TRANSCRIPTION FACTOR BINDING SITE RECOGNITION BY REGULARITY MATRICES BASED ON THE NATURAL CLASSIFICATION METHOD

Vityaev E.E.*[1,2], Lapardin K.A.[2], Khomicheva I.V.[3], Proskura A.,L.[3]
[1]Institute of Mathematics SB RAS, Novosibirsk, Russia
[2]Novosibirsk State University, Novosibirsk, Russia
[3]Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia
email: vityaev@bionet.nsc.ru, fax: +7-383-333-25-98
*Corresponding author

Abstract

A principally new approach to the classifications of nucleotide sequences based on the "natural" classification concept is proposed. As a result of "natural" classification of the nucleotide sequences, we obtain regularity matrices, where nucleotides are interconnected by regularities. Method, algorithm and software system DNANatClass for performing the "natural" classification have been developed. Experimental results comparing weight matrices with regularity matrices are presented. In this experiment, site recognition by regularity matrices appears to be more accurate than by weight matrixes.

1. Introduction

Position weight matrix is the most common method for transcription factor binding site (TFBS) recognition. In this paper we present the regularity matrices that arise from the concept of "natural" classification in its application to nucleotide sequences. The concept of "natural" classification was investigated and developed in the previous papers [25, 27-29]. The main property of regularity matrices is that each of the nucleotides A, T, G, C in each position of the matrix is characterized by its regularities connecting it with nucleotides in other positions, whereas weight matrices estimate the contribution of each nucleotide taken separately without any interconnectivity.

Numerous principles of constructing classifications are currently known. Classifications are based on the hypothesis of compactness and various measures of closeness in a feature space, on resemblance of standards, supertargets, various criteria of classification quality and quality functionals, separation of distribution mixtures, etc. [5]. In contrast to the above-listed classifications, the objective of "natural" classification is discovering the laws of nature. There are different definitions of "natural" classification put forward by naturalists in different times (see overview [32]). We propose a definition of the "natural" classification that is in accordance with the definitions by naturalists: "Objects should be divided into classes in accordance with the regularities satisfied by the objects. Objects of one class should obey one group of regularities, and objects of different classes should obey different groups of regularities. Objects of one class should also possess some integrity which is understood as mutual prediction of object properties" [25].

2. Criteria of "natural" classification

Zabrodin [32] systematized criteria of the "naturalness" of classification, put forward by naturalists at various times. Let us overview these notions. Below, we will provide a definition of natural classification and systematics, explaining all these features.

1. V.Yu. Zabrodin [32]: "A natural classification is such and only such classification that expresses a law of the nature".

2. Wavell's criterion [16]: “The more general sentences can be deduced from a classification, the more natural is it”.
3. A.A. Lyubishchev's criterion [32]: “The perfect system is such one where all traits of the object are determined by its position in the system. The closer is a system to this ideal, the less artificial is it. A system should be called natural if the number of features of an object functionally linked to its position in the system is the greatest. Ideally, these are all its features”.
4. S.A. Schreider [20]: “Two types of regularities can be found in the manifold of object forming a "natural" classification:
   - correlations, linking the "brief" description of the archetype, sufficient for diagnostics of the object's belonging to the corresponding class, with the "full" description. In fact, they are laws, which allow prediction of all features of an object on the base of the fact that it belongs to a certain natural class.
   - rules illustrating how features of objects change with transition to neighboring classes. Just these rules ensure the possibility of transfer of knowledge from a single object to all objects belonging to the class and, with somewhat more difficulties, to objects of neighboring classes”.
5. E.S. Smirnov's criterion [21]: “The taxonomical problem is in "indication": we should pass from an infinitely great number of attributes to their limited number, which would replace all the attributes”.
6. L. Rutkovsky [19]: “The more essential features are things to be compared similar in, the more is it probable that they are similar in other respects”.
7. V.L. Kozhara [10] “Therefore, it is natural to suggest that the stability of taxonomic structures (TS) depends on the number of bases N (number of features), the more, the more stable. With increasing N, the stability is likely to increase with retardation, so that at a certain sufficiently large N it scarcely increases”.
8. The following principle of constructing "natural" classifications was put forward in [25, 27-29]: “Division into classes should be made so that objects of one class would obey the same regularities”.

3. Definition of Natural Classification and Systematics

Let us define a regular model M_a of object a. It includes the set of all concepts, attributes, parameters, and values that can be applied to the object and take certain values on it: (truth, numerical, probabilistic, etc.). Let Ω_a be a list of attributes and their values. Also let us define a system of laws including analytical expressions describing the linkage between some concepts, laws establishing interrelations between values, the set of induction dependences (regularities) establishing interrelations between the potentially infinite set of attributes. Isolate from the system of laws the subset Z_a of laws and regularities that allow deduce some concepts, attributes, parameters, and values from others for this object. Here we do not specify the form of the laws, because various forms of regularities will be considered below. Not all laws and regularities are included. For example, regularities of type IF…THEN… cannot be applied to an object and cannot be used for prediction if the premise of the rule is not fulfilled on the object. Subset Z_a provides the regular structure of the object. We call the model $M_a = \langle \Omega_a, Z_a \rangle$ the *regular model of the object*.

Consider a certain class ς of objects. Define the regular model of the class $M_\varsigma = \langle \Omega_\varsigma, Z_\varsigma \rangle$ as an association of all regular models of objects of the class. In this association, attributes with identical values remain unchanged, and other attributes form combinations of values.

