0.10, 0.10, 0.10, 0.10, 0.10, 0.10, 0.10, 0.10,
CI  -38   -37   -36   -35   -34   -33   -32   -31   -30   -29
CW  0.10, 0.10, 0.10, 0.10, 0.11, 0.11, 0.11, 0.12, 0.12, 0.12,
CI  -28   -27   -26   -25   -24   -23   -22   -21   -20   -19
CW  0.13, 0.14, 0.14, 0.15, 0.16, 0.17, 0.19, 0.20, 0.22, 0.24,
CI  -18   -17   -16   -15   -14   -13   -12   -11   -10    -9
CW  0.26, 0.29, 0.32, 0.36, 0.39, 0.43, 0.47, 0.52, 0.57, 0.63,
CI   -8    -7    -6    -5    -4    -3    -2    -1     0     1
CW  0.68, 0.76, 0.82, 0.88, 0.93, 0.98, 0.98, 0.95, 0.86, 0.73,
CI    2     3     4     5     6     7     8     9    10    11
CW  0.53, 0.36, 0.21, 0.14, 0.10, 0.10
CI   12    13    14    15    16    17
//------------------------------------------------------
CF  CONTEXTUAL FEATURE
CT  Conformational
CD  Mean(Direction)
CU  0.50
CR  -4 16
PN  Direction
CP  -154.,    0.,    2.,  143.,    0.,  154.,   64., -120.,
CC    AA     AT     AG     AC     TA     TT     TG     TC
CP   120., -143.,   57.,  180.,  -64.,   -2.,    0.,  -57.
CC    GA     GT     GG     GC     CA     CT     CG     CC
//------------------------------------------------------
CF  PREDICTION ACTIVITY
CT  Multiple Linear Regression
CL  0.91
CD  Weight(ASM)
CD  Mean(Direction)
CF  F= 0.3 + 0.6* Weight(ASM) + 0.0008* Mean(Direction)
//
```

Fig. 6. Presentation of the revealed DNA and RNA properties in the knowledge base of significant site properties (for *E. coli* promoter strength). Upper part, weighted concentration of ASM trinucleotide: type (CC), description (CD), ASM trinucleotide (CO), utility (CU), and significance function $w(i)$ at position i of the promoter (CW and CI, respectively); central part, **Direction** angle: similar CT, CD, and CU fields, site region for averaging the property (CR), and its values for each dinucleotide (CP and CC); lower part presents equation (1) for predicting *E. coli* promoter strength from weighted concentration of ASM trinucleotide and **Direction** angle (CF).

the weighted concentration of ASM trinucleotides and the mean **Direction** angle is exemplified in Fig. 6. Figure 5c demonstrates reliable correspondence between the predicted and experimental values of promoter strength ($r = 0.91$, $\alpha < 10^{-6}$).

The source code of the program for calculating site activity $F(S)$ for arbitrary sequence S is automatically generated by ACTIVITY for the optimized equation (1). It is stored in the library of programs for predicting site activity and is available via Internet.

Table 2. Examples of functional sites included in the *Activity* database on DNA and RNA site activity

No.	Site			Activity				Ref.
