On Solvability, Regularity, and Locality of the Problem of Genome Annotation

I. Yu. Torshin

Russian Satellite Center of the Trace Element Institute for UNESCO, Bol'shoi Tishinskii per. 26, stroen. 15–16, Moscow, 109652 Russia
e-mail: tiy135@yahoo.com

Abstract—Determination of the nucleotide sequences of hundreds of organisms (in the first place, the human genome) is a significant technical achievement of modern biology. The next stage of studying the genome is to determine the functions of each gene and the corresponding protein: the so-called genome annotation. The existing methods of classifying the biological roles of proteins on the basis of the amino acid sequence are restricted to searching for similar sequences in a database and, as a result, have limited applicability. In this paper, a formalism is introduced for studying this problem in the framework of the algebraic approach and the solvability, locality, and regularity of the problem and monotonicity of the condition of solvability are considered. The proposed formalism enables one to study systematically the hypothesis of locality of various biological roles of proteins.

Key words: algebraic approach to pattern recognition, data mining, genome annotation, postgenomic bioinformatics.

DOI: 10.1134/S1054661810030156

INTRODUCTION

A genome is the nucleotide sequence of the entire DNA of an organism, and DNA is the databank of a cell. Intricate macromolecular complexes of a cell produce all the proteins necessary for the life of an organism on the basis of the data encoded in the DNA. In other words, all possible physiological processes in a given organism are encoded in the genome, so the data in the genome has everlasting significance both for fundamental medical research and for practical medicine.

For basic research and practical applications in biomedicine, the studied genome must be annotated, i.e., (i) all proteins that can be produced on the basis of the genome DNA and (ii) all biological roles (or biological functions, in other terminology) of these proteins must be determined. Then, the set of lists of all roles of all proteins will be a full annotation of the genome.

At the present time, experimental determination of all biological roles of all proteins is practically impossible, and development of effective theoretical methods for solving the annotation problem is necessary. Bioinformatics methods existing in biology are based solely on the dogma of the similarity of biological roles of proteins if their amino acid sequences are similar. These methods are characterized by rather restricted applicability to real objects, because (i) they arbitrarily use the conception of similarity and (ii) there are many proteins with close functions and dissimilar amino acid sequences. In any case, all methods known in modern biology make it possible to annotate at most half the human genome [1].

In this work, a formalism for consideration of genome annotation from the viewpoint of the algebraic approach to recognition problems is presented. This formalism employs some elements of the formalism developed earlier for solving the problem of recognition of the protein secondary structure [2, 3]. All numerical experiments described in the present work were performed on the basis of a stable version of human genome annotation (NCBI genome build 36, www.ncbi.nlm.nih.gov).

1. INITIAL DEFINITIONS

Let there be given an alphabet A corresponding to the set of all amino acids forming proteins: $A = \{a_1, a_2, ..., a_n\}$, $n > 0$. Denote the set of words of length k in this alphabet by A^k. Then, the set of all initial words in the alphabet A is $A^* = \bigcup_{l=1}^{\infty} A^l$. A protein or an amino acid sequence will be understood as a word P^l of length l, $P^l \in A^*$.

Let there be defined a dictionary T of m terms t_i describing all known roles or biological functions of proteins: $T = \{t_1, t_2, ..., t_m\}$. If T^k is the set of all combinations of k terms t_i, then the set of annotations T^*

Received April 15, 2010

over the dictionary T is $T^* = \bigcup_{l=1}^{\infty} T^l$. An annotation of a protein P will be understood as a list $t \subset T^*$ of k terms: $t = \{t_1, t_2, ..., t_k\}$, $t_j \in T$.

Let Δ denote the uncertainty. Let us introduce the Δ-extended alphabet $\tilde{A} = A \cup \{\Delta\}$, the Δ-extended terminology dictionary $\tilde{T} = T \cup \{\Delta\}$, and the Δ-extended sets $\tilde{A}^* = \bigcup_{l=1}^{\infty} \tilde{A}^l$ and $\tilde{T}^* = \bigcup_{l=1}^{\infty} \tilde{T}^l$, respectively. The genome annotation problem is defined as finding the function F that maps the set of proteins to the set of annotations: F: $\tilde{A}^* \longrightarrow \tilde{T}^*$.

Let us take a set $\mathbf{P}$ of n words in the alphabet A: $\mathbf{P} = \{P_1, P_2, ..., P_n\}$. These can all be proteins encoded by a sequence of a certain genome (the so-called proteome) or an arbitrary set of proteins (say, a sample of protein from a database). The set of words $\mathbf{P}$ will be referred to as annotated if each protein $P_i \in \mathbf{P}$ is assigned with a list $t^i \subset T^*$, $t^i = \{t^i_1, t^i_2, ..., t^i_{k_i}\}$, $t^i_j \in T$, $k > 0$.

A set of precedents Pr will be understood as a set $\Pr \subseteq \tilde{A}^* \times \tilde{T}^*$. Thus, a precedent is a pair consisting of a word $P_i \in \mathbf{P}$ and the corresponding annotation $t^i \subset \tilde{T}^*$, $(P_i, t^i) \in \Pr$. A function F is correct on the set of precedents if $\forall_{\Pr} (P, t)$: $F(P) = t$. A set of precedents Pr is consistent if the function F exists on the given Pr and is inconsistent if F does not exist.

Theorem 1. *F exists if and only if the following conditions of consistency of a set of precedents takes place*:

$$\forall_{\Pr} (P_i, t^i), (P_j, t^j), \quad i \neq j: (P_i = P_j) \Rightarrow (t^i = t^j). \quad (1)$$

Proof. According to (1), the existence of F is estimated on a definite Pr. Assume that, in the chosen Pr, there is a pair of precedents (i, j) such that, for $P_i = P_j = P$, the lists t^i, $t^j \subset T^*$ are not equal: $t^i \neq t^j$, i.e., $t^i \cup t^j \neq t^i \cap t^j$. Then, $F(P) = t^i \neq t^j = F(P)$, which does not agree with the definition of a function. Hence, condition of consistency (1) of the set of precedents is the condition of existence of F. The theorem is proven.