Remember that, according to Smirnov's criterion “The taxonomical problem is in "indication": we should pass from an infinitely great number of attributes to their limited number, which would replace all the attributes”. Note that the diversity of classes is incommensurably less than the diversity of combinations of attribute values; hence, there are lots of regular linkages between attribute values. For example, if the number of classes equals 16000 and the attributes are binary, only about 14 attributes can be independent: $16*1024 = 2^4*2^{10} = 2^{14}$. Scientists classifying animals, plants,

soils, etc. can use an enormous, potentially infinite set of attributes and parameters. However, only about ten attributes can be independent to an extent, and other attributes can be deduced from regularities. Search for attributes allowing prediction of all others is the problem of indication. In the regular model of class M_a, such attribute values are generative sets of attribute values. From the set of values of generative attributes $\langle x_{i1}=x^{i1}{}_{j1}, x_{i2}=x^{i2}{}_{j2}, \dots, x_{im}=x^{im}{}_{jm}\rangle$ and regularities from Z_ς, values of all other attributes in M_a can be predicted for objects of the class. Obviously, the set of values of generative attributes is nonuniquely defined.

Assume that we know all classes $\{\varsigma_{i\in I}\}$ and all regular models of these classes M_{ς_i}. Let us consider the problem of constructing systematics. We look for such generative sets of attributes $x_{i1}, x_{i2}, \dots, x_{iN}$ whose set of values is generative for each class of $\{\varsigma_{i\in I}\}$. This means that for each class there is a set of values of these attributes $\langle x_{i1}=x^{i1}{}_{j1}, x_{i2}=x^{i2}{}_{j2}, \dots, x_{iN}=x^{iN}{}_{jN}\rangle$ that is generative for this class. The attribute set $S = \langle x_{i1}, x_{i2}, \dots, x_{iN}\rangle$ is called system-forming if for each class of $\{\varsigma_{i\in I}\}$ the generative sets of system-forming attribute values $\langle x_{i1}=x^{i1}{}_{j1}, x_{i2}=x^{i2}{}_{j2}, \dots, x_{iN}=x^{iN}{}_{jN}\rangle$ are different. In this case, each class is unambiguously defined by the set of values of system-forming attributes. Obviously, sets of system-forming attributes are also nonuniquely defined. The task of the scientist is just to find the most compact and informative set of system-forming attributes. The essence of systematics is to present the pattern of change of system-forming attribute values with transition from objects of one class to objects of another class in any way, e.g., by tabulation. The change of system-forming attribute values can obey a law; therefore, it is appropriate to present systematics in a special way so that this law would be clearly expressed. Let us define a *regular model of a systematics* as $M_S = \langle S, Z_S\rangle$, where S is the set of system-forming attributes, and Z_S is the law of the systematics: the law of change of attribute values from S with transition from class to class. Generally, it can be expressed by a table, each of whose rows contains the name of a class and the set of system-forming attribute values. A certain class $M_\varsigma = \langle \Omega_\varsigma, Z_\varsigma\rangle$ corresponds to each set of system-forming attribute values S. Then the systematics law Z_S is a metalaw with regard to the regularities of class Z_ς. The systematics law Z_S is linked to laws of classes as indicated in the definition of systematics put forward by S.A. Schreider. Regularities of the first type are regularities of the corresponding class Z_ς, and regularities of the second type are the systematics law Z_S.

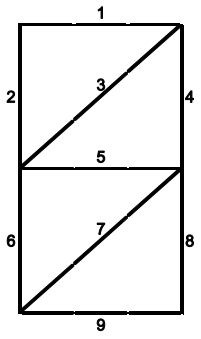
Our definition meets Lyubishchev's criterion because all attributes of an object are determined by interaction of two types of laws: first, the systematics law Z_S, which determines the class of the object and values of system-forming attributes of this class from the position of the object in the system, and then, other properties of the object are deduced from the regularities of class Z_ς.

Let us define a systematics as the set $\Sigma = \langle S, Z_S, \{Z_{\varsigma i}\}_{i\in I}\rangle$. The goal of a scientist is to choose the best system, explaining the properties of objects in the simplest way.

Now let us suggest that we do not know the division of objects into classes. How should we construct a systematics in this case? The above definition is also applicable to the case when only regular models of objects are known. The task of constructing systematics reduces to finding such division of the set of objects into classes that the systematics based on these classes is the best: all properties of an object are predictable from its position in the systematics.

4. Example of systematics construction. Recognition of postal code digits.

Consider ten postal code digits: 0, 1, 2, 3, 4, 5, 6, 7, 8, and 9. Number attributes of indices as shown in the fig. 1. Determine predicates $P_1, \dots, P_9$ implying the presence (true) or absence (false) of the i^{th} element in the image of a digit. Then postal code digits can be represented as shown in the fig. 2. We treat the digits as classes $\{\mathfrak{C}_{i\in I}\}$, $I = \{0,..,9\}$. Let us find regular models of these classes. For this purpose, we find the set of regularities for these digits in the form of laws, whose definition is provided in definition 2 below.



Consider the *empirical system* M = {A, Q}, Q = $\{P_i,\ldots, P_9\}$, the set of predicates of the signature Ω = $\{P_1,\ldots, P_9\}$, determined on $A=\{0,..,9\}$.

Find all laws (definition 2) on the empirical system M = {A, Q}. This results in 3743 laws found by program [29]. Then for each digit laws obeyed on it are determined. For example, 529 regularities are obeyed on digit 2.

For each digit, a regular model of the class can be obtained from the set of attributes and the set of laws. Find minimal generative sets for each class. For example, in the case of digit 2, it is the set $\{P_2,P_3\}$. The values of other attributes are restored according to the following laws:

Fig. 1. Attributes of postal code digits.