	name	*n*	variants	designation	scale	min	max	
1	*Cro*-binding site of *E. coli*	7	Natural	Association rate constant, k_A	ln	19.1	19.9	[35]
2	*Cro*-binding site of *E. coli*	7	Natural	Dissociation rate constant, k_D	ln	19.1	19.9	[35]
3	*Cro*-binding site of *E. coli*	7	Natural	*Cro*–DNA affinity	ln	22.9	27.5	[35]
4	CRP-binding site of *E. coli*	10	Natural	CRP–DNA affinity	ln	–3.2	3.2	[23]
5	*E. coli* promoters	27	Natural	Promoter strength	–log	0.26	2.1	[14]
6	Eukaryotic TATA box	8	Natural	hTFIID–DNA affinity	ln	–3.0	0.0	[36]
7	Eukaryotic TATA box	9	Mutant	Lifetime	min	1	185	[37]
8	Eukaryotic TATA box	9	Mutant	DNA bend	(°)	33	106	[37]
9	Eukaryotic TATA box	19	Synthetic	yTBP–DNA affinity	–ln	11.8	24.2	[38]
10	Eukaryotic *INR* element	115	Mutant	Pre-mRNA yield	ln	–5.3	0.9	[39]
11	Eukaryotic *Oct-1* element	10	Mutant	Pre-mRNA yield	ln	–2.3	0.63	[40]
12	Eukaryotic *USF* element	14	Synthetic	USF–DNA affinity	ln	3.8	100	[41]
13	Mouse *PE1B* element	10	Mutant	Pre-mRNA yield	ln	–1.4	1.4	[42]
14	Eukaryotic *IL-1* element	18	Mutant	Pre-mRNA yield	ln	–1.9	4.1	[43]
15	Pre-mRNA 3'-processing site	16	Mutant	mRNA yield	%	3	289	[44]
16	Pre-mRNA splicing site	22	Mutant	mRNA yield	%	18	100	[45]
17	*E. coli* ribosome binding site	185	Synthetic	Protein yield	ln	0.0	8.06	[13]
18	2AP-induced mutations	26	Natural	Mutation frequency	ln	0.0	5.6	[46]

Note: *n*, number of variants; *Cro*, *E. coli Cro*-repressor; CRP, *E. coli* catabolite activator protein; 2AP, 2-aminopurine; hTFIID, human transcription factor IID; yTBP, yeast TATA-binding protein.

STUDY OF ACTIVITY OF FUNCTIONAL DNA AND RNA SITES

During formation of the ACTIVITY system, 70 samples of sites with experimentally measured specific activities were included in the database. Such samples are exemplified in Table 2. The number of sequences in a database varies from 7 to 185, being 20 on average, while their total number in the database is over 1500. The database includes sequences of natural sites with different locations in the genomic DNA (RNA) [14, 23, 35, 36, 46], the sites generated by directed or random mutagenesis [37, 39, 40, 42–45], and synthesized sequences [13, 38, 41]. Specific activity of sites is expressed as kinetic and thermodynamic constants of the site–protein complex [23, 35, 36, 38, 41]; lifetime [37]; rate of a gene product synthesis, e.g., pre-mRNA (for transcription-regulating sites [39, 40, 42, 43]), processed mRNA (for 3'-terminal processing [44] and splicing [45] sites), or protein (translation initiation sites [13]). In addition, other quantitative site properties depending on their nucleotide context are determined, e.g., the DNA bending angle [37] and the frequency of mutations induced by a particular mutagen [46]. Table 2 demonstrates that the activity levels of sites of the same type located in different regions of DNA (RNA) may differ by several orders of magnitude. For instance, *E. coli* promoter strength and affinity of *E. coli* operators for *Cro* repressor differ by three [14] and two [35] orders of magnitude, respectively; the affinity of TATA boxes for human TFIID differs 20-fold [36].

Significant statistical, physicochemical, and conformational properties were revealed for the sites included in the database, and methods for predicting their activity were built. Table 3 exemplifies such properties and methods for predicting the activity for certain sites. Let us consider some of them in detail.