Along with solvability, the modern theory of recognition [4–12] usually studies the regularity of a problem, i.e., the solvability of a given problem combined with the solvability of problems in a certain vicinity of this problem in the studied set of problems. Following the ideas of the scientific school of Academician Yu.I. Zhuravlev, we will define the vicinity of a problem Z with the set of precedents $\Pr = \{(P_1, t^1), (P_2, t^2), ..., (P_n, t^n)\}$ as the set of problems Z' with the set of precedents $\Pr' = \{(P_1, t^{1'}), (P_2, t^{2'}), ..., (P_n, t^{n'})\}$ with arbitrary $t^{1'}, t^{2'}, ..., t^{n'}$. Obviously, in this case, the problem Z will be regular on the set of precedents Pr under the condition

$$\forall_{\Pr} (P_i, t^i), (P_j, t^j), \quad i \neq j \Rightarrow (P_i \neq P_j). \quad (2)$$

Several remarks should be made concerning solvability and regularity of the problem under study on real sets of precedents. From the viewpoint of the problem area, proteins belonging to different organisms can have different functions, i.e., $P_i = P_j$ will correspond to $t^i \neq t^j$. Therefore, the existence of the function F in the case of an annotation problem is guaranteed by selecting Pr from proteins of the same organism (i.e., obtained from the same genome). The latter removes inconsistent sets of precedents, typical for the problem of protein secondary structure recognition [3]. Moreover, in the case of proteins of one genome, $\forall (i, j)$: $P_i \neq P_j$; i.e., the problem is also regular.

In the case of the annotation problem, the alphabet A is determined uniquely as $A = \{$A, C, D, E, F, G, H, I, K, L, M, N, P, R, S, T, V, W, Y$\}$. Another essential problem is determination of the terminology dictionary T. In the general case, an annotation of a known protein is a set of definite key words established from published data. Therefore, it is evident that there are a large number of methods for determining the terminology dictionary T. Recently, in biology the unified system of terms known as GO (Gene Ontology) [13, 14], which is used for describing biological functions of proteins, has gained increasing acceptance.

GO is a hierarchical system of terms. For example, the property of a protein molecule to bind a magnesium ion is assigned in the GO dictionary with a special indicator (GO:0000287), a key word pointing to the corresponding biological function (magnesium ion binding), and definition of the term (interacting selectively with magnesium ions). The term "binding metal ions" (GO:0046872) is situated higher in the term hierarchy. If the same protein not only binds magnesium but also takes part in magnesium ion transport through the cell membrane, it is described by the term GO:0015693 "magnesium ion transport." Thus, the annotation of one protein or another in the GO dictionary may be understood as a set of corresponding identifiers {GO:0046872, GO:0000287, GO:0015693, ...} and key words and definitions unambiguously assigned to these identifiers.

The GO database (www.geneontology.org) contains more than 23100 such terms applicable for