$\neg P_3\&\neg P_2 \Rightarrow P_1$
$\neg P_3\&\neg P_2\&P_1 \Rightarrow P_4$
$P_4\&\neg P_2\&P_1 \Rightarrow \neg P_5$
$\neg P_3\&\neg P_2\&P_1 \Rightarrow \neg P_6$
$\neg P_6\&\neg P_5\&P_4\&P_1 \Rightarrow P_7$
$P_7\&\neg P_3\&P_1 \Rightarrow \neg P_8$
$P_8\&\neg P_6\&\neg P_5\&\neg P_2 \Rightarrow P_9$

	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8	P_9
0	1	1	0	1	0	1	0	1	1
1	0	0	1	1	0	0	0	1	0
2	1	0	0	1	0	0	1	0	1
3	1	0	1	0	1	0	1	0	0
4	0	1	0	1	1	0	0	1	0
5	1	1	0	0	1	0	0	1	1
6	0	0	1	0	1	1	0	1	1
7	1	0	1	0	0	1	0	0	0
8	1	1	0	1	1	1	0	1	1
9	1	1	0	1	1	0	1	0	0

Fig. 2. Postal code digits representation.

As mentioned above, generative sets are defined nonuniquely. For example, $\{P_5,P_7\}$ is also a generative set, from which the values of other attributes are restored according to the following laws:

$P_7\Rightarrow P_1$
$P_7\&\neg P_5\Rightarrow\neg P_2$
$P_7\&\neg P_5\Rightarrow P_4$
$P_4\&\neg P_2\& P_1\Rightarrow\neg P_3$
$\neg P_3\&\neg P_2\Rightarrow P_9$
$P_4\&\neg P_2\Rightarrow\neg P_6$
$P_9\&\neg P_6\& P_4\Rightarrow \neg P_8$

From the number of laws it is evident that the regular model of digit 2 for the generative $\{P_5,P_7\}$ is simpler. It looks as follows: $M_2 = \langle\{1,0,0,1,0,0,1,0,1\}, \{P_7,\neg P_5, P_7\Rightarrow P_1, P_7\&\neg P_5\Rightarrow\neg P_2, P_7\&\neg P_5\Rightarrow P_4, P_4\&\neg P_2\& P_1\Rightarrow\neg P_3, \neg P_3\&\neg P_2\Rightarrow P_9, P_4\&\neg P_2\Rightarrow\neg P_6, P_9\&\neg P_6\& P_4\Rightarrow \neg P_8\}\rangle$. The minimum generative set for each digit allows constructing its regular model.

We are coming to the construction of the regular model of systematics. Its law Z_S can be expressed as a table each of whose rows presents a digit of the class and values of generative attributes. For choosing the minimum generative set of the systematics, consider various combinations of generative sets for classes. The generative set with the maximum number of attributes of all minimum generative sets for all digits is that for digit 8 (The minimum number of attributes is 3). Thus, the generative set of the systematics consists of no less than three attributes. The minimum generative sets of classes do not necessarily allow determining the minimum generative set of the systematics. For example, the minimum generative sets for digit 3 are $\{P_3, P_7\}$ and $\{\neg P_4, P_7\}$, whereas the generative sets containing of three attributes for digit 3 do not include the 7^{th} attribute. Hence, not only all generative sets of two attributes but also generative sets of three attributes deserve consideration.

As $2^3 = 8$ is less than the number of classes, three attributes are insufficient for unambiguous restoration of a class. Therefore, we consider all possible combinations of four attributes. We find that the minimum generative set of attributes for the systematics is $\{P_4, P_5, P_6, P_7\}$. In this case, it is determined by a single way.

The systematics for postal code digits can be expressed by figure 3, where values of attributes $\{P_4, P_5, P_6, P_7\}$ and the minimum generative sets of attributes are shown for each digit.

	P_4	P_5	P_6	P_7	Generative sets
0	1	0	1	0	$\{P_4, P_5, P_6\}$
1	1	0	0	0	$\{P_5, P_6, P_7\}$
2	1	0	0	1	$\{P_5, P_7\}$
3	0	1	0	1	$\{P_4, P_7\}$
4	1	1	0	0	$\{P_4, P_5, P_6, P_7\}$
5	0	1	0	0	$\{P_4, P_6, P_7\}$
6	0	1	1	0	$\{P_4, P_5, P_6\}$
7	0	0	1	0	$\{P_4, P_5\}$
8	1	1	1	0	$\{P_4, P_5, P_6\}$
9	1	1	0	1	$\{P_4, P_5, P_7\}$

Fig. 3. Systematics for postal code digits.

From attribute values, the class is determined, and from the minimum generative set, the values of all other attributes are restored.

5. Subject domain theory discovery

In this section, we introduce the notion of the subject domain theory. Let us introduce the first-order logic L of the signature $\Im = \langle P_1,..., P_k, c_1,\ldots,c_n\rangle$, where $P_1,..., P_k$ are the predicate symbols of the arity $n_1..., n_k$ and $c_1,\ldots,c_n$ are constants. An *empirical system* is taken to mean a finite model $M = \langle B,W\rangle$ of the signature $\Im$, where B is the basic set of the empirical system, $W = \langle P_1,...,P_k\rangle$ is the tuple of predicates of the signature $\Im$ defined on B.

Let us represent the *subject domain* by an empirical system $M = \langle A, W\rangle$. By the *subject domain theory* Th(M) we mean the set of all sentences of L that are true on M.

