The Study of the Solvability of the Genome Annotation Problem on Sets of Elementary Motifs

I. Yu. Torshin

Moscow Institute of Physics and Technology, Institutski per 9, Dolgoprudnyi, 141700 Russia
Russian Branch of the Trace Elements Institute for UNESCO, bol. Tishinski per. 26, Moscow, 109652 Russia

Abstract—The problem of genome annotation (i.e., the establishment of the biological roles of proteins and corresponding genes) is one of the major tasks of postgenomic bioinformatics. This paper reports the development of the previously proposed formalism for the study of the local solvability of the genome annotation problem. Here, we introduce the concepts of elementary motifs, positional independence of motifs, heuristic evaluation of informativeness, and solvability on the sets of elementary motifs. We show that introduction of a linear order in a set of elementary motifs allows us to calculate the irreducible motif sets. The formalism was used in experiments to compute the sets of the most informative motifs for several protein functions.

Keywords: algebraic approach to pattern recognition, irreducible set of motifs, data mining, genome annotation problem, post-genomic bioinformatics.

DOI: 10.1134/S1054661811040171

1. INTRODUCTION

After the determination of the nucleotide sequence of the human genome in 2001–2002, the need to develop efficient theoretical methods for solving the annotation problem became apparent. Annotation of genes is done through the annotation of proteins encoded by those genes. Existing bioinformatics methods are characterized by limited applicability, because they cannot annotate more than half of the proteins of the human genome [1].

Previously, the formalism to study the solvability and locality of the task of the annotation of the genome was proposed [2]. *Protein* P is determined as an element of a set $A^* = \bigcup_{l=1}^{\infty} A^l$, where A is an alphabet of amino acids, $A = \{$A, C, D, E, F, G, H, I, K, L, M, N, P, R, S, T, V, W, Y$\}$. *Annotation* $\mathbf{t}$ of protein P is determined as an element of a set $T^* = \bigcup_{l=1}^{m} T^l$, where $T = \{t_1, t_2, ..., t_m\}$ is a *terminology dictionary* reporting so-called biological functions (biological roles) of proteins. The set T^* bijectively images itself in *a set of Boolean vectors of annotations* $\tilde{T}$, whose elements $\{t_1(P), t_2(P), ..., t_j(P), ..., t_m(P)\}$ are such that $t_j(P) = 1$ if the term t_j belongs to the annotation of the protein P and 0, if it does not. The solution of the annotation problem is the correct algorithm, studied on *a set of precedents* $Pr \subseteq A^* \times \tilde{T}$ and assigning the given protein P in accordance with its annotation $\mathbf{t} = \{t_1, t_2, ..., t_j, ...\}$, $t_j \in T$, $\mathbf{t} \subset T^*$. Let us note that the term t_j generates a partition of the set of the precedents into two nonintersecting classes c^{t_j} and its inverse $\bar{c}^{t_j}$, such that $c^{t_j} \cup \bar{c}^{t_j} = Pr$, $c^{t_j} \cap \bar{c}^{t_j} = \varnothing$.

In terms of the development of so-called formalism, the annotation problem can be reduced without loss of accuracy to the construction of $m = |T|$ *correct local t-classifiers* f_j:

$$\bigvee_{T} t_j \bigvee_{Pr} P \exists_{i}^{|P|} i: f_j(\eta(i, \hat{m}_\Sigma(M), P)) = t_j(P), \qquad (1)$$

where Pr is a set of precedents; i is the position of the amino acid sequence P; $\eta(i, \hat{m}, U)$ is an operator for selection of the subsequence of word U on *a mask* $\hat{m} = \{\mu_1, \mu_2, ..., \mu_m\}$, $\mu_i \in Z$, and $\mu_1 < \mu_2 < ... < \mu_m$; $M = \{\hat{m}_1, \hat{m}_2, ..., \hat{m}_{|M|}\}$ is a system of masks; $\hat{m}_\Sigma(M) = \bigcup_{k=1}^{|M|} \hat{m}_k$ is a combined mask of system M; and $t_j(P) = (P \in c^{t_j})$. Let us denote the parameter m of the mask $\hat{m} = \{\mu_1, \mu_2, ..., \mu_m\}$ as $|\hat{m}|$ and call it the dimension of the mask $\hat{m}$; $[\hat{m}] = \mu_m - \mu_1 + 1$ is an extension of mask $\hat{m}$.

Received July 19, 2011

The local t_j-classifier f_j exists if and only if for given *Pr* and *M the condition of local solvability* is satisfied [2]:

$$\bigvee_{Pr}(P_1, t_j(P_1)), (P_2, t_j(P_2)), P_1 \neq P_2$$

$$\exists_{\substack{i = L\ldots|P_1|-R \\ k = L\ldots|P_2|-R}} (i,k)\left(\bigvee_{l=1}^{|M|} \hat{m}_l\colon \eta(i, \hat{m}_1, P_1) = \eta(k, \hat{m}_1, P_2)\right) \quad (2)$$

$$\Rightarrow t_j(P_1) = t_j(P_2),$$

where $L = L(M) = \max(-\min_{k=1,N} \mu_1^k, 0)$ and $R = R(M) = \max(\max_{k=1,N} \mu_{|m_k|}^k, 0)$ define the definitional domain of operator η. In the following discussion, we assume the existence of restrictions on $L(M)$ and $R(M)$. *Iirreducible* is such a system of masks M that condition (2) is violated for any $M' \subset M$.

Let us consider the possibility of using (2) for the experiments. We assume that *Pr* is consistent, i.e., *Pr*, $\bigvee_{Pr}(P_i, \mathbf{t}^i), (P_j, \mathbf{t}^j), i \neq j\colon (P_i = P_j) \Rightarrow (\mathbf{t}^i = \mathbf{t}^j)$. In contrast to the problem of recognition of a secondary structure of proteins, where the determination of a consistent set of precedents is a another research problem [3], the consistency of *Pr* in the annotation problem is provided by the regularity of any *Pr* built on a sample of proteins encoded by the same genome [2]. Therefore, a irreducible system of masks, which provides local solvability of the problem (or at least the boundary parameters of irreducible systems of masks) would be, above all, a nontrivial result of experiments for a given term t_j.