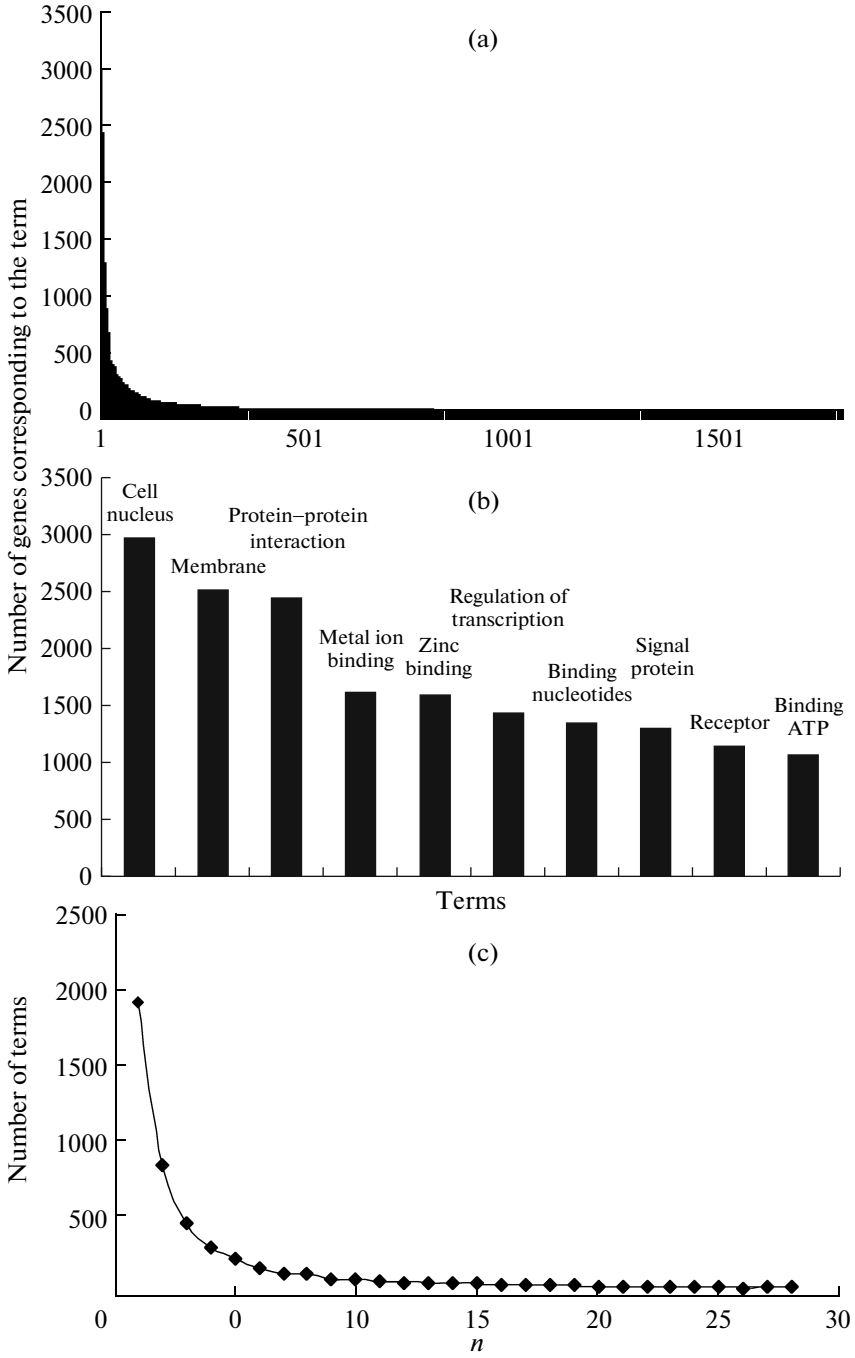

Fig. 1. Annotation of the human genome by terms from the GO dictionary. (a) Frequency of occurrence of different terms. Data for the terms occurring more than three times are shown; (b) examples of most frequently occurring terms; (c) the number of different terms occurring in the considered sample at most n times.

describing any known biological function of proteins of one organism or another. In the case of the human genome, only 5000 terms of these 23100 are used (Fig. 1). The most frequently used terms are shown in Fig. 1b.

It should be noted that more than 3000 of these 5000 terms are unique: they occur only in the existing annotation of the human genome less than 3 times (Fig. 1c). Such a low frequency of occurrence does not allow one to construct an acceptable set of precedents. Apparently, the problem of unique terms can be solved by using consistent sets of precedents including proteins with similar biological roles found in genomes of different organisms.

2. BELONGING OF A PROTEIN TO CLASSES OF BIOLOGICAL FUNCTIONS OF PROTEINS AND *t*-CLASSIFIERS

A term dictionary T of m terms, $T = \{t_1, t_2, ..., t_m\}$, corresponds to m classes of proteins C^{t_1}, C^{t_2}, ..., C^{t_m}, $C^{t_i} \subset A^*$. The inversion of a class C^{t_i} will be understood as the set of proteins $\bar{C}^{t_i}$ not belonging to C^{t_i}; i.e., $C^{t_i} \cap \bar{C}^{t_i} = \varnothing$. If each protein $P_i \in \mathrm{Pr}$ is assigned with a list $t^i \subset T^*$, $t^i = \{t^i_1, t^i_2, ..., t^i_{k_i}\}$, then the annotated protein P_i belongs to k_i classes C^{t_1}, C^{t_2}, ..., C^{t_k}. Thus, the term $t_j \in T$ corresponds to the partition of the set of precedents into two nonintersecting subsets $c^{t_j} \subseteq C^{t_j}$ and $\bar{c}^{t_j} \subseteq \bar{C}^{t_j}$ such that $c^{t_j} \cup \bar{c}^{t_j} = \mathrm{Pr}$ and $c^{t_j} \cap \bar{c}^{t_j} = \varnothing$.

In the general case, classes C^{t_i} overlap, so that $\forall i, j = 1, ..., m$: $C^{t_i} \cap C^{t_j} \neq \varnothing$. The overlapping of classes suggests a certain interrelation between phenomena described by the terms t_i and t_j. The degree of this interrelation can be estimated by comparing the powers of the subsets c^{t_i} and c^{t_j} of the set of precedents. The linkage between two functional terms t_i and t_j will be understood as the parameter s_{ij} defined as

$$s_{ij} = \mathrm{sgn}(|c^{t_i}| - |c^{t_j}|)\frac{|c^{t_i} \cap c^{t_j}|}{\min(|c^{t_i}|, |c^{t_j}|)}. \qquad (3)$$

For $s_{ij} = 1$, $c^{t_j} \subseteq c^{t_i}$; for $s_{ij} = -1$, $c^{t_i} \subseteq c^{t_j}$; i.e., $|s_{ij}| = 1$ corresponds to totally overlapping subsets and the sign of the linkage s_{ij} indicates the order of inclusion of the ith subset to the jth one. The value $s_{ij} = 0$ corresponds to totally nonoverlapping subsets c^{t_i} and c^{t_j}. It should be noted that the calculation of linkages between terms will be necessary for estimating the completeness and consistency of the annotation of each $P_i \in \mathrm{Pr}$.

Since $c^{t_j} \cup \bar{c}^{t_j} = \mathrm{Pr}$ and $c^{t_j} \cap \bar{c}^{t_j} = \varnothing$, the solution to the annotation problem can be reduced to solving m problems on the membership of P_i to each of the m given classes of biological roles. In this case, the annotation t^i of a protein P_i, which is a list of k_i terms, is uniquely translated to a Boolean vector of annotations $\mathbf{t}^i$ such that $\mathbf{t}^i = \{t_1(P_i), ..., t_j(P_i), ..., t_m(P_i)\}$, where $t_1(P_i) = 1$ if $P_i \in C^{t_j}$ and $t_1(P_i) = 0$, otherwise. The function $f_{t_j}(P_i)$, which determines membership in the class C^{t_j}, will be referred to as the t_j-classifier or the jth t-classifier. One t-classifier or another relates P_i to one of two nonintersecting subsets: $c^{t_j} \subseteq C^{t_j}$ or $\bar{c}^{t_j} \subseteq \bar{C}^{t_j}$. A t-classifier is correct when

$$\forall P \in \mathrm{Pr}: f_{t_j}(P) = \begin{cases} 1, & \text{if } P \in c^{t_j} \\ 0, & \text{if } P \notin c^{t_j},\ P \in \bar{c}^{t_j} \\ \Delta, & \text{if algoritm has not calculated membership.} \end{cases} \qquad (4)$$