The task of the discovery of the *subject domain theory* consists in the *discovery* of the theory Th(M). We will assume that the theory Th(M) is universally axiomatizable.

It is known that the theory Th(M) of universal formulas can be reduced by logically equivalent transformations to the set of rules:

$$C = (A_1\&...\&A_k \Rightarrow A_0), k \geq 0, \quad (1)$$

where $A_0, A_1, \dots, A_k$ are literals. Therefore, we can assume that the theory Th(M) is a set of rules. Thus, the task of discovery of the subject domain theory is reduced to one dealing with the discovery of rules (1).

What can we say about the truth of rules (1) on the empirical system M, if guided by the logical analysis of axioms?

- Rule $C = (A_1\&...\&A_k \Rightarrow A_0)$ is true on M, if the premise of the rule is always false on M;
- Rule C is true on M, if some of its logically stronger subrule, containing only a part of the premise and the same conclusion, is true on M.

Let us clarify the logically stronger subrules from which the truth of the rule follows.

Theorem 1. [26]. Rule $C = (A_1\&...\&A_k \Rightarrow A_0)$ logically follows from any rule of the form:

$(A_{i1}\&...\&A_{ih} \Rightarrow A_0)$, where $\{A_{i1},...,A_{ih}\} \subset \{A_1,...,A_k\}$, $0 \le h < k$, and

$(A_{i1}\&...\&A_{ih} \Rightarrow A_0) \vdash (A_1\&...\&A_k \Rightarrow A_0)$,

$\vdash$ – provability in propositional calculus.

Definition 1. By a *subrule* of rule C is meant a logically stronger rule defined in theorem 1 for rule C.

Corollary 1. If a subrule of rule C is true on M, then rule C is also true on M.

Definition 2. By the *law* on the set of experimental results M, is meant rule C, which is true on M, and none of its subrules is true on M.

Let L be the set of all laws on M. It can be proven that from the set of laws L the subject domain theory Th(M) is inferred.

Theorem 2. [26] $L \vdash Th(M)$.

Therefore, the task of the discovery of the subject domain theory Th(M) is reduced to the discovery of the set of laws L.

6. Probabilistic laws on PRM

Let us generalize the notion of law on probabilistic case. Let us define the probability on $M = \langle A, W\rangle$ and logical expressions. For the sake of simplicity we introduce a discrete probability function on A as a mapping $\mu: A \to [0,1]$ such that [7]

$$\sum_{a\in A} \mu(a)=1 \text{ and } \mu(a) \neq 0,\ a \in A.$$
$$\mu(B)=\sum_{b\in B} \mu(b),\ B \subseteq A \qquad (2)$$

The discrete probability function μ^n on the product $(A)^n$ will be thereby defined by taking

$$\mu^n(a_1,...,a_n)=\mu(a_1)\times...\times\mu(a_n)$$

A more general case of the probability function μ definition is considered in [7]. Let us define the interpretation of language L on the empirical system $M = \langle A, W\rangle$ as mapping $I: \Im \to W$, which associates with every signature symbol $P_j \in \Im$, $j = 1,...,k$, the predicate P_j from W of the same arity. Let $X = \{x_1, x_2, x_3, ...\}$ be the variables of language L. By the valuation ν we mean the function ν: $X \to A$.

Let us define the probability for sentences of language L. Let $U(\Im)$ be the set of all atomic formulas of language L of the form $P(x_1,\dots,x_n)$; $\Re(\Im)$ is the set of all the sentences of the language L, obtained by the closure of the set $U(\Im)$ relative to the logical operations $\&,\vee,\neg$. By $\nu I\varphi$, $\varphi \in \Re(\Im)$ will be defined the formula φ where the predicate symbols from $\Im$ are substituted by the predicates from W by interpretation I and variables of the formula φ are substituted by objects from A by the valuation ν. The probability η of the sentence $\varphi(x_1,\dots,x_n) \in \Re(\Im)$ on M is defined as follows

$$\eta(\varphi) = \mu^n(\{(a_1,\dots,a_n) \mid M \vDash \nu I\varphi,\ \nu(x_1) = a_1,\dots,\ \nu(x_n) = a_n\}),$$
$$\text{where } \vDash \text{ is the truth on M} \qquad (3)$$

Now we revise in terms of probability the concept of law on M. Let us do it in such a way that the concept of the law on M would be a particular case of this definition. The law is a rule true on M whose subrules are false on M. Let us revise the concept of the law on the PR_M. The law is such a rule true on M, which cannot be made simpler or logically stronger without losing the truth. This property of the law "not to be simplified" allows stating the law not only in terms of truth but also in terms of probability.

Definition 3. By a probabilistic law on PR_M, we designate a rule $C = (A_1\&...\&A_k \Rightarrow A_0)$, $k \geq 0$, satisfying the condition: the conditional probability of the rule $\eta(A_0/A_1\&...\&A_k)$, $\eta(A_1\&...\&A_k) > 0$ is strictly more than the conditional probability of each of its subrules.

We denote the set of all probabilistic laws on M as LP.

Proposition 1. $L \subset LP$.

Definition 4. By the Strongest Probabilistic Law (SLP-rule) on PR_M, we designate such a probabilistic law $C = (A_1\&...\&A_k \Rightarrow A_0)$, which is not a subrule of any other probabilistic law.

We define as SPL the set of all SPL-rules.

Proposition 2. $L \subset SLP \subset LP$.

7. Semantic Probabilistic Inference of the Set of Laws L and LP

In this section we define the semantic probabilistic inference of the sets of laws L, probabilistic laws LP and SLP.