The practical applicability of the condition for the solvability of the form (2) is substantially limited (a) by the need to search all possible M, and (b) by significant loss of information when removing masks from M. Let M_n^m be a system of masks formed by all combinations of m out of n possible positions in the combined mask (i.e., m is the dimension of each mask M_n^m, and n is the length of the combined mask), so that $|M_n^m| = C_n^m$. In practically interesting cases $m = 3$–8 and $n = 8$–30, so that an exhaustive search of all $\sum_{m=3}^{8}\sum_{n=8}^{30}\sum_{j=1}^{C_n^m} C^j_{C_n^m}$ subsets M_n^m is not feasible. In addition, a removal of any mask out of M entails the removal of all subsequences $|A|^m$, generated by that mask. When $m = 3$–8, $|A| = 20$, $|A|^3 = 8000$, and $|A|^8 = 25.6 \times 10^9$, so that the significant loss of information on protein subsequences when deleting even a single mask becomes obvious.

These problems can be solved in terms of the classification of attribute values [4, 5], developed in the scientific school of Academician Zhuravlev [6–8]. In the transition from *the attributes* (masks of M or pairs "a mask is a position sequence") to *the values of attributes* (subsequences formed by a given mask in a certain position), the study of the solvability of the criterion of monotony allows one to operate with individual subsequences. The introduction of heuristic evaluations of informative subsequences [5] can reduce an exhaustive search for the solution and makes the practical implementation of algorithms for testing the solvability possible.

Table 1. Terms of the GO dictionary, most common in the annotation of human genome (total number of genes >25000, number of annotated genes is 14000)

Term	GO ID	Number of genes/proteins
t_1—"Nucleus"	5634	2890
t_2—"Membrane"	16020	2450
t_3—"Integral membrane protein"	16021	2400
t_4—"Protein—protein interactions"	5515	2390
t_5—"Binding of metal ions"	46872	1620
t_6—"Zinc binding"	8270	1580
t_7—"Transcription regulation"	6355	1430
t_8—"Binding of nucleotides"	166	1260
t_9—"Receptor activity"	4872	1170
t_{10}—"ATP binding"	5524	1050

It follows from condition (2), which includes operator $\eta(i, \hat{m}, U)$ for choosing a subsequence, that a particular feature of sequence U can be considered as mask $\hat{m}$, as well as a complex of the mask and the position of the sequence, $(i, \hat{m})$. Before the further development of formalism, it is necessary to choose the most appropriate method for the generation of attributes in terms of data available in the area of concern. For this, let us we consider the specific subsequences of amino acids known in biology as "functional sites" or "amino acid motifs," which correspond to specific biological roles of proteins.

2. BIOLOGICAL ROLES OF PROTEINS, FUNCTIONAL SITES, AND AMINO ACID MOTIFS

In this paper, the biological roles of proteins are described using *terminology dictionary* T, in which capacity, for example, the system of standardized terminology GO (Gene Ontology) [9] can be used. To

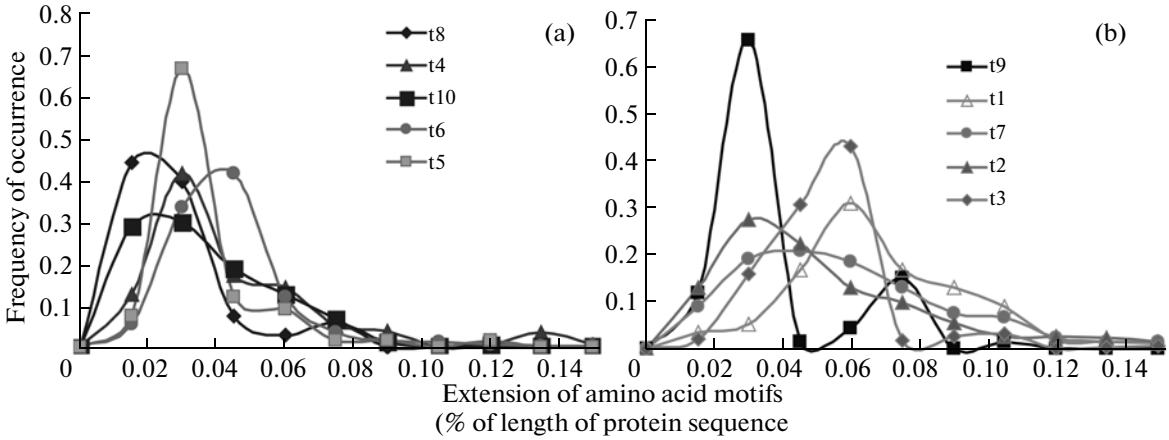

Fig. 1. Frequency of values of the relative extension of sites formed by amino acid motifs for various terms of the GO dictionary (DB PROSITE data).

date, the DB GO (www.geneontology.org) has more than 23 100 terms; the most common terms are listed in Table 1.

The particular biological role or "biological function" of protein P is realized by a complex of specific amino acid residues in the protein sequence. These residues form "sites," which are certain areas of a three-dimensional structure of the protein, which ensure the fulfillment of that biological role of the protein.

Within the developed formalism, such a site, uniquely corresponding to a particular biological role of the protein described by term t_j, is a subsequence of protein P selected in the ith position P on a mask $\hat{m}$, so that $|\hat{m}|$ is *the dimension of the site* $S_j(P) = \eta(i, \hat{m}, P)$, $[\hat{m}]$ *is the extension of the site*, *and the ratio* $[\hat{m}]/|P|$ *is the relative extension of the site*, which reflects the degree of localization of the site in protein P. The biological function of the protein is *localized*, if the maximum extension of the appropriate site is much smaller than the amino acid sequence of the protein [2], so that *the degree of localization* of the biological function t in protein P is $\mathrm{loc}(t, P) = 1 - \max[\hat{m}]/|P|$.

To describe the sites of such a specific type, the so-called amino acid motifs are used in biology: they are specific (usually relatively short: 8–30 letters) sequences, "patterns" of amino acids. For example, one of the motifs, corresponding with the term t_8 "nucleotide binding" (GO code 166) is [AG]-X(4)-GK-[ST], where "[AG]" means "A or G," X is any letter of the alphabet A, and $X(4)$ is a sequence of four letters. The database PROSITE (PROtein SITEs) [10], compiled by experts on the basis of biochemistry and molecular biology data, contains more than 1800 such motifs of amino acid sequences.

We note that each term t_j corresponds to several different amino acid motifs. Thus, term t_8 corresponds to not only the above-mentioned motif [AG]-X(4)-G-K-[ST] (ID PS00017 in the PROSITE database) but also the motifs [LIVM]-X-[LIVM](2)-[HEA]-[TI]-X-D-X-H-[GSA]-X-[LIVMF] (PS00785), [FYPH]-X(4)-[LIVM]-G-N-H-E-F-[DN] (PS00786] and others. The terms in Table 1 correspond to 198 amino acid motifs in the PROSITE database. These motifs differ by the extension and the degree of the localization of the generated sites, as well as by the location in the amino acid sequences. In Fig. 1, the data on the extension of the amino acid motifs corresponding to the terms in Table 1 is summarized; the data on the relative positions of these motifs in different proteins of the human genome can be seen in Fig. 2.