Theorem 2. *The correctness of each t-classifier is the necessary and sufficient condition for correctness of function F.*

Proof. F maps the set of words in the alphabet A to the set of annotations: F: $\tilde{A}^* \longrightarrow \tilde{T}^*$. F is correct if, on a consistent set of precedents, $\underset{\mathrm{Pr}}{\forall} (P_i, t^i)$: $F(P_i) = t^i$. It can be written in the annotation vector form as $\underset{\mathrm{Pr}}{\forall} (P_i, \mathbf{t}^i)$: $F(P_i) = \mathbf{t}^i$. If all t-classifiers f_{t_j} are correct, the annotation vector $\mathbf{t}^i$ is calculated as $\mathbf{t}_v = \{f_{t_1}(P_i), ..., f_{t_j}(P_i), ..., f_{t_m}(P_i)\}$, $\mathbf{t}_v = \mathbf{t}^i$, so that $F(P_i) = \mathbf{t}^i$. If at least one of the t-classifiers is incorrect, $\mathbf{t}_v \neq \mathbf{t}^i$ and, hence, the condition of correctness is violated: $F(P_i) \neq \mathbf{t}^i$. Therefore, the correctness of any t-classifier is a necessary condition and the correctness of m (i.e., all) t-classifiers is a sufficient condition for correctness of F. The theorem is proven.

Sets of precedents can differ in the degree of completeness of the annotation of each protein. From the biological point of view, a reliable determination that a protein P belongs to only one class is, certainly, an incomplete description of P. According to the presently available annotation of the human genome, the majority of proteins and the corresponding genes, indeed, belong to at least three classes or have at least three different biological functions (Fig. 2). At the same time, the number of classes to which a protein belongs (i.e., $|t^i|$) does not exceed 40.

The linkage between terms t_i and t_j ($s_{ij} \in [-1, 1]$) calculated on the given set of precedents reflects the degree of similarity between the biological functions of proteins from the classes C^{t_i} and C^{t_j}. There can be both mutually exclusive classes of biological functions ($s_{ij} \approx 0$) and classes interrelated to one degree or another ($|s_{ij}| > 0$). Different variants of linkage between terms and the corresponding biological functions are illustrated by the data presented in the table. In this example, negative values of linkage are absent because the terms t_1, ..., t_{11} are arranged in the order of

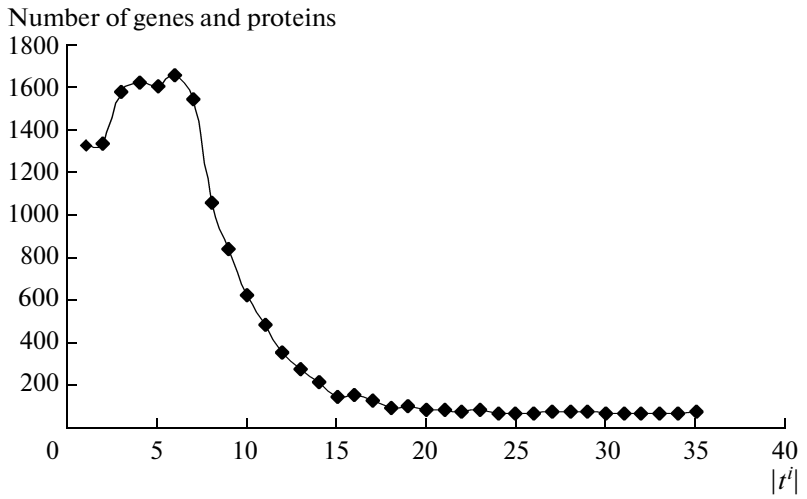

Fig. 2. The number of human genes belonging to $|t^i|$ functional classes.

descending frequency of their occurrence in the genome annotation (as in Fig. 1b).

The calculated linkages agree well with the qualitative relationships known in biology between different kinds of biological roles of proteins. Let us consider the linkage of the term t_1 (cell nucleus). The high positive values of the linkage with the terms t_6 (regulation of transcription) and t_{11} (transcription, i.e., the process of gene activation and RNA synthesis) reflect the fact that practically all processes of transcription pro-

The values of linkage (s_{ij}) for the terms most frequently occurring in genome annotation

t_i \ t_j	t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8	t_9	t_{10}	t_{11}
t_1	1.00	0.03	0.34	0.54	0.62	0.93	0.24	0.10	0.03	0.23	0.95
t_2		1.00	0.21	0.10	0.06	0.01	0.14	0.30	0.47	0.15	0.01
t_3			1.00	0.27	0.27	0.23	0.23	0.17	0.12	0.22	0.28
t_4				1.00	0.80	0.38	0.05	0.05	0.02	0.06	0.50
t_5					1.00	0.45	0.05	0.05	0.02	0.06	0.48
t_6						1.00	0.04	0.06	0.02	0.05	0.85
t_7							1.00	0.11	0.07	0.84	0.05
t_8								1.00	0.58	0.10	0.05
t_9									1.00	0.08	0.02
t_{10}										1.00	0.05
t_{11}											1.00

Term	Identifier
t_1: cell nucleus	GO: 0005634
t_2: membrane	GO: 0016020
t_3: protein–protein interaction	GO: 0005515
t_4: metal ion binding	GO: 0046872
t_5: zinc binding	GO: 0008270
t_6: regulation of transcription	GO: 0006355
t_7: binding nucleotides	GO: 0000166
t_8: signal protein	GO: 0007165
t_9: receptor activity	GO: 0004872
t_{10}: binding adenosine triphosphate (ATP)	GO: 0005524
t_{11}: transcription	GO: 0006350

ceed in the cell nucleus. The DNA-binding sections of many proteins, which are so-called transcription factors that directly interact with DNA, can bind a DNA molecule only by virtue of a zinc (metal) ion; hence, we have the linkage with the terms t_4 (binding metal ions) and t_5 (binding zinc). At the same time, for example, receptor molecules are situated on the external cell membrane, which is physically separated from the nucleus; hence, the linkages of the term t_1 (cell nucleus) with the terms t_2 (membrane) and t_9 (receptor activity) are close to zero.