Definition 5. By the *semantic probabilistic inference* (SP-inference) of some SPL-rule we mean such a sequence of probabilistic laws, which we designate as $C_1 \sqsubset C_2 \sqsubset ... \sqsubset C_n$, that:

$$C_1,C_2,...,C_n \in LP,\ C_n - \text{SPL rule},\ C_i = (A^i_1 \& ... \& A^i_{ki} \Rightarrow G),\ i = 1,2,...n,\ n \geq 1,$$
$$\text{rule } C_i \text{ is a subrule of rule } C_{i+1},\ \eta(C_{i+1}) > \eta(C_i),\ i = 1,2,...n-1, \qquad (5)$$

Proposition 3. Any probabilistic law belongs to some SPI-inference.

Proposition 4. There is some SPI-inference for any SPL-rule.

Corollary 2. For any law from L there is some SPI-inference of that law.

Let us consider the set of all SP-inferences of some sentence G. This set constitutes the semantic probabilistic inference tree of sentence G figure 4.

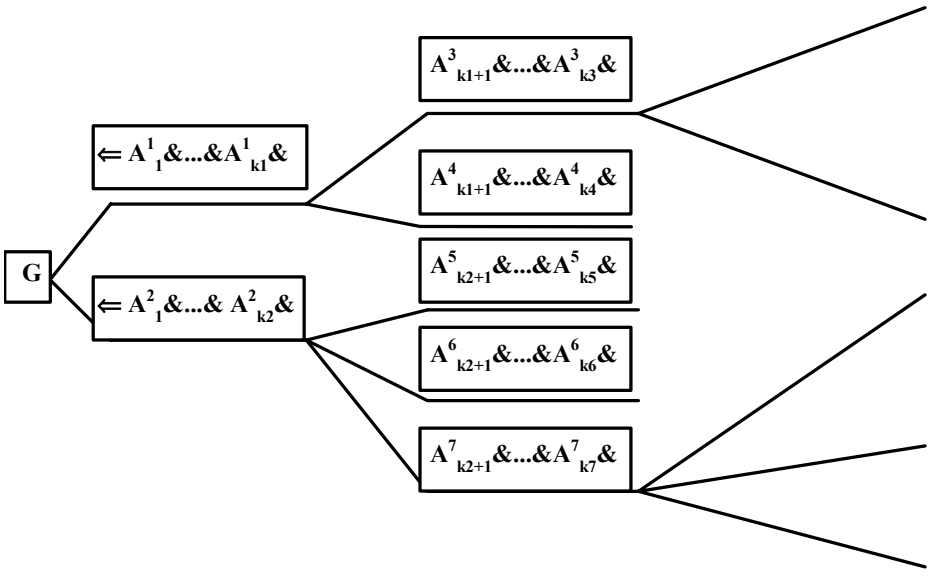


Fig. 4. Semantic probabilistic inference tree.

Definition 6. By the *maximum specific rule* MS(G) of sentence G we mean the SPL-rule of the semantic probabilistic inference tree of sentence G having the maximum value of conditional probability.

The set of all maximum specific rules for any atom G is defined as MSR.

Proposition 5. $T \subset MSR \subset SPL \subset TP$.

8. Models of predictions

The next question that follows the discovery of the subject domain theory Th(M) is how to employ the theory. The main employment of the theory is predictions. Now we consider the models of predictions and the statistical ambiguity problem.

One of the major results of the Philosophy of Science is the so-called *covering law model* that was introduced by Hempel in the early sixties in his famous article 'Aspects of scientific explanation' [8-9]. The basic idea of this covering law model is that a fact is explained/predicted by subsumption under a so-called *covering law,* i.e. the task of an explanation/prediction is to show that a fact can be considered as an instantiation of a law. In the covering law model two types of explanation/predictions are distinguished: *Deductive-Nomological* (D-N) explanations/predictions and *Inductive-Statistical* (I-S) explanations/predictions. In D-N explanations/predictions the law is *deterministic*, whereas in I-S explanations the law is *statistical.*

The Deductive-Nomological explanations/predictions of some observed phenomenon G are inferred by the rule:

$$\frac{\frac{L_1,\ldots,L_m}{C_1,\ldots,C_n}}{G}$$

That satisfies the following conditions:

i. $L_1,\ldots,L_m$ are universally quantified sentences (having at least one universally quantified formula), $C_1,\ldots,C_n$ has no quantifiers or variables;

ii. $L_1,\ldots,L_m, C_1,\ldots,C_n \Rightarrow G$;

iii. $L_1,\ldots,L_m, C_1,\ldots,C_n$ is consistent;

iv. $L_1,\ldots,L_m \not\Rightarrow G$; $C_1,\ldots,C_n \not\Rightarrow G$;

We assume that for deductive-nomological inference of predictions we will use laws from L. In this case, due to theorem 3, we may infer any predictions that follow from the subject domain theory Th(M).

Hempelian inductive-statistical explanations/predictions of some observed phenomenon G are inferred by the analogous rule:

$$\frac{\frac{L_1,\ldots,L_m}{C_1,\ldots,C_n}\ [r]}{G}$$

where [r] is the probability of inference.

In addition to items (i-iv) it satisfies the following Requirement of Maximal Specificity RMS:

v. RMS: All laws $L_1,\ldots,L_m$ are maximally specific.

In Hempel [8-9] the RMS is defined as follows.