The data summarized in Fig. 1 is a good illustration of the desirability of introducing the hypothesis of locality. Thus, the amino acid motifs for the terms t_4, t_5, t_6, t_8, and t_{10} take, as a rule, 3–5% of the length of the sequence in all investigated sequences of proteins, and amino acid motifs for t_1, t_2, t_3, t_7, and t_9 take 3–8%. In this case, the length of all investigated motifs is rarely more than 10% of the length of the protein sequence.

The data in Fig. 2 allow us to conclude that the distribution of amino acid motifs along the amino acid sequences is in some way irregular. For example, the motifs corresponding to t_1 and t_{10}, occur mainly at the beginning of the amino acid sequences (the first 20–30% of the positions of the sequence, or loci 0.2–0.3 in Fig. 2); t_3 is mainly in the centers of the sequences (loci 0.3–0.5); t_9 is at the ends of sequences (loci 0.7–0.8); and motifs t_7 meet with similar frequencies (on the average, 0.1) in different parts of the chain. It is important to note that despite the irregular distribution along the sequence, the considered amino acid motifs are found in almost all parts of the sequences, although with varying frequency. Similar results were obtained for other motifs collected in the PROSITE database.

Thus, the analysis of the known amino acid motifs in biology points to the practicability of considering the two hypotheses in the search for the solution to the annotation problem: (A) the hypothesis of locality (the biological function described by the term t_j is realized by a relatively short section of the sequence) and (B) the hypothesis of positional independence (the motif corresponding to t_j can be found in any part of the sequence. These two hypotheses are the basis for further progress of the developed formalism.

3. THE SOLVABILITY CONDITION ON THE SET OF ELEMENTARY MOTIFS

From the available data on amino acid motifs, it follows that each term $t_j \in T$ corresponds, in general, to several motifs of various dimensions and extensions. Previously introduced concepts for describing the annotation problem in a local form (a mask, an η operator, a mask system, a combined mask of the mask system) allow us to represent an arbitrary site $S_j(P)$ by the complex of the sets and the subsequences of smaller dimension and extension and then reformulate the condition of local solvability in terms of some subsequences of symbols, which are available or missing in protein P.

Let us call element of $\mathbf{K} = \{(\hat{m}, V) | \hat{m} \in M, |V| = |\hat{m}|\}$ as *the elementary motif* κ. In terms of the hypothesis about the positional independence of the motifs, we say that an elementary motif $\kappa = (\hat{m}, V)$ *is* in the sequence of $P(\kappa \subset P)$, if the following condition is executed:

$$\kappa \subset P \Leftrightarrow \exists_{i=1}^{|P|} i: \eta(i, \hat{m}, P) = V. \quad (3)$$

For an arbitrary pair of proteins, P_1 and P_2, motif κ will be called *distinguishing* if κ is presented in one of the objects and not the second one. If *distinguishing* motif $\kappa \subset P \in c^{t_j}$, call it *permissive*; motif $\kappa \subset P \in \bar{c}^{t_j}$ is called *prohibitive*. Let $K(Pr, M)$ be the set of all motifs that are present in sequences from Pr in the given mask system M.

Theorem 1. *The existence of a distinguishing motif for any pair of precedents consistent with Pr is a necessary and sufficient condition for local solvability of the annotation problem.*

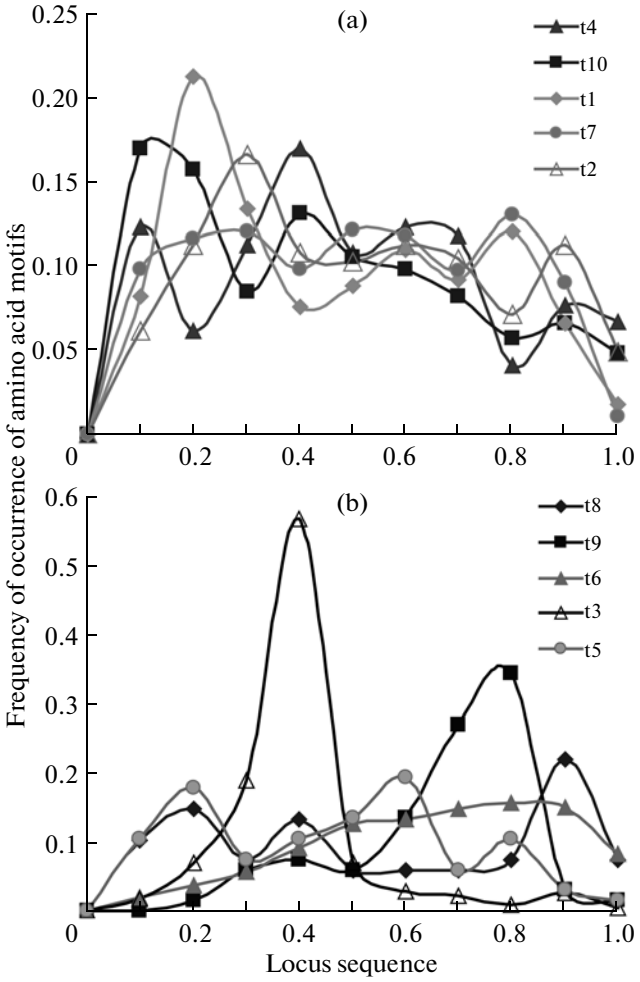

Fig. 2. Relative location of amino acid motifs for various terms of the GO dictionary.

Proof. Applying the logical operator **NOT**, we get the reverse form of the assertion (2) for a given t_j:

$$\forall_{Pr} (P_1, t_j(P_1)), (P_2, t_j(P_2)), P_1 \neq P_2: t_j(P_1) \neq t_j(P_2)$$
$$\Rightarrow \exists (i, k) \exists_{l=1}^{|M|} \hat{m}_l: \eta(i, \hat{m}_l, P_1) \neq \eta(k, \hat{m}_l, P_2). \quad (2.1)$$

The term $t_j \in T$ leads to the partition of Pr into classes c^{t_j} and $\bar{c}^{t_j}$. We proceed from $t_j(P_1)$, $t_j(P_2)$ to an expression of belonging to classes:

$$\forall_{Pr} (P_1 \in c^{t_j}) \neq (P_2 \in c^{t_j})$$
$$\Rightarrow \exists (i, k) \exists_{l=1}^{|M|} \hat{m}_l: \eta(i, \hat{m}_l, P_1) \neq \eta(k, \hat{m}_l, P_2). \quad (2.2)$$

Replacing the mask and the operators of choice of the subsequence by the motifs considering the positional

independence (3) and going from a set of masks M to a set of motifs $K(Pr, M)$, we get *the solvability condition on the set of motifs*:

$$\bigvee_{Pr}(P_1 \in c^{t_j}) \neq (P_2 \in c^{t_j}) \Rightarrow \exists_{K(Pr, M)} \kappa\colon (\kappa \subset P_1) \neq (\kappa \subset P_2). \quad (2.3)$$

Condition (2.3) proves what is required. The sufficiency is proved by contradiction. Assume that (2.3) is not executed, and for a certain pair of precedents involving the proteins P_1 and P_2, a distinguishing motif with P_1 and P_2 belonging to different classes does not exist. Then, $K(Pr, M)$ does not have such a mask and a subword that satisfy (2.1), which could satisfy the solvability condition (2). The theorem is proved.