Two significant statistical properties were revealed for 19 synthetic analogs of TATA boxes with known affinity for TBP: (1) weighted concentration of TV dinucleotide with the highest contribution to affinity for TBP in the site center (Fig. 3b), and (2) weighted concentration of WR dinucleotide with the highest significance at the site ends (Fig. 3c). Utility U of these properties (0.35 and 0.41, respectively) is above the threshold $U_0 = 0$, which allowed them to be selected for predicting the affinity for TBP. The correlation coefficients between these properties and TBP–TATA affinity are 0.73 and 0.76, respectively ($\alpha < 0.01$). Testing of the method of predicting the affinity

Table 3. Examples of significant site context properties included in the ACTIVITY database

Site				Property			Significance		
name	position 1	n	activity, F	X_k	region	property	U	r	α
E. coli promoters	Transcription start	27	Promoter strength, $-\log(P_{bla})$	X_1	Fig. 3a	[ASM]	0.59	0.86	10^{-2}
				X_2	–4; 14	Direction	0.50	0.71	10^{-2}
				$F = 0.3 + 0.6 \times X_1 + 0.0008 \times X_2$				0.91	10^{-4}
TATA box (synthetic oligoDNA)	First nucleotide of the site	19	yTBP–DNA affinity	X_1	Fig. 3b	[TV]	0.35	0.73	10^{-2}
				X_2	Fig. 3c	[WR]	0.41	0.76	10^{-2}
				$F = 14.5 + 2.5 \times X_1 + 0.9 \times X_2$				0.77	10^{-2}
TATA box (mutant)	Transcription start	9	DNA bending in TBP–TATA	X_1	0; 9	Inclination	0.19	0.76	0.05
				$F = 120.15 + 70.32 \times X_1$				0.76	0.05
PE1B of mouse αA-crystallin gene promoter	Transcription start	10	Pre-mRNA yield	X_1	–32; –25	P_{nucl}	0.36	–0.77	10^{-2}
				X_2	–29; –19	DIST	0.41	0.86	10^{-3}
				X_3	–31; –25	Tilt	0.38	–0.78	10^{-2}
				$F = -39 - 0.1 \times X_1 + 12 \times X_2 - X_3$				0.89	10^{-4}
USF element of eukaryotic promoters	First nucleotide of the site	14	USF–DNA affinity	X_1	11; 15	Depth	0.22	–0.78	10^{-3}
				X_2	11; 20	Twist	0.23	–0.86	10^{-4}
				$F = 170 - 16.3 \times X_1 - 0.7 \times X_2$				0.91	10^{-5}
SV40 pre-mRNA 3'-processing site	Cutting point	16	mRNA yield	X_1	Fig. 3a	[VUKK]	0.24	0.88	10^{-4}
				$F = -301.72 + 216.16 \times X_1$				0.88	10^{-4}
2-Aminopurine-induced mutations	Mutation point	26	Mutation frequency	X_1	–1; 2	T_{melt}	0.20	0.90	10^{-5}
				$F = -8.5568 + 0.1585 \times X_1$				0.90	10^{-5}
E. coli Cro-binding site	First nucleotide of the site	7	*Cro*–DNA affinity	X_1	1; 16	WIDTH	0.55	0.97	10^{-3}
				X_2	6; 19	Roll	0.44	0.90	10^{-3}
				X_3	6; 19	Rise	0.41	0.92	10^{-2}
				$F = -72 + 4 \times X_1 + X_2 + 13 \times X_3$				0.99	10^{-5}
E. coli CRP-binding site	Center of repeat in the site	10	CRP–DNA affinity	X_1	–15; 14	Rise	0.15	–0.86	10^{-2}
				X_2	–17; 12	Width	0.06	0.78	10^{-2}
				$F = 190 - 66.8 \times X_1 + 7.5 \times X_2$				0.87	10^{-2}

Note: n, number of site variants; X_k, revealed context-dependent property significant for site activity; Fig. 3, weight functions $w(i)$ of site position significance used for calculating weighted oligonucleotide concentrations from equation (2); $F = F_0 + S_{k-1,K} F_k \times X_k$, canonical equation (1) optimized for predicting site activity using the revealed significant properties.

for TBP based on these properties has demonstrated a reliable correlation between the predicted and experimental affinities of DNA for TBP (Fig. 7a: $r = 0.77$, $\alpha < 0.01$).