An Inductive-Statistical (I-S) inference

$$\frac{\frac{p(G;F) = r}{F(\boldsymbol{a})}\ [r]}{G(\boldsymbol{a})}$$

is an acceptable I-S prediction with respect to a knowledge state K, if the following requirement of maximal specificity is satisfied. For any class H for which the following two sentences are contained in K

$\forall x(H(x) \Rightarrow F(x))$,
$H(\boldsymbol{a})$, (4)

there exists a statistical law p(G;H) = r' in K such that r = r'. The basic idea of RMS is that if F and H both contain the object ***a***, and H is a subset of F, then H provides more specific information about the object ***a*** than F, and therefore the law p(G;H) should be preferred over the law p(G; F).

For inductive-statistical inference of predictions we may use probabilistic laws LP.

9. Requirement of Maximal Specificity. Solution of the statistical ambiguity problem.

Now present the definition of the requirement of maximum specificity for a probabilistic case and the corresponding definition of maximum specific rules that solve the problem of statistical ambiguity.

We suppose that the class H of objects in (4) is defined by some sentence $H \in \Re(\Im)$ of language L. In this case the RMS says that p(G;H) = p(G;F) = r for this sentence. In terms of probability it means that $\eta(G/H) = \eta(G/F) = r$ for any $H \in \Re(\Im)$, satisfying (4).

Definition 7. Rule $C = (F \Rightarrow G)$ satisfies the *Probabilistic Requirement of Maximal Specificity* (PRMS) iff:

the equations $\eta(G/F\&H) = \eta(G/F) = r$ for the rules $C = (F \Rightarrow G)$ and $C' = (F\&H \Rightarrow G)$ follows from: $H \in \Re(\Im)$ and $F(\mathbf{a})\&H(\mathbf{a})$ (in that case the sentence $\forall x(F(x)\&H(x) \Rightarrow F(x))$ is true and $\eta(F\&H) > 0$),

In other words, PRMS means that there is no other sentence $H \in \Re(\Im)$ that increases (or decreases, see lemma 1 below) the conditional probability $\eta(G/F) = r$ of rule C by adding it to the premise.

Lemma 1. [24] If the sentence $H \in \Re(\Im)$ decreases the probability $\eta(G/F\&H) < \eta(G/F)$ then the sentence $\neg H$ increases it: $\eta(G/F\&\neg H) > \eta(G/F)$.

Lemma 2. [24] For any rule $C = (B_1\&...\&B_t \Rightarrow A_0)$, $\eta(B_1\&...\&B_t) > 0$ of the form (2) there is a probabilistic law $C' = (A_1\&...\&A_k \Rightarrow A_0)$ on M which is a subrule of rule C and $\eta(C') \geq \eta(C)$.

Theorem 3. [24] Any MS(G) rule satisfies PRMS.

Corollary 3. [24] Any law on M satisfies the PRMS requirement.

Theorem 4. [24] The I-S inference is consistent for any laws $L_1,\ldots,L_m \in$ MSR.

It follows from the theorem that after discovering the set MSR of all maximum specific rules we can predict by the I-S inference without contradictions.

10. Prediction in the semantic probabilistic inference

The semantic probabilistic inference allows more precise determination of a prediction. Define data D for obtaining predictions as a set of facts (letters) $\{C_1,\ldots,C_n\}$, true on the empirical system M, $D = \{C_1,\ldots,C_n\} = \{\nu IA$, if $M \vDash \nu IA$, and $\neg\nu IA$, if $M \vDash \neg\nu IA \mid A \in U(\Im)\}$.

Definition 8. For any sentence $G(\mathbf{a})$, $G \in \Re(\Im)$, $\mathbf{a} = \langle a_1,\ldots,a_n\rangle$, and data D the rule C of the form (2) is the *rule best for prediction* for the sentence $G(\mathbf{a})$ iff:

1. literals of the premise $\nu I(A_1 \& ... \& A_k)$ of rule C belong to data D under some evaluation ν and $\nu IG = G(\mathbf{a})$;
2. rule C has the maximal value of conditional probability η(C) among all rules that satisfy condition 1;
3. there are no subrules of rule C that satisfy conditions 1,2.

Corollary 4. Any rule C best for prediction for any sentence $G(\mathbf{a})$ and data D is a probabilistic law.

Definition 9. By the *Semantic Probabilistic prediction* (SP-prediction) of some sentence G(**a**), G $\in \Re(\Im)$, **a** = $\langle a_1,\ldots,a_n\rangle$ by data D and the set of probabilistic laws LP, we mean the prediction of that sentence by such a rule C = $(A_1 \,\&\ldots\&\, A_k \Rightarrow G)$ (SP-rule) that:

1. the premise $\nu I(A_1 \,\&\ldots\&\, A_k)$ of the rule C belongs to data D under some evaluation ν and νIG = G(**a**);
2. rule C has the maximum value of conditional probability among the rules that satisfies condition 1 (under some evaluation ν) and it belongs to the semantic probabilistic inference tree of sentence G;
3. if the semantic probabilistic inference tree of sentence G is empty (has no rules) or condition 1 is not satisfied for the rules from that tree under any evaluation ν, then the prediction is not defined;
4. estimation [r] of prediction G(**a**) is defined as the conditional probability of rule C, r = η(C);

The Semantic Probabilistic prediction/explanation may be presented as an I-S inference:

$$\frac{\text{SP-rule C}\quad A_1(\mathbf{a})\&\ldots\&A_k(\mathbf{a})}{G(\mathbf{a})}\ [\eta(C)]$$

Let us present another definition of SP-rule C. We proof that rule C in some sense is the rule best for prediction.