3. LOCALITY OF THE PROBLEM ON MEMBERSHIP OF A PROTEIN IN A FUNCTIONAL CLASS

Different classes of biological functions of proteins can relate either to the entire amino acid sequence of a protein P_i or to a rather small section of it. For example, selective binding of a zinc ion (term GO:0008270) or magnesium ion (term GO:0000287) is provided by the existence of one specific subsequence or another (so-called motif) of 7–10 amino acids that exist. The same may be said about the proteins' property defined by the term "binding ATP" (GO:0005524). It is known that ATP (adenosine triphosphate) is the energy substrate of a cell, which interacts with many proteins. The motif of an amino acid sequence that can be written as (A $\vee$ G)–X–X–X–X–G–K–(S $\vee$ T), where $\vee$ denotes "or" and X is any letter of the alphabet A, binds ATP with high selectivity (motif PS00017 in the PROSITE database). At the same time, protein–protein interactions (term GO:0005515) can involve as many as half the amino acids of one protein P_i or another. In the case of transmembrane proteins (GO:0016020, membrane), the entire amino acid sequence P_i often takes part in the interaction with the membrane.

In other words, one biological function of a protein or another can be localized or delocalized in the sequence for the protein P_i. No systematic study of this question has been made, although there are many known particular cases. For example, the PROSITE (PROtein SITEs) database [15] contains more than 1500 biologically important motifs of amino acid sequences similar to the above-presented motif of ATP binding. Most of these motifs, which enable one to recognize biological functions of the corresponding proteins, are significantly shorter than the amino acid sequences of these proteins. Therefore, it seems reasonable to develop a formalism for studying the locality of various biological functions of proteins.

The proposed formalism is based on the approach developed earlier for studying the locality of the problem of protein secondary structure recognition [3]. Let there be given a word $\vec{U} = \{u_1, u_2, ..., u_n\}$ of length n corresponding to the amino acid sequence of a certain protein P. Let us determine a certain leading position i, $1 \le i \le n$. Also assume that there is a mask $\hat{m} = \{\mu_1, \mu_2, ..., \mu_m\}$, where $\mu_i \in Z$, $\mu_1 < \mu_2 < ... < \mu_m$. We will refer to μ_i as positions. The parameter m will be referred to as the dimension of the mask $\hat{m}$ and denoted by $|\hat{m}|$; the value of the expression $\mu_m - \mu_1 + 1$ will be referred to as the length of the mask and denoted by $[\hat{m}]$. Let us define the subword selection operator $\eta(i, \hat{m}, \vec{U})$ as

$$\eta(i, \hat{m}, \vec{U}) = \begin{cases} u_{i+\mu_1} u_{i+\mu_2} \dots u_{i+\mu_m} & \text{if } i + \mu_1 \ge 1, i + \mu_m \le n \\ \varnothing & \text{otherwise.} \end{cases} \quad (5)$$

A subword or $(\hat{m}, i)$-subword will be understood as a particular value of the operator η at a definite position i of a certain word $\vec{U}$, selected by virtue of the mask $\hat{m}$. From the point of view of the algorithm for calculation of estimates [4–7], the pair $(\hat{m}, i)$ may be considered as an analog of the support set, and the $(\hat{m}, i)$-subword, as an analog of representative selection. A motif will be understood as the set consisting of a mask of a certain word $\vec{U}$, selected by virtue of the mask $\hat{m}$ and an $(\hat{m}, i)$-subword, i.e., a pair $(\hat{m}, \eta(i, \hat{m}, \vec{U}))$.

Let there be given a system of masks $M = \{\hat{m}_1, \hat{m}_2, ..., \hat{m}_N\}$. Assume that $\hat{m}_1 = \{\mu_1^1, \mu_2^1, ..., \mu_{|m_1|}^1\}$, ..., $\hat{m}_N = (\mu_1^N, \mu_2^N, ..., \mu_{|m_N|}^N)$. We will define a single-element system of masks $\hat{M}_\Sigma(M)$ as a united mask $\hat{m}$ such that $\hat{m} = \bigcup_{k=1}^{|M|} \hat{m}_k$. Obviously, $\bigvee_{k=1}^{|M|} \hat{m}_k$: $(|\hat{m}_k| \le |\hat{M}_\Sigma|)$ & $([\hat{m}_k] \le [\hat{M}_\Sigma])$.

Words P_i in the set of precedents Pr have a finite length, so the domain of the operator η with the given system of masks M should be written with allowance for the boundary conditions caused by finite lengths of amino acid sequences P_i. Let $L, R \in N \cup \{0\}$. We will denote the values of L and R calculated on the system of masks M by $L(M)$ and $R(M)$, respectively. $L(M)$ and $R(M)$ are defined as the minimum setoffs from the edges (left and right, respectively) of the upper word at which all masks from M are applicable. In this case, $L(M) + 1 = \min(i)$: $\bigvee_{k=1}^{N} (i + \mu_1^k \ge 1)$ and $R(M) = |\vec{U}| - \max(i)$: $\bigvee_{k=1}^{N} (i + \mu_{|m_k|}^k \le |\vec{U}|)$. Then, $L(M) =$

$\max(-\min_{k=1,N}\mu_1^k, 0)$ and, analogously, $R(M) = \max(\max_{k=1,N}\mu_{|m_k|}^k, 0)$. Thus, the operator $\eta(i, \hat{m}, \vec{U})$ is defined on $i \in \{L(M), |\vec{U}| - R(M)\}$. Henceforth, we will assume that, for the given M, there are definite values $L(M)$ and $R(M)$ at which the operator η is applicable.