Analysis of the sample [37] containing nine sequences with known DNA bend within the TBP–TATA complex has shown that this bend increases with the free DNA **Inclination** angle (Fig. 7b: $r = 0.76$, $\alpha < 0.05$). This result agrees with the experimentally determined structure of the TBP–TATA complexes (reviewed in [25]) demonstrating that DNA bending results from intercalation of four TBP phenylalanine residues between neighboring base pairs from the minor groove side of TATA box. **Inclination** is known to describe the rotation angle of a base pair along its short axis, and increased **Inclination** widens the minor groove [30], which favors the phenylalanines intercalation from the minor groove side and the subsequent DNA bending.

We studied the sample [42] containing 10 sequences of mouse αA-crystallin promoters generated by mutagenesis and containing neighboring TATA box and PE1B signal. The highest utility ($U = 0.36$) of all analyzed physicochemical properties of this region was observed for the frequency of contacts between the component dinucleotides and nucleosome core proteins ($r = –0.77$, $\alpha < 0.01$). The correlation between this property and transcription activity is

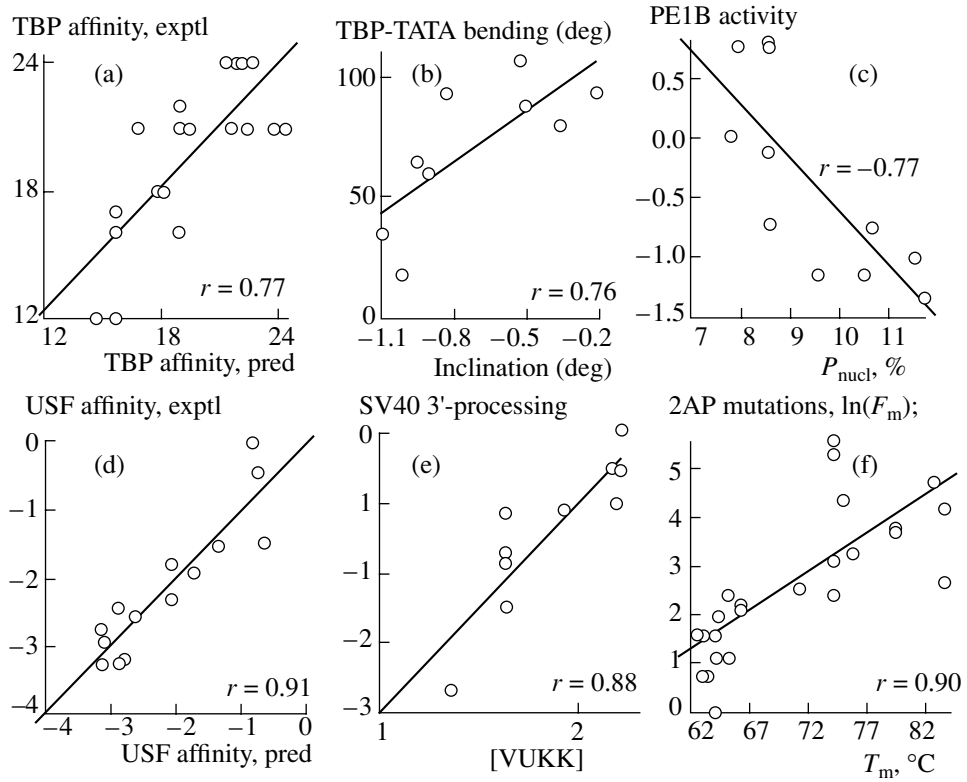

Fig. 7. Examples of the revealed properties significant for site activity: (a) synthetic TATA-box–TBP affinity [38]; (b) DNA bending in the TBP–TATA complex [37]; (c) *in vivo* transcriptional activity of the PE1B region of mouse αA-crystallin promoter [42]; (d) synthetic DNA affinity for USF [41]; (e) mRNA yield during 3'-terminal processing of SV40 pre-mRNA [44]; (f) frequency of 2-aminopurine-induced mutations [46].