Theorem 5. [23] Any SP-rule C of some sentence G(**a**) by data D and laws LP is the rule best for prediction for the sentence G(**a**) and data D.

Thus, the semantic probabilistic inference not only provides us with the sets of laws LP and L, but also provides us with semantic probabilistic predictions that are fulfilled by the rules best for prediction.

11. Natural classification of subject domain objects

Consider an object (set of objects) **a** = $\langle a_1,\ldots,a_n\rangle$ of an empirical system M. Assume that for these objects $\langle a_1,\ldots,a_n\rangle$ some data are available in the form of a set of literals $D_{\mathbf{a}} = \{A_1,\ldots,A_n\} = \{\nu IA,$ if $M \models \nu IA \parallel \neg\nu IA,$ if $M \models \neg\nu IA \mid A \in U(\Im), \nu\colon X \to \langle a_1,\ldots,a_n\rangle\}$ true on objects from **a**. A set of probabilistic laws from LP can be fulfilled (see definition below) on the data D.

Definition 10. A probabilistic law is applicable to data D if the literals of the premise of the law belong to $D_{\mathbf{a}}$.

Definition 11. A probabilistic law can be fulfilled on data D if the literals of both premises and conclusions of the law belong to $D_{\mathbf{a}}$.

From the data $D_{\mathbf{a}}$ and the set of probabilistic laws LP some facts G(**a**) can be predicted for objects from **a** by semantic probabilistic predictions. For the reason of statistical ambiguity, these predictions can cause contradictions, i.e., predictions of both G(**a**) and ¬G(**a**). Were we sure that each best for prediction rule for the sentence G(**a**) by data $D_{\mathbf{a}}$ is the maximum specific rule, no contradictions would arise according to theorem 4. If there is a contradictory prediction G(**a**) and ¬G(**a**), this means that the maximum specific rules MS(G(**a**)) and MS(¬G(**a**)) for predicting letters G(**a**) and ¬G(**a**) are not simultaneously applicable to the data $D_{\mathbf{a}}$. This means that data $D_{\mathbf{a}}$ lack some information required for prediction of G(**a**), ¬G(**a**).

Regular models of objects and classes should reflect the integrity of objects and classes; therefore, they should not have contradictions.

Definition 12. A regular model $M_{\mathbf{a}} = \langle D_{\mathbf{a}}, Z_{\mathbf{a}}\rangle$ of a class (object) **a** is such data $D_{\mathbf{a}}$ on this class (object) and regularities $Z_{\mathbf{a}} \subset$ LP that:

1. data $D_{\mathbf{a}}$ are closed with regard to semantic probabilistic predictions, i.e., if a sentence G(**a**) or ¬G(**a**) is predicted by semantic probabilistic predictions by data $D_{\mathbf{a}}$ then G(**a**) and/or ¬G(**a**) belong to $D_{\mathbf{a}}$.
2. set $D_{\mathbf{a}}$ is consistent and does not include simultaneously G(**a**) and ¬G(**a**);
3. $Z_{\mathbf{a}}$ is the set of SP rules that predict sentences from $D_{\mathbf{a}}$ by semantic probabilistic predictions by data $D_{\mathbf{a}}$;
4. all letters from $D_{\mathbf{a}}$ are present in at least one rule from $Z_{\mathbf{a}}$.

Define the task of classification of objects of a subject domain as finding of all regular models of objects and classes.

If a classification is made on data (sample) D from an empirical system M, we do not know the probabilities. Probabilistic laws can be found on sample D by checking the Fisher test for contingency tables as described in [12], section 4.8.3. Fisher test. According to data D, we can find a set LP_α of probabilistic laws with a certain level of confidence α, which does not precisely match the set of probabilistic laws LP. Correspondingly, we cannot find the set of all maximum specific rules MSR. We can find only a set SPL_α of probabilistic laws detected with the level of confidence α that have maximum conditional probability values. Therefore, we cannot avoid contradictions, which are indicated in item 2 of class definition 12. In our opinion, the best approximation of definition 12 is the class definition presented in the following section.

12. Classification of DNA sequences

Let us consider the problem of classification and recognition of transcription factor binding sites. Consider a sample D of such sites. Sites of length n can be represented by nucleotide sequences of length n, where each position i is occupied by attribute $x_i \in \{A,T,G,C\}$, i = 1, …, n. Class description will involve sets of values of some attributes $x_{s_1}, ..., x_{s_m}$, representing the set $\{Y_{s_1}, ..., Y_{s_m}\}$, $Y_{s_t} \subset \{A,T,G,C\}$, $Y_{s_t} \neq \varnothing$, t = 1,...,m.

For aligned nucleotide sequences, we introduce predicates $P^{\varepsilon_t}_{i_t j_t}$, where index i_t denotes the position number, j_t denotes one of the nucleotides {A,T,G,C}, and ε = 0/1 means that the predicate has(/has not) negation. For example, the predicate $P^1_{i_k A}$ means that the position i_k is occupied by the nucleotide A. Let $\mu(C) = \mu(P^{\varepsilon_0}_{i_0 j_0} / P^{\varepsilon_1}_{i_1 j_1} \& ... \& P^{\varepsilon_k}_{i_k j_k})$ be the conditional probability of the rule $C = (P^{\varepsilon_1}_{i_1 j_1} \& ... \& P^{\varepsilon_k}_{i_k j_k} \Rightarrow P^{\varepsilon_0}_{i_0 j_0})$. For the sample of sequences we discover the set of probabilistic laws LP_α, with confidence level α on data D. By the estimation of the law C we mean the value $\mu(C) = -\ln(1-\mu(C))$.