A definite biological role of a protein P_i is performed by a set of definite amino acid residues in the sequence for this protein. These amino acid residues form sites, i.e., certain sections in the three-dimensional structure of a protein that are responsible for performing this biological role of the protein. In the framework of the developed formalism, such a site, which uniquely corresponds to a definite biological role described by a term t_j, is a certain set of leading positions $S_j(P_i) = \{i_i^i, \ldots, i_k^i\}$. The dimension of $S_j(P_i)$ will be understood as the number of positions in it: $|S_j(P_i)| = k$. The length of the site will be understood as the length of the section of the sequence for the protein P_i that covers this site, i.e., $[S_j(P_i)] = (i_k^i - i_1^i + 1)$.

In the proposed formalism, any set of leading positions can be described as one leading position and a set of positions corresponding to a certain system of masks M. Therefore, the localization of a biological function in an amino acid sequence is described using a certain mask. Assume that t is an arbitrary term, $t \subset T$. Let us choose an M such that, for this t, the condition

$$\forall_{c^t} P, [S_t(P)] < [M_\Sigma(M)] \tag{6}$$

is fulfilled.

The hypothesis of locality of the studied problem is formulated below as a hypothesis of the existence of a certain local function. According to Theorem 2, the sought for function F can be uniquely represented as a set of t-classifiers $f_t(P)$. The correct local classifier will be understood as a function f_t^{loc} such that

$$\forall_T t \forall_{\Pr} P \exists_{k=L(M)+1}^{|P|-R(M)} k: f_t^{\text{loc}}(\eta(k, M_\Sigma, P)) = t(P). \tag{7}$$

Since the problem of annotation can be reduced to m problems of constructing exact t-classifiers f_{t_j}, henceforth, we will consider the problem in the form corresponding to condition (7).

The condition of solvability of a local t-classifier f_t^{loc} will be understood as the following condition:

$$\begin{gathered}\forall_{j=1}^{m} j \forall_{\Pr}(P_1, t_j(P_1)), (P_2 t_j(P_2)) \forall_{i,k\in N}(i,k): \\ \eta(i, \hat{M}_\Sigma(M), P_1) = \eta(k, \hat{M}_\Sigma(M), P_2) \\ \Rightarrow t_j(P_1) = t_j(P_2).\end{gathered} \tag{8}$$

Theorem 3. *A local t-classifier f_t^{loc} exists if and only if the condition of solvability for f_t^{loc} is fulfilled.*

The proof is obvious and is conducted ad absurdum.

An important corollary of this theorem is the existence of solvability at separate masks from M. Let us introduce the condition of local solvability with the use of separate masks as

$$\begin{gathered}\forall_{j=1}^{m} j \forall_{\Pr}(P_1, t_j(P_1)), (P_2, t_j(P_2)) \forall_{i,k\in N}(i,k) \\ \times\left(\forall_{l=1}^{|M|} \hat{m}_l: \eta(i, \hat{m}_l, P_1) = \eta(k, \hat{m}_1, P_2)\right) \\ \Rightarrow t_j(P_1) = t_j(P_2).\end{gathered} \tag{8'}$$

Corollary. *A local t-classifier exists under condition* (8').

Any two subwords such as $\nu^1 = \eta(i, \hat{M}_\Sigma(M), P_1)$ and $\nu^2 = \eta(k, \hat{M}_\Sigma(M), P_2)$ in (8) are equal if the letters at each positions coincide. In other words, for $\nu^1 = \{\nu_1^1, \nu_2^1, \ldots, \nu_m^1\}$ and $\nu^2 = \{\nu_1^2, \nu_2^2, \ldots, \nu_m^2\}$, the expression $\eta(i, \hat{M}_\Sigma(M), P_1) = \eta(k, \hat{M}_\Sigma(M), P_2)$ corresponds to the system of equalities $S \equiv \{\forall\mu \in M_\Sigma(M): \nu_{i+\mu}^1 = \nu_{j+\mu}^2\}$. Any $\hat{m}_k \subset M_\Sigma(M)$ corresponds to the subsystem of equalities $s_k \equiv (\eta(i, m_k, \vec{V}_1) = \eta(j, m_k, \vec{V}_2))$. Since $M_\Sigma(M) = \bigcup_{k=1}^{|M|}\hat{m}_k$, $S = \bigcup_{k=1}^{|M|}s_k$ and the equivalence of the expressions $\eta(i, \hat{M}_\Sigma(M), \vec{V}_1) = \eta(j, \hat{M}_\Sigma(M), \vec{V}_2)$ in (8) and $\left(\forall_{k=1}^{|M|}\hat{m}_k: \eta(i, \hat{m}_k, \vec{V}_1) = \eta(j, \hat{m}_k, \vec{V}_2)\right)$ in (8') is evident.

In the modern theory of recognition [4–12], special attention is paid to the regularity of the problem of recognition, because this property of the problem