Corollary 1. *Solvability can be guaranteed by both permissive and prohibitive motifs.* Condition (2.3) can be represented as a conjunction of conditions (2.3.1) and (2.3.2):

$$\bigvee_{Pr} P_1 \in c^{t_j}, \quad P_2 \in \bar{c}^{t_j} \Rightarrow \exists_{K(Pr, M)} \kappa\colon \kappa \subset P_1, \kappa \not\subset P_2, \quad (2.3.1)$$

$$\bigvee_{Pr} P_1 \in c^{t_j}, \quad P_2 \in \bar{c}^{t_j} \Rightarrow \exists_{K(Pr, M)} \kappa\colon \kappa \not\subset P_1, \kappa \subset P_2. \quad (2.3.2)$$

Obviously, (2.3.1) corresponds to the permissive motifs (i.e., motifs, belonging to the objects of class c^{t_j} motifs), and (2.3.2) corresponds to prohibitive motifs (i.e., motifs belonging to the inversion of class c^{t_j} motifs).

Corollary 2. *Motifs* $\kappa_1 = (\hat{m}_1, V_1)$ *and* $\kappa_2 = (\hat{m}_2, V_2)$ *are called shift-equivalent if they are formed by shift-equivalent masks* ($\hat{m}_1 = \{\mu_1^1, \mu_2^1, \ldots, \mu_m^1\}$, $\hat{m}_2 = \{\mu_1^1 + \delta, \mu_2^1 + \delta, \ldots, \mu_m^1 + \delta\}$, $\delta \in Z$) and $V_1 = V_2$. Let $\boldsymbol{\kappa}_{se}^{\hat{m}, V}$ *be the set of shift-equivalent motifs for the given* $\hat{m}$ *and* V. *Then, every motif of* $\boldsymbol{\kappa}_{se}^{\hat{m}, V}$ *ensures the execution of condition* (2.3) *if at least one motif of* $\boldsymbol{\kappa}_{se}^{\hat{m}, V}$ *is a distinguishing motif.* The proof is obvious from the definition of the entry of motif (3) and the criterion for local solvability in the form (2.3).

Note. Execution of condition (2.3) implies its validity for every pair of objects of the set Pr^2. In experimental testing the solvability of some Pr, M, and $K \subset K(Pr, M)$, condition (2.3) can only be executed on some $pr(K, Pr) \subset Pr^2$. It is natural to call the ratio $r(K, Pr) = |pr(K, \Pr)|/|Pr^2|$ *the satisfiability of the condition for solvability* when using the given Pr and K.

Theorem 1 and its corollaries serve as the basis for the practical application of the developed formalism, whose meaning is to establish a minimum set of motifs that provide solvability.

In [2, 3], the monotony of the condition for the solvability on the mask systems was analyzed, the phenomenon of irreducible mask systems was studied, and a search algorithm of irredundant mask systems was formulated. Accordingly, for the search for minimal sets of motifs satisfying (2.3), one should consider the boundaries of the monotony of the condition for solvability (2.3) by varying K.

K variation is reduced to adding or removing individual motifs. On the one hand, the addition of the motifs to K (of course, with constant $L(M)$ and $R(M)$, see (2)) does not violate the validity of (2.3), i.e., *condition* (2.3) *is monotone on K when* $K \subseteq K'$. On the other hand, the set of motifs K for which the condition (2.3) is executed may be redundant in the sense that solvability will remain when removing some motifs. If condition (2.3) is executed for K, but not satisfied for any $K' \subset K$, then such a set of motifs is called *irreducible*.

Generally speaking, the definition of the irreducible sets of motifs K of an irredundant system of masks M can be solved by an exhaustive search. However, in the case of mask systems of type M_n^m, searching all the subsets of $C_n^m|A|^m$ motifs is not feasible. Reducing the exhaustive search is possible through the classification of the attribute values and accentuation in the set of all values of all studied attributes of the subset of the "most informative." Then, in the study of the monotony of condition (2.3), one should leave the "highly informative" motifs and delete the motifs with "fairly low" informativeness.

4. HEURISTIC EVALUATION OF THE INFORMATIVENESS OF ELEMENTARY MOTIFS

The evaluation of the informativeness of the motifs D: $\mathbf{K} \longrightarrow \mathbf{R}_+$ can be administered in various ways so that greater *informativeness* of a motif will correspond to larger values of D. A rigorous set-theoretic study of the form of the appropriate functional is beyond the scope of this article and is another area for investigation. Here, we introduce some heuristic evaluations of

motif informativeness based on the occurrence rate of the elementary motifs.

Let $K(Pr, M)$ be a set of motifs for given Pr and M. As before, we consider the single term t_j and the partition generated by it of Pr into c^{t_j} and $\bar{c}^{t_j}$, and $n_1 = |c^{t_j}|$ and $n_2 = |\bar{c}^{t_j}|$. Each motif $\kappa_\alpha \in K(Pr, M)$ is a part of the N_Σ^α precedents Pr, $N_\Sigma^\alpha = n_1^\alpha + n_2^\alpha$, so that *the frequency of occurrence* of the motif in the objects of class c^{t_j} is defined as $\nu_1^\alpha = n_1^\alpha / N_\Sigma^\alpha$ and motif κ_α is associated with vector $(\nu_1^\alpha, N_\Sigma^\alpha)$.

Let the frequency of objects (precedents) in the class c^{t_j} be ν_1^0. We assume that the *informativeness* of motif κ_α is proportional to $|\nu_1^\alpha - \nu_1^0|$: i.e., the more ν_1^α differs from ν_1^0, the more informative the motif. Then, it is natural to define D^α, the evaluation of informativeness of the αth motif on class c^{t_j}, as some V-shaped function with a single minimum when $\nu_1^\alpha = \nu_1^0$ and such that $D^\alpha = 1$ when $\nu_1^\alpha = 1.0$ and $\nu_1^\alpha = 0$. This requirement is satisfied, for example, by a piecewise linear function (Fig. 3).