Let us introduce a criterion for mutual conformity of probabilistic laws LP_α on data D. We designate that the regularity $(P^{\varepsilon_1}_{i_1 j_1} \& ... \& P^{\varepsilon_k}_{i_k j_k} \Rightarrow P^{\varepsilon_0}_{i_0 j_0})$ *is applied* to the set $\{Y_{s_1}, ..., Y_{s_m}\}$, if $\{i_1,...,i_k\} \subset \{s_1, ..., s_m\}$ and also $x_{i_t j_t} \in Y_{i_t}$ if $\varepsilon_t = 1$ and $x_{i_t j_t} \notin Y_{i_t}$ if $\varepsilon_t = 0$, t = 1,...,k.

If the regularity is applied to the set $\{Y_{s_1}, ..., Y_{s_m}\}$ and the conclusion of the rule $P^{\varepsilon_0}_{i_0 j_0}$ is fulfilled for that set, e.d. $\{i_0\} \subset \{s_1, ..., s_m\}$ and $x_{i_0 j_0} \in Y_{i_0}$ if $\varepsilon_0 = 1$ and $x_{i_0 j_0} \notin Y_{i_0}$ if $\varepsilon_0 = 0$, then we say that the regularity is *satisfied* for that set, but if the conclusion is not fulfilled, then we say that the regularity is *falsified* for that set. By the criterion of regularity interconnection on the set $\{Y_{s_1}, ..., Y_{s_m}\}$ we designate the value:

$$\Gamma(\{Y_{s_1}, ..., Y_{s_m}\}) = \sum_{C \in S} \mu(C) - \sum_{C \in F} \mu(C)$$

where S is the set of probabilistic laws LP_α for the set $\{Y_{s_1}, ..., Y_{s_m}\}$ satisfied and F is the set of falsified ones.

Definition 2. The set $\{Y_{s_1}, ..., Y_{s_m}\}$, for which the criterion Γ reaches the local maximum relative to the modifications of the sets $Y_{s_1}, ..., Y_{s_m}$ on any one element is called the regular model M = $\langle \{Y_{s_1}, ..., Y_{s_m}\}, Z \rangle$ of the class. By the set of regularities Z, describing the class, we designate the set S∪F of all probabilistic laws LP_α, that are applied to the class. By the description of the class we mean the set $\{Y_{s_1}, ..., Y_{s_m}\}$.

The set $\{Y_{s_1}, ..., Y_{s_m}\}$ can be presented as a matrix. For example the sequence [A][A][C][A][G][-C][T][A][C][A][G][G][T][A][A][G][G][G][C][T] can be presented as matrix $M(Y_{s_1}, ..., Y_{s_m})$ (table 1).

Table 1. matrix $M(Y_{s_1}, ..., Y_{s_m})$

A	1	1	0	1	0	0	0	1	0	1	0	0	0	1	1	0	0	0	0	0	1
T	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0
G	0	0	0	0	1	0	0	0	0	0	1	1	0	0	0	1	1	1	1	0	0
C	0	0	1	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0

In addition to the matrix $M(Y_{s_1}, ..., Y_{s_m})$ we define the regularity matrix $R(Y_{s_1}, ..., Y_{s_m})$ as the matrix of predictions of cells of the matrix $M(Y_{s_1}, ..., Y_{s_m})$ by regularities. The sum of values of the regularity matrix $R(Y_{s_1}, ..., Y_{s_m})$ is equal to the criterion $\Gamma(\{Y_{s_1}, ..., Y_{s_m}\})$. Also we use the involvement matrix $I(Y_{s_1}, ..., Y_{s_m})$ to show the involvement of all predicates of the regularities in there interconnection, which have estimation μ(C) for each predicate of the regularity.

Recognition. Given the control set B of sequences and the regular model M = $\langle \{Y_{s_1}, ..., Y_{s_m}\}, Z \rangle$ of the class, we can recognize the positive and control samples by calculating the score $\Gamma(\{Y_{s_1}, ..., Y_{s_m}\})$ of every training and control sequence. When we define some threshold of the score, we can calculate the true/false positive rates for the training and control sets.

13. Implementation and Results

For the TFBSs analysis and recognition we have chosen the samples of sites of steroidogenic factor-1 (SF1), early growth response factor 1 (EGR1), sterol regulatory element binding protein (SREBP). The train data sets (sequences of TFBSs with flanks) were extracted from the TRRD database [11]. The more detailed description of results is presented in the following subsections.

14.1 EGR1 binding sites analysis.

The data set contained 22 sequences of EGR1 binding sites (BSs). Because of the limited data set the analysis and recognition accuracy of Natclass system in comparison with the optimized positional weight matrix (PWM) was performed according to the standard jackknife procedure [6]. First of all, we tried the PWM on different sequences lengths as it was described in [14] to reach the highest PWM recognition accuracy. When the optimal sequence length for PWM were found to be 10, we prepared positive training set containing sequences of EGR1 BSs of the same length. During each jackknife iteration methods were trained on the data set of 21 sequences leaving exactly the one for the control. The trained methods were applied to the rest sequence (calculation of the recognition score), estimating the false positive (FP) error; the control sample was randomly generated

with the nucleotide frequencies as in the positive samples and contained 100 000 sequences. Totally we performed 22 jackknife iterations that correspond to the number of sequences in the data set. We arranged the control sites according to the false positive rates. Figure 5 depicts the correlations between the false positive/negative rates. Natclass system outperforms PWM at any error cutoff.