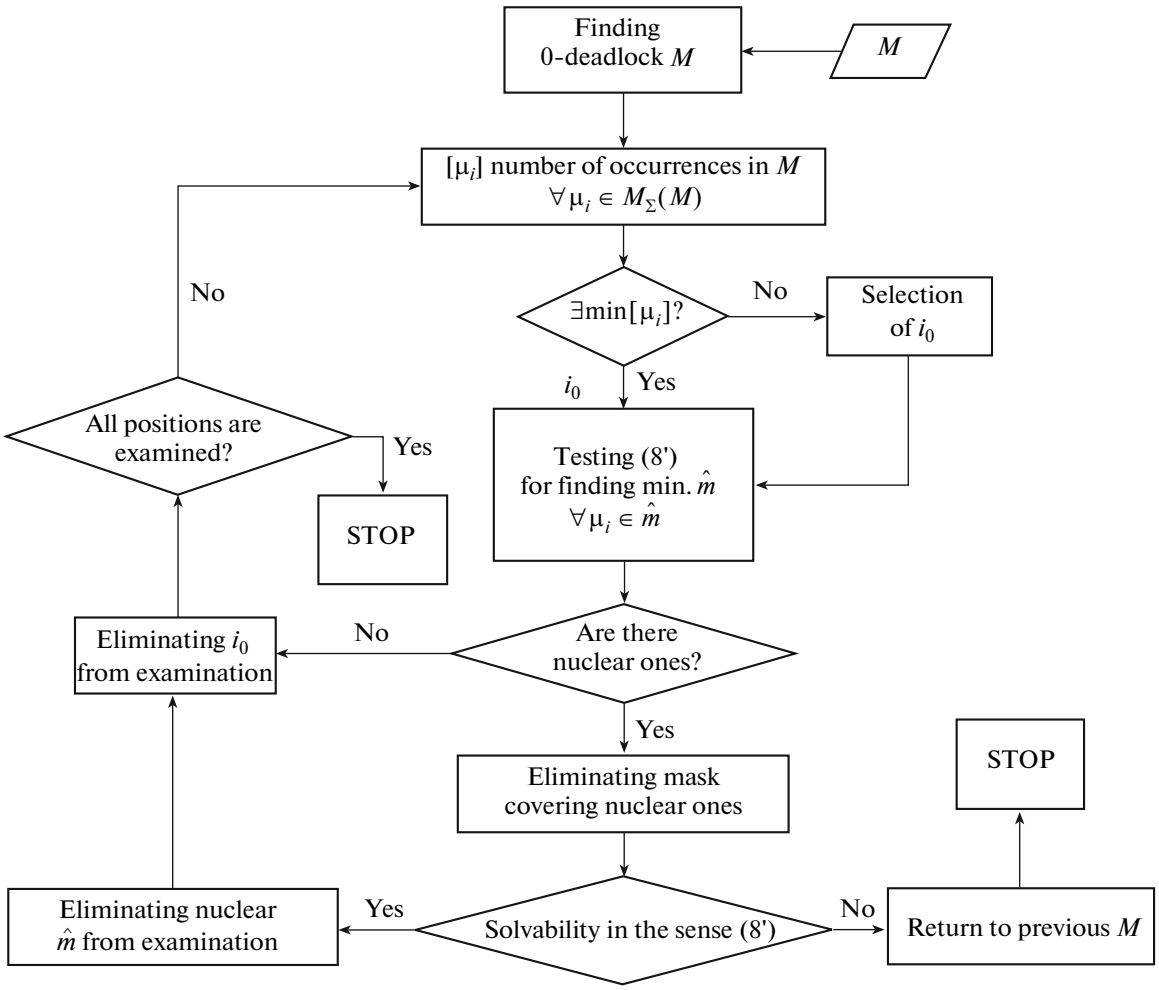

Fig. 3. A variant of the algorithm for finding irredundant systems of masks.

guarantees its solvability. For the problem in form (7), locality is defined as

$$\underset{\Pr}{\forall}(P_1, t_j(P_1)), (P_2, t_j(P_2)) \underset{i,k \in N}{\forall}(i, k): \quad \eta(i, \hat{M}_\Sigma(M), P_1) \neq \eta(k, \hat{M}_\Sigma(M), P_2). \tag{9}$$

Condition (9) is satisfied if the selected united mask $M_\Sigma(M)$ has a sufficiently large dimension. For the problem in local form (7) to become regular, it suffices that $|\hat{M}_\Sigma(M)| \geq 7$, because the existing sets of precedents are comparatively small, i.e., $\sum_{\Pr}|P| \ll |A|^7$.

The biological function of a protein is localized if the length of the corresponding site is much smaller that that of the amino acid sequence of this protein. In the proposed formalism, the biological function corresponding to the term t can be considered as localized on the set of precedents if, for $\hat{m} = \hat{M}_\Sigma(M)$, the following condition is fulfilled:

$$\underset{\Pr}{\forall} P \overset{|P|}{\underset{i=1}{\exists}} \eta(i, \hat{m}, P), [\hat{m}] \ll |P|: \quad \eta(i, \hat{m}, P) \subseteq P \Rightarrow P \in c^t. \tag{10}$$

Then, the degree of localization of a biological function in a given protein P will be understood as the ratio between the length of the mask $\hat{m}$ and the length of the amino acid sequence for P:

$$\mathrm{loc}(t, P) = \frac{[\hat{m}]}{|P|}. \tag{11}$$

A prior estimate of the degree of localization of an arbitrary biological function of a protein described by the term t_j is practically impossible: for the existing sets, regularity is attained for $|\hat{M}_\Sigma(M)| \leq 7$. Therefore,

using $|\hat{M}_\Sigma(M)|$, it is impossible to make at least rough estimations of the degree of localization. For an arbitrary t_j, the degree of localization can be estimated (i) using a significantly greater set of experimental data or (ii) by constructing a correct t_j-classifier and analyzing local solvability of the problem for different lengths of $\hat{M}_\Sigma(M)$.

Conditions (8) and (8') make it possible to conduct experiments on estimating the solvability of local t-classifiers. The most important parameter in conducting these experiments is the system of masks M. Therefore, it is necessary to consider the possibilities of varying M.

4. MONOTONICITY OF THE CONDITION OF SOLVABILITY OF LOCAL T-CLASSIFIERS

The varying of M consists in adding and removing separate masks. In the general case, condition (8') of solvability of local t-classifiers is not monotonic with respect to M; i.e., the existence of solvability for M does not imply solvability for an arbitrary M' such that $M \in M'$. In other words, we cannot exclude the possibility of finding a mask $\hat{m} \notin M$ such that, after including it into M, the quantities $l(M)$ and $r(M)$ will change so that the conditions of solvability will be violated because the requirement of $(l(M), r(M))$-correctness will not be fulfilled. Therefore, it is appropriate to study the monotonicity of the condition of solvability under the condition that $l(M)$ and $r(M)$ are invariable.