A V-shaped piecewise linear function of the form D^α is found by the equation $y = kx + b$ for the two sets of dots $\{(0, 1), (\nu_1^0, 0)\}$ and $\{(\nu_1^0, 0), (1, 1)\}$, so that

$$D^\alpha = \begin{cases} 1 - \nu_1^\alpha / \nu_1^0 & \text{at } \nu_1^\alpha \le \nu_1^0 \\ (\nu_1^\alpha - \nu_1^0)/(1 - \nu_1^0) & \text{at } \nu_1^\alpha > \nu_1^0. \end{cases} \quad (4)$$

The D^α value indicates how often the αth motif could be found in class c^{t_j} or, in other words, reflects the distribution of entries of the motif in objects of different classes. For example, $D^\alpha = 1.0$ corresponds to the fact that the motif occurs only among the objects of class c^{t_j} or, conversely, only among objects of $\bar{c}^{t_j}$. An important option for evaluation of D^α is the assessment of D'^α:

$$D'^\alpha = \begin{cases} 0 & \text{at } \nu_1^\alpha \le \nu_1^0 \\ (\nu_1^\alpha - \nu_1^0)/(1 - \nu_1^0) & \text{at } \nu_1^\alpha > \nu_1^0. \end{cases} \quad (4')$$

When using D^α, both the permissive and prohibitive motifs could be more informative, while using D'^α, only the permissive motifs will be among the more informative, i.e., motifs ensuring the solvability by condition (2.3.1). We note that in the annotation problem, in contrast to the problem of recognition of secondary structure [3], the use of prohibitive motifs is inappropriate in terms of the locality of the problem. In fact, the presence of a prohibitive motif at one locus does not prohibit, literally, the presence of a permissive motif in another locus of the chain.

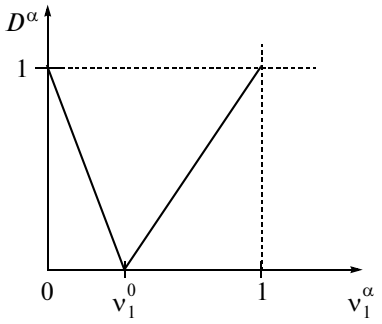

Fig. 3. Piecewise-linear V-shaped function D^α for evaluation of the informativeness of the αth motif.

In addition to comparative evaluations of the distribution of the objects among the classes, the frequency of its occurrence among the objects affects the informativeness of the motif. In other words, for fixed D^α, we assume a motif with large N_Σ^α as being more informative. Using this notation, we can offer at least three ways to evaluate the total informativeness of the αth motif:

- $D_1(\alpha) = D^\alpha$,
- $D_2(\alpha) = N_\Sigma^\alpha D^\alpha$,
- $D_2'(\alpha) = N_\Sigma^\alpha D'^\alpha$
- $D(\alpha, D_0) = \begin{cases} N_\Sigma^\alpha & \text{at } D_1(\alpha) > D_0, \\ 0 & \text{at } D_1(\alpha) \le D_0. \end{cases}$

In addition to the heuristic evaluations of the motif informativeness stated above, others also may be offered. It is intuitively clear that an informative motif must devote many of objects of class c^{t_j} and sufficiently small objects of class $\bar{c}^{t_j}$ [11]. In [12], more than a dozen different heuristic evaluations of informativeness are given, representing all sorts of heuristic functions from a pair of values similar to n_1 and n_2, such as entropy criterion of informational gain, well-known statistical criteria of xi-square, and Fisher et al.'s exact [11, 12]. As a part of the problem, heuristic evaluations of the informativeness of motifs are necessary for finding irreducible sets of motifs that take into consideration the criterion of solvability of the problem.

5. INFORMATIVENESS OF MOTIFS AND SOLVABILITY CONDITION

Let D be a heuristic evaluation of motif informativeness, D: $\mathbf{K} \longrightarrow \mathbf{R}_+$. The function D assigns to each set of motifs $K(Pr, M)$ its informativeness of a particular subset $\mathbf{R}_+$. The order relation on $\mathbf{R}_+$ induces a linear

order on the set of motifs K. Having the ordered set of motifs, the selection of the most informative may be implemented as (1) removal from K of the least informative motifs, as long as solvability remains or (2) the selection of the most informative motifs, as long as the solvability on all pairs of objects is not reached.

Let us consider the second search option of irredundant K—the selection of the most informative motifs. The introduction of a linear order on the set of motifs allows using data about the informativeness of motifs for testing solvability conditions (2.3). The principle of the selection of motifs is that for every pair of objects from Pr a distinguishing motif with the highest information content is found. The selected motifs in such a way form *a set of distinguishing motifs* K^0 with the highest informativeness *such that* $K^0 \subseteq \mathrm{K}(Pr, M)$. Let us formulate the conditions when K^0 is irreducible.

Theorem 2. *A set K^0 is irreducible if and only if for every motif from K^0 to Pr there is at least one pair of objects for which this motif is the only different one.*

Proof. First we prove sufficiency. Any two motifs $\kappa_\alpha = (\hat{m}_\alpha, V_\alpha)$ and $\kappa_\beta = (\hat{m}_\beta, V_\beta)$ can be arranged in accordance with the values of $D(\alpha)$ and $D(\beta)$. We enumerate all the elements $K = K(Pr, M)$, so that the linear order of motifs correspond to decreasing values of D: $\kappa_1, \kappa_2, \kappa_3, \ldots, \kappa_\alpha, \ldots, \kappa_{|K|}$, $D(\kappa_\alpha) \geq D(\kappa_{\alpha+1})$.

On the initial set of motifs K, let the condition of solvability be executed (2.3). Let us define the function $K_f(i, j)$, which locates the single motif with the highest D (and, hence, with a minimum number of motif α), which will help to distinguish the ith and jth objects (precedents):

$$K_f(i,j) = \min_{1\ldots|K|}\alpha\colon (\kappa_\alpha \subset P_i) \neq (\kappa_\alpha \subset P_j). \qquad (5)$$

Then, a minimal set of motifs K^0 on which the solvability remains, $K^0 \subseteq K(Pr, M)$, is determined by *the characteristic function* $T(\alpha)$:

$$T(\alpha) = \begin{cases} 1 \equiv \bigsqcup\limits_{Pr}(i,j)\colon (K_f(i,j) = \alpha) \\ 0 \quad \text{otherwise.} \end{cases} \qquad (6)$$

For each pair of ith and jth precedents $K_f(i, j)$ is the most informative distinguishing motif for the corresponding sequences. For all these motifs $T(\alpha) = 1$, i.e., these motifs form K^0. After computing $T(\alpha)$ for all pairs of precedents, each ith precedent corresponds to n_i^{rm} distinguishing motifs from K^0, $n_i^{rm} = \left|\{T(\alpha) = 1\}_i\right|$. Objects with $n_i^{rm} = 0$ are called zero-objects and objects with $n_i^{rm} = 1$ are called unit-objects. Obviously, the distinguishing motif is single only in pairs of objects composed of a zero-object and a unit-object (i.e., $n_i^{rm} + n_j^{rm} = 1$).