negative (Fig. 7c), indicating that the higher is the interaction between this region and nucleosomes, the lower is the αA-crystallin transcription. This conclusion agrees with the experimental data [47, 48] demonstrating that TBP–TATA binding is preceded by nucleosome release from the TATA-containing promoter region. We have also shown [49] that TATA-containing promoter regions feature a lower twist of B-DNA as compared with nucleosomal binding sites.

The size of the major groove DIST and the angle between the neighboring basepair planes along their short axis (tilt) also proved to be significant for transcription activity of the TATA/PE1B-containing region of mouse αA-crystallin gene promoter (Table 3: $U = 0.41$ and $U = 0.38$, respectively). All the exposed significant properties were used to build the methods for predicting the transcription activity of the mouse αA-crystallin gene from the sequence of TATA/PE1B-containing promoter region ($r = 0.89$, $\alpha < 10^{-4}$).

The depth of the minor groove ($r = -0.78$, $\alpha < 10^{-3}$) and the DNA helix twist ($r = -0.86$, $\alpha < 10^{-4}$) proved to be the most significant for binding between synthetic DNA and USF protein. Both properties allow reliable prediction of USF–DNA affinity (Fig. 7d).

The weighted concentration of VUKK tetranucleotide to the right of the 3'-terminal pre-mRNA processing point of SV40 proved to be the only significant property for mRNA production (Fig. 7e: $r = 0.88$, $\alpha < 10^{-4}$). This is a *G*/*U*-rich oligonucleotide (since $V = \{A, C, G\}$ and $K = \{U, G\}$), which agrees with the known observation that the sites of 3'-terminal pre-mRNA processing have a *G*/*U*-rich region subsequent to the cutting point [44].

The proposed approach proved to be applicable to a wide range of experimental data on "nucleotide sequences vs. their specific activity." For instance, it was used to show that the frequency of $C \rightarrow T$ mutation induction by 2-aminopurine [46] depends on the melting temperature of DNA around the mutation point (Fig. 7f).

CONCLUSION

Note that in each case ACTIVITY analyzes the information obtained under specific experimental conditions. Hence, it reveals the properties of sites significant for these experimental conditions and generates a method predicting activity only for these conditions. One of ACTIVITY advantages for studying the functional sites is its accessibility via Internet,

fully automated functioning, and the minimal possible volume of experimental data [31]. The data obtained demonstrate the applicability of the proposed approach to a wide range of experimental data on site activity. Further development of the approach will go by accumulation of experimental data on the functional site activity, extending the range of conformational, statistical, and physicochemical properties of DNA and RNA, and complementing the linear additive model of activity prediction by more complex ones accounting for interrelation of the properties significant for site activity.

ACKNOWLEDGMENTS

This work was supported by Russian Foundation for Basic Research (projects 95-04-12469, 96-04-49957, 96-04-50006, 97-04-49740, and 97-07-90309), Russian program Human Genome (project 12312 GCh-5), Russian State Committee for Science and Technology, and the Integration Program of the Siberian Division of the Russian Academy of Sciences.

REFERENCES

1. Neidle, S., *DNA Structure and Recognition*, Oxford: IRL Press, 1994.
2. Kolchanov, N.A., *Mol. Biol*., 1997, vol. 31, pp. 581–583.
3. Ignat'eva, E.V., Merkulova, T.I., Vishnevskii, O.V., and Kel', A.E., *Mol. Biol*., 1997, vol. 31, pp. 684–700.
4. Kel', O.V., Kel', A.E., Romashchenko, A.G., Vingender, E., and Kolchanov, N.A., *Mol. Biol*., 1997, vol. 31, pp. 601–615.