In the general case, condition of solvability (8') is not monotonic as well and, for $M' \subseteq M$, the existence of solvability of the problem $Z(\Pr, M)$ does not imply the solvability of the problem $Z(\Pr, M')$, where M' is obtained by removing masks from M.

In the general case, the system of masks M for which the problem is solvable can be redundant in the sense that the solvability persists after removing some masks from M. It is especially important to consider monotonicity in finding irredundant systems of masks. This question was thoroughly studied in [3] in connection with secondary structure recognition. The key conceptions of 0-deadlockness, deadlockness, and nuclearness of a system of masks have been introduced:

M is 0-deadlock if condition (8') is satisfied for M but not satisfied for any $M' \subset M$ such that $M_\Sigma(M') \subset M_\Sigma(M)$;

M is deadlock if condition (8') is violated for any $M' \subset M$;

$$M \text{ is nuclear if } \forall_{i=1}^{|M|}\, i\, \exists_{\hat{m}_i}\mu\colon \forall_{j=1}^{|M|,\, i\neq j}\, j(\mu \in \hat{m}_j).$$

In [3], it has been shown that the exhaustive search for systems of masks can be reduced by analogy with the search for minimum DNF (disjunctive normal forms) [16] and principles of constructing algorithms of search for irredundant systems have been proposed.

A variant of the algorithm developed on the basis of the principles set forth in [3] is presented in Fig. 3. At the first stage, a 0-deadlock system of masks is found. Then, less redundant systems of masks are sought on the basis of nuclearness: a nuclear subsystem enters into all deadlock systems of masks, and masks covered by a nuclear subsystem are eliminated.

CONCLUSIONS

A formalism for the analysis of solvability and locality of the genome annotation problem has been proposed. It has been shown that the annotation problem can be reduced to solving a number of simpler problems on membership in classes of biological functions of proteins by constructing correct local t-classifiers. The latter is performed by virtue of finding irredundant systems of masks satisfying the condition of solvability of local t-classifiers on the basis of a consistent set of precedents. An algorithm for conducting experiments on estimating the solvability of local classifiers has been formulated.

ACKNOWLEDGMENTS

This work was supported by the Russian Foundation for Basic Research, project nos. 09-07-12098, 09-07-00212-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. K. V. Rudakov and I. Yu. Torshin, "On Solvability of the Formal Problem of Protein Secondary Structure Recognition," in *Mathematical Methods of Pattern Recognition-14, Suzdal, September 21–25, 2009*.
3. K. V. Rudakov and I. Yu. Torshin, "Questions on Solvability of the Problem of Protein Secondary Structure Recognition," Informatics and its Application **4** (2), (2010).
4. Yu. I. Zhuravlev, "Well-Posed Algebras over Sets of Ill-Posed (Heuristic) Algorithms. I," Kibernetika, no. 4, 5–17 (1977)
5. Yu. I. Zhuravlev, "Well-Posed Algebras over Sets of Ill-Posed (Heuristic) Algorithms. II," Kibernetika, no. 6, 21–27 (1977)
6. Yu. I. Zhuravlev, "Well-Posed Algebras over Sets of Ill-Posed (Heuristic) Algorithms. III," Kibernetika, no. 2, 35–43 (1978)
7. Yu. I. Zhuravlev, "On the Algebraic Approach to Solving Problems of Recognition and Classification," in *Problems of Cybernetics* (Nauka, Moscow, 1978), issue 33, pp. 5–68 [in Russian].

8. Yu. I. Zhuravlev and K. V. Rudakov, "On the Algebraic Correction of Procedures for Data Processing," in *Problems of Applied Mathematics and Informatics*, (Nauka, Moscow, 1987), pp. 187–198 [in Russian].
9. K. V. Rudakov, "Universal and Local Restrictions in the Problem of Correcting Heuristic Algorithms," Kibernetika, no. 2, 30–35 (1987).
10. K. V. Rudakov, "Completeness and Universal Restrictions in the Problem of Correcting Heuristic Classification Algorithms," Kibernetika, no. 3, 106–109 (1987).
11. K. V. Rudakov, "Symmetric and Functional Restrictions in the Problem of Correcting Heuristic Classification Algorithms," Kibernetika, no. 4, 73–77 (1987).
12. K. V. Rudakov, "On Application of Universal Constraints for Studying Classification Algorithms," Kibernetika, no. 1, 1–5 (1988).
13. *Gene Ontology: Tool for the Unification of Biology*, in (The Gene Ontology Consortium, 2000; Nature Genet.25: 25–29).
14. D. P. Hill, A. P. Davis, J. E. Richardson et al., "Program Description: Strategies for Biological Annotation of Mammalian Systems: Implementing Gene Ontologies in Mouse Genome Informatics," Genomics **74** (1), 121–128 (May 15, 2001).
15. N. Hulo, C. J. Sigrist, V. Le Saux, et. al., "Recent Improvements to the PROSITE Database," Nucleic Acids Res. **32** (Database issue), 134–137 (January 1, 2004).
16. Yu. I. Zhuravlev, "Set Theoretical Methods in Logic Algebra," Problemy Kibernetiki **8** (1) 25–45 (1962).

Ivan Yur'evich Torshin. Born 1972. Graduated from the Chemistry Department of Moscow State University in 1995. Received PhD degree in 1997. Visiting professor at Thomas Jefferson University and at Georgia University (United States) 1999–2004. Since 2007, Leading Researcher in the Russian Satellite Center of the Trace Element Institute for UNESCO. Over 75 publications in Russian and international science journals on computational biology, chemistry, medicine, and informatics.