Now let us imagine that from K^0, the αth motif found in the N_Σ^α objects is removed. If $n_i^{rm} > 1$ for all N_Σ^α objects, than $n_i^{rm} + n_j^{rm}$ a fortiori is more than 1 and the removal of the motif may or may not lead to the loss of solvability. When $n_i^{rm} = 1$ for one of the N_Σ^α objects, then, when comparing this object with the arbitrary zero-object of another class of the αth object will be the only motif in this pair of objects and the removal of this motif will inevitably lead to the loss of solvability. K^0 cannot be irreducible when the last statement is true for all motifs.

The requirement is proved by contradiction. Let K^0 be a irreducible set of motifs. The condition for irreducibleness of K^0 is the loss of solvability when removing the arbitrary motif. In accordance with (2.3), solvability is lost when for objects of different classes there are no distinguishing motifs, i.e., $n_i^{rm} + n_j^{rm} = 0$. Assume that an arbitrary αth motif of irreducible K^0 occurs in N_Σ^α objects and for all of these objects $r_i^{rm} > 1$ (in other words, for the αth motif there is no pair of objects for which this motif is the only distinguishing motif). Then, when deleting the αth motif $n_i^{rm} + n_j^{rm} > 0$; i.e., there is a possibility of arbitrary motif removal from K^0 without losing solvability and, consequently, K^0 is not irreducible. The theorem is proved.

Corollary 3. *The set of K^0 calculated by* (6) *is irreducible.* K^0 is irreducible when it is corresponds with each motif with at least one pair of objects with a single distinguishing motif. In the construction of K^0, function $K_f(i, j)$ (5) selects *the only distinguishing motif* for any pair of objects.

Corollary 4. *Finding all zero-objects in one class is a necessary condition for solvability.* Let us assume that all zero-objects, except for the ith objects, are concentrated in class c^{t_j}, and the ith object is the only zero-object in class $\bar{c}^{t_j}$. Then, when comparing the ith zero-object with any zero-object from c^{t_j} a loss of solvability will occur.

Corollary 5. *Finding all zero-objects in one class is a necessary condition for K^0 to be irreducible.* K^0's irreducibleness implies solvability of the problem. When violating a necessary condition for solvability (Corollary 4) the deadlockness is not feasible.

Note. Irreducible sets of motifs, obtained on different Pr, may significantly differ from each other. Let us divide c^{t_j} into n nonintersecting pairs of sets of precedents Pr_i. In a fixed system of masks M, for each Pr_i, $K_i(Pr_i, M)$ is calculated, and then, $K_i^0(Pr_i, M)$ is con-

structed. Let $K = K(c^{t_j}, M) = \bigcup_{i=1}^{n} K_i(Pr_i, M)$. Then, let us call the ratio of the number of sets $K_i^0(Pri, M)$ in which this motif came as a distinguishing motifs to the sets of precedents Pr_i as *the fullness* of the αth motif, i.e.,

$$z_\alpha = \frac{\left|\{K_i^0, \kappa_\alpha \in K_i^0\}\right|}{n}. \quad (7)$$

Value $z_\alpha = 1$ indicates the entry of the most informative distinguishing motif κ_α into $K_i^0(Pr_i, M)$, built on an arbitrary Pr_i; $z_\alpha = 0$ corresponds to the fact that the αth motif is not the most informative distinguishing motif in any of the Pr_i. It is clear that the motifs with *the maximum fullness* ($z_\alpha = 1$) are of particular interest for the selection of the most informative attributes and construction of the correct recognition algorithms.

Theorem 2 and its corollaries allow us to calculate the irreducible sets of the most informative motifs. The core of the developed formalism is based on two fundamental assumptions, whose analysis is a promising direction for further research.

1. Solvability on the set of motifs is defined through the introduction of heuristic evaluations of the informativeness of motifs. The conduction of a rigorous theoretic-set justification of the possible forms of the corresponding functional, generating *D*-function is necessary.

2. Condition $D(\kappa_\alpha) \geq D(\kappa_{\alpha+1})$ in the process of calculating $K_f(i, j)$ (expression 5) corresponds to a certain arbitrariness in the selection of the motif when $D(\kappa_\alpha) = D(\kappa_{\alpha+1}) = D(\kappa_{\alpha+2}) = ...$, etc. Arbitrariness in the selection of the motif raises the question about the problems of retraining recognition algorithms, which will be built using the irreducible K^0, built on different samples of objects. Variation of the occurrence of motifs in different samples of objects also makes it necessary to introduce a combinatorial evaluation of the values of *D*.

6. EXPERIMENTAL TESTING OF A CRITERION FOR LOCAL SOLVABILITY

Condition (2.3) and Theorem 2 make it possible to conduct experiments on the evaluation of the solvability of the local *t*-classifiers. All numerical experiments whose results are described in this paper were based on the stable version of the annotation of the human genome (NCBI genome build 36, www.ncbi.nlm.nih.gov). This version of the annotation of the genome consists of almost 14 000 of the 29 875 proteins with annotations (i.e., attributed to certain terms of *Gene Ontology*).

Since the calculation of the characteristic function on expression (6), N^2, is a difficult task, experiments were performed only for the two most common terms: t_1 (nucleus, $|c^{t_1}| = 2890$, $|\bar{c}^{t_1}| = 26985$) and t_2 (membrane, $|c^{t_2}| = 2450$). For each t_j, test samples *Pr* were formed for a calculation of characteristic functions of $T(\alpha)$ by random selection of objects without replacement. Samples of 2, 5, and 10% of the objects from c^{t_j} and an equal number of objects from $\bar{c}^{t_j}$ were examined, by 10 samples for each of the three values given above. Calculations were also carried out for samples which included 20% of c^{t_j}, by 5 samples for t_1 and t_2. Further, samples of a certain size are denoted as 10% *Pr*, 20% *Pr*, etc. The study of larger samples of the precedents at the present time are not possible due to significant computational difficulties (for example, the experiments described in this paper took 12 weeks on a 2-nuclear PC 2.7 GHz).