5. Bukher, F., *Mol. Biol*., 1997, vol. 31, pp. 616–625.
6. Kel', A.E., Kolchanov, N.A., Kel', O.V., Romashchenko, A.G., Anan'ko, E.A., Ignat'eva, E.V., Merkulova, T.I., Podkolodnaya, O.A., Stepanenko, I.L., Kochetov, A.V., Kolpakov, F.A., Podkolodnyi, N.L., and Naumochkin, A.A., *Mol. Biol*., 1997, vol. 31, pp. 626–636.
7. Gelfand, M.S., *J. Comp. Biol*., 1995, vol. 2, pp. 87–115.
8. Berg, O.G. and von Hippel, P.H., *J. Mol. Biol*., 1987, vol. 193, pp. 723–750.
9. Podkolodnaya, O.A. and Stepanenko, I.L., *Mol. Biol.*, 1997, vol. 31, pp. 671–683.
10. Mulligan, M.E., Hawley, D.K., Entriken, R., and McClure, W.R., *Nucleic Acids Res*., 1984, vol. 12, pp. 789–800.
11. Stormo, G.D., Schneider, T.D., and Gold, L., *Nucleic Acids Res.*, 1986, vol. 14, pp. 6661–6679.
12. Berg, O.G. and von Hippel, P.H., *J. Mol. Biol.*, 1988, vol. 200, pp. 709–723.
13. Barrick, D., Villanueba, K., Childs, J., Kalil, R., Schneider, T.D., Lawrence, C.E., Gold, L., and Stormo, G.D., *Nucleic Acids Res*., 1994, vol. 22, pp. 1287–1295.
14. Jonsson, J., Norberg, T., Carlsson, L., Gustafsson, C., and Wold, S., *Nucleic Acids Res*., 1993, vol. 21, pp. 733–739.
15. Kraus, R.J., Murray, E.E., Wiley, S.R., Zink, N.M., Loritz, K., Gelembiuk, G.W., and Mertz, J.E., *Nucl. Acids Res.*, 1996, vol. 24, pp. 1531–1539.
16. Ponomarenko, M.P., Kolchanova, A.N., and Kolchanov, N.A., *J. Comput. Biol*., 1997, vol. 4, pp. 83–90.
17. Ponomarenko, M.P., Savinkova, L.K., Ponomarenko, Yu.V., Kel', A.E., Titov, I.I., and Kolchanov, N.A., *Mol. Biol*., 1997, vol. 31, pp. 726–732.
18. Kel, A.E., Ponomarenko, M.P., Likhachev, E.A., Orlov, Y.L., Ischenko, I.V., Milanesi, L., and Kolchanov, N.A., *Comput. Appl. Biosci*., 1993, vol. 9, pp. 617–627.
19. Ponomarenko, M.P., Ponomarenko, J.V., Kel, A.E., and Kolchanov, N.A., *Proc. 1997 Pacific Symp.*, Altman, R., Ed., Singapore: World Sci., 1996, pp. 340–351.
20. Ponomarenko, M.P., Kel, A.E., Orlov, Y.L., Benjukh, D.N., Ischenko, I.V., Bokhonov, V.B., Likhachev, E.A., and Kolchanov, N.A., *Computer Analysis of Genetic Macromolecules: Structure, Function and Evolution*, Kolchanov, N. and Lim, H., Eds., Singapore: World Sci., 1994, pp. 35–65.
21. Etzold, T. and Argos, P., *Comput. Appl. Biosci*., 1993, vol. 9, pp. 49–57.
22. Hogan, M.E. and Austin, R.H., *Nature*, 1987, vol. 329, pp. 263–266.
23. Gartenberg, M.R. and Crothers, D.M., *Nature*, 1988, vol. 333, pp. 824–829.
24. Sugimoto, N., Nakano, S., Yoneyama, M., and Honda, K., *Nucleic Acids Res*., 1996, vol. 24, pp. 4501–4505.
25. Juo, Z.S., Chiu, T.K., Leiberman, P.M., Baikalov, I., Berk, A.J., and Dickerson, R.E., *J. Mol. Biol*., 1996, vol. 261, pp. 239–254.