Each of *the mask systems used* had a fixed dimension of all masks. The solvability of nested mask systems, which are elements of form $M_m^m \subset M_{m+1}^m \subset ... \subset M_{n-1}^m \subset M_n^m$, was tested. The maximum extension of masks in all systems was 8 positions. The studied mask systems were based on the mask system with the dimension of all masks equal to 2 (system M_8^2, $|M_8^2| = C_8^2 = 28$) and dimension 3 (M_8^3, $|M_8^3| = C_8^3 = 56$). By Theorem 1, Corollary 2, the removal of the shift-equivalent masks will not lead to the loss of solvability; therefore, for the calculation of $T(\alpha)$ the mask systems $M_8'^2$, $|M_8'^2| = 11$ and $M_8'^3$, and $|M_8'^3| = 25$ were used, obtained by reduction of the shift-equivalent mask systems M_8^2 and M_8^3, respectively.

The feasibility of using *heuristic evaluations* of *the informativeness of the motifs* $D_1(\alpha)$, $D_2(\alpha)$, $D_2'(\alpha)$, and $D(\alpha, D_0)$ was investigated. Preliminary experiments showed that the evaluation of $D_2'(\alpha)$ leads to irreducible sets of motifs of the smallest dimension.

Calculations $T(\alpha)$ showed that current formalism allows performing an effective reduction of the set of motifs $K(Pr, M)$ to irreducible K^0 without the loss of solvability. For example, each of the sets of the motifs $K(Pr, M_8'^3)$, built for the 10% *Pr* sample, contained 176000–177000 motifs. The number of selected motifs, i.e., $|K^0|$, was 190–250, which is less than 0.2% of the original $K(Pr, M_8'^3)$.

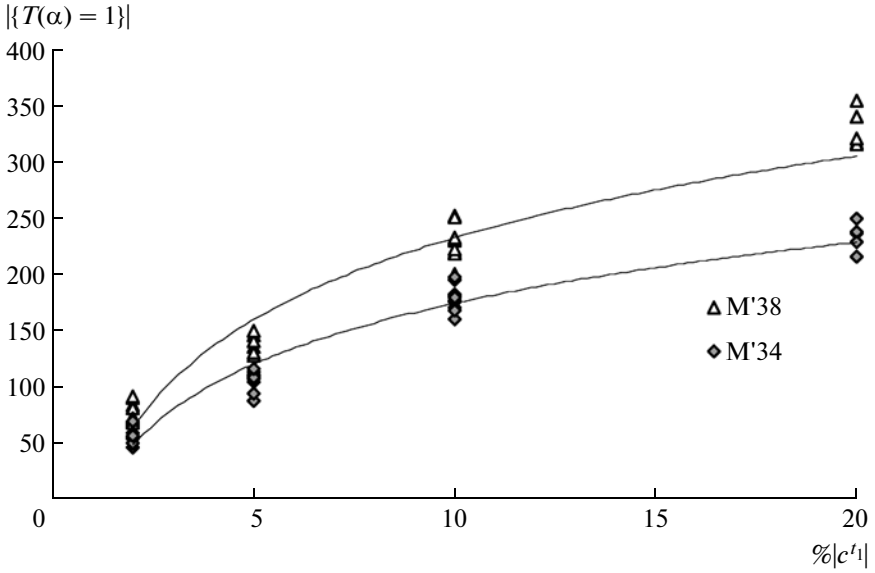

Fig. 4. Dependence of the number of selected motifs on the size of the set of precedents.

The logarithmic nature of the relationship between the number of selected motifs from $|Pr|$ (Fig. 4) suggests that the efficiency of reduction of the set of motifs will remain high at larger size samples of the precedents. As expected, the reasons for K^0 (i.e., motifs with $T(\alpha = 1)$ are most common among motifs with high informativeness (i.e., the lowest α, Fig. 5).

Let us consider the dependence of the number of pairs of objects on which the solvability is reached (maximum, $|Pr^2|$) on the number of motifs with the highest informativeness. Since the number of motifs in K^0 is dependent on the number of objects (Fig. 4), we use percentages to compare the results obtained for the Pr of different sizes (Fig. 6).

The results presented in Fig. 6, indicate the existence of a core in a set of irreducible motifs. Motifs belonging to such a core provide solvability in most pairs of objects. For example, in the irreducible K^0, built on the mask system $M_4^{'3}$, only 20% of the most informative motifs K^0 ensure the solvability of more than 90% of pairs of objects (i.e., $r = 0.9$). Complete solvability is achieved by adding to this core of the set of low-informative motifs, each of which provides solvability to the relatively small number of pairs of objects.

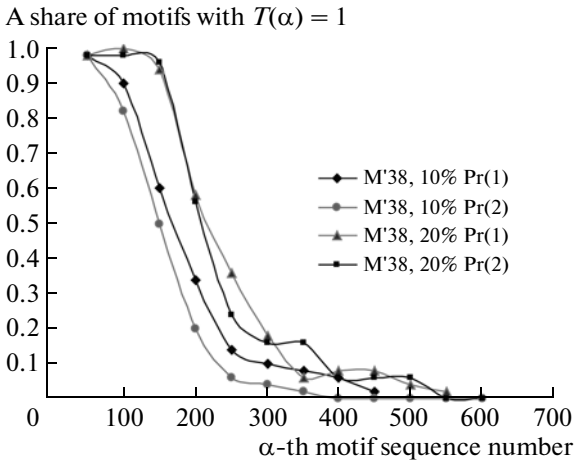

Fig. 5. Motifs of irreducible K^0 are found mostly among common motifs with high informativeness. Index α is a sequence number of a motif in $K(Pr, M)$.

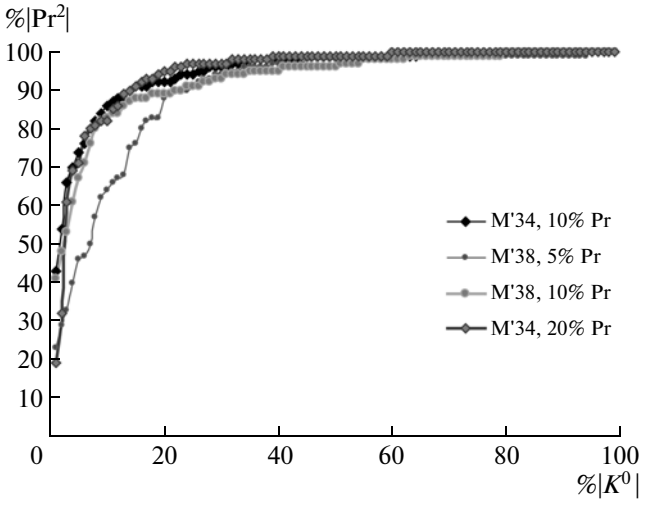

Fig. 6. Percentage of recognizable pairs of objects for the Pr of various sizes.

Table 2. Satisfactoriness of the condition for the solvability for t_1 on sets of motifs with maximum fullness for different M and Pr. For various Pr of the same size, the average error in the values of $r(K, Pr)$ was 0.01 and did not exceed 0.03

Size of a set of precedents	M_3^3	$M_3'^3$	$M_6'^3$	$M_8'^3$
2% Pr (58 × 2 objects)	0.47	0.56	0.28	0.30
5% Pr (145 × 2 objects)	0.90	0.90	0.32	0.39
10% Pr (289 × 2 objects)	0.98	0.98	0.39	0.35
20% Pr (578 × 2 objects)	0.99	1.00	0.75	0.66

Table 3. Satisfactoriness of the condition for solvability of t_2 on the sets of motifs with maximum fullness

Size of a set of precedents	M_3^3	$M_3'^3$	$M_6'^3$	$M_8'^3$
2% Pr (49 × 2 objects)	0.94	0.94	0.56	0.84
5% Pr (123 × 2 objects)	0.97	0.97	0.53	0.78
10% Pr (245 × 2 objects)	0.99	0.99	0.88	0.98
20% Pr (490 × 2 objects)	0.99	1.00	0.98	0.99

One can find particular interest in the study of the solvability of such subsets of the irreducible sets of motifs, which consist of motifs with a maximum fullness ($z_\alpha = 1$). Calculations were carried out on *the satisfactoriness of the condition for solvability* $r(K_{z=1}, Pr)$, $K_{z=1} = \{\kappa_\alpha, T(\alpha) = 1, z_\alpha = 1\}$ for terms t_1 and t_2, using irreducible sets of motifs, built for various mask systems in different sizes of the studied samples of precedents (Tables 2 and 3).

The experimental results summarized in the tables show that many motifs with a maximum fullness ensure 99–100% satisfactoriness of the condition for solvability. Both the sample size of the precedents and the parameters of the mask system have a significant impact on the satisfactoriness (2.3). The $r(K_{z=1}, Pr)$ values increase to 1.0 when the size of the set of precedents is increased and when the extension of the masks is reduced. It is clear that the use of mask systems M_n^m with $n \approx m$ and $m = 3$, for 99% satisfactoriness of the condition for solvability, it is sufficient to use a set of precedents, including not more than 10% of c^{t_j}.

7. CONCLUSIONS

In this work, the development of the formalism to study the local solvability of the genome annotation problem was carried out. It is shown that the ordering of the set of the motifs via heuristic evaluation of informativeness allows effective reduction of the set of the motifs without the loss of solvability. The designed formalism allowed us to experiment to find the irreducible sets of the most informative motifs, and to establish the most appropriate sizes of the set of the precedents and the optimal parameters of the mask systems. Long-term directions of further investigations are formulated: a theoretic-set justification of the evaluations of informativeness and combinatorial evaluations of values of D. Finding the irreducible sets of the most informative motifs is essential for the next phase of this study—the synthesis of algorithms in terms of the algebraic approach to pattern recognition.

ACKNOWLEDGMENTS

This work was supported by the Russian Foundation for Basic Research (grants 09-07-12098, 09-07-00 212-a, and 09-07-00211-a).

REFERENCES

1. I. Yu. Torshin, *Bioinformatics in the Post-Genomic Era: Sensing the Change from Molecular Genetics to Personalized Medicine* (Nova Biomedical Books, New York, 2009).
2. I. Yu. Torshin, "On Solvability, Regularity, and Locality of the Problem of Genome Annotation," Pattern Recogn. Image Anal. **20** (3), 386–395 (2010).
3. K. V. Rudakov and I. Yu. Torshin, "Solving Ability Problems of Protein Secondary Structure Recognition," Informat. Prim. **4** (2), 25–35 (2010).
4. K. V. Rudakov, "Signs Values Classification Problems in Recognition Problems," in *Proc. Int. Conf. "Intellectualization of Information Processing" IIP-8* (Paphos, Oct. 17–23, 2010).
5. I. Yu. Torshin, "Motive Analysis in the Problem of Protein Secondary Structure Recognition on the Base of Solvability Criterion," *Proc. Int. Conf. "Intellectualization of Information Processing" IIP-8* (Paphos, Oct. 17–23, 2010).
6. Yu. I. Zhuravlev, "Set–Theoretical Methods for Logic Algebra," Probl. Kibernet. **8** (1), 25–45 (1962).
7. Yu. I. Zhuravlev, "On Algebraic Approach for Solving Classification and Recognition Problems," in *Cybernetic Problems* (Nauka, Moscow, 1978), Issue 33, pp. 5–68 [in Russian].
8. K. V. Rudakov, "The Way to Use the Universe Limitations for Researching the Classification Algorithms," Kibernet., No. 1, 1–5 (1988).
9. M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese,

J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock, "Gene Ontology: Tool for the Unification of Biology. The Gene Ontology Consortium," Nature Genet. **25**, 25–29 (2000).

10. N. Hulo, C. J. Sigrist, V. Le Saux, P. S. Langendijk-Genevaux, L. Bordoli, A. Gattiker, E. De Castro, P. Bucher, and A. Bairoch, "Resent Improvements to the PROSITE Database," Nucl. Acids Res. **1** (32 Database Issue), D134-7 (2004).
11. K. V. Vorontsov, "Combinatorial Reliability Theory for Precedent Learning," Doctoral Dissertation in Mathematics and Physics (Vychislitel'nyi Tsentr RAN, Moscow, 2010).
12. J. Furnkranz and P. A. Flach, "Roc 'n' Rule Learning–Towards a Better Understanding of Covering Algorithms," Mach. Learn. **58** (1), 39–77 (2005).

Ivan Yur'evich Torshin. Born 1972, graduated from the Chemistry Department of Moscow State University (MSU) in 1995. Received Candidate's degree at the Chemistry Department of Moscow State University in 1997. Senior researcher of the Russian Branch of the Institute of Trace Elements for UNESCO, lecturer at Moscow Institute of Physics and Technology (MIPT) and MSU, a member of the Center of Forecasting and Recognition. Author of 84 papers in reference journals in biology, chemistry, medicine, and computer science, including 3 monographs of the series Bioinformatics in the Post-Genomic Era" (Nova Biomedical Publishers, NY, 2006–2009).