26. Karas, H., Knuppel, R., Schulz, W., Sklenar, H., and Wingender, E., *Comput. Appl. Biosci*., 1996, vol. 12, pp. 441–446.
27. Suzuki, M., Yagi, N., and Finch, J.T., *FEBS L*., 1996, vol. 397, pp. 148–152.
28. Shpigelman, E.S., Trifonov, E.N., and Bolshoy, A., *Comput. Appl. Biosci*., 1993, vol. 9, pp. 435–440.
29. Gorin, A.A., Zhurkin, V.B., and Olson, W.K., *J. Mol. Biol.*, 1995, vol. 247, pp. 34–48.
30. EMBO Workshop, *EMBO J*., 1989, vol. 8, pp. 1–5.
31. Forster, E. and Ronr, B., *Methoden der Korrelations und Regressions Analyse*, Berlin: Verlag Die Wirtschaft, 1979.
32. Leman, E., *Proverka statisticheskikh gipotez* (Verification of Statistical Hypotheses), Moscow: Nauka, 1979.
33. Zadeh, L.A., *Information and Control*, 1965, vol. 8, pp. 338–353.
34. Fishburn, P.C., *Utility Theory for Decision Making*, New York: Wiley, 1970.
35. Kim, J.G., Takeda, Y., Matthews, B.W., and Anderson, W.F., *J. Mol. Biol*., 1987, vol. 196, pp. 149–158.
36. Wiley, S.R., Kraus, R.J., and Mertz, J.E., *Proc. Natl. Acad. Sci. USA*, 1992, vol. 89, pp. 5814–5818.
37. Starr, D.B., Hoopes, B.C., and Hawley, D.K., *J. Mol. Biol.*, 1995, vol. 250, pp. 434–446.
38. Sokolenko, A.A., Sadomirskii, I.I., and Savinkova, L.K., *Mol. Biol*., 1996, vol. 30, pp. 279–285.

39. Javahery, R., Khachi, A., Lo, K., Zenzie-Gregory, B., and Smale, S.T., *Mol. Cell. Biol*., 1994, vol. 14, pp. 116–127.

40. Boyd, D.C., Turner, P.C., Watkins, N.J., Gerster, T., and Murphy, S., *J. Mol. Biol.*, 1995, vol. 253, pp. 677–690.

41. Bendall, A.J. and Molloy, P.L., *Nucl. Acids Res*., 1994, vol. 22, pp. 2801–2810.

42. Sax, C.M., Cvelk, A., Kantorow, M., Gopal-Srivastava, R., Iligan, J.G., Ambulos, N.P., and Piatigorsky, J., *Nucl. Acids Res*., 1995, vol. 23, pp. 442–451.

43. Kretsovali, A. and Papamatheakis, J., *Nucl. Acids Res.*, 1995, vol. 23, pp. 2919–2928.

44. McDevitt, M.A., Hart, R.P., Wong, W.W., and Nevins, J.R., *EMBO J.*, 1986, vol. 5, pp. 2907–2913.

45. Lesser, C.F. and Guthrie, C., *Genetics*, 1993, vol. 131, pp. 851–863.

46. Coulondre, C., Miller, J.H., Farabaugh, P.J., and Gilbert, W., *Nature*, 1978, vol. 274, pp. 775–780.

47. Godde, J.S., Nakatani, Y., and Wolffe, A.P., *Nucl. Acids Res.*, 1995, vol. 23, pp. 4557–4564.

48. Edmondson, D.G. and Roth, S.Y., *FASEB J.*, 1996, vol. 10, pp. 1173–1182.

49. Ponomarenko, M.P., Ponomarenko, Yu.V., Kel', A.E., Kolchanov, N.A., Karas, Kh., Vingender, E., and Sklenar, Kh., *Mol. Biol*., 1997, vol. 31, pp. 733–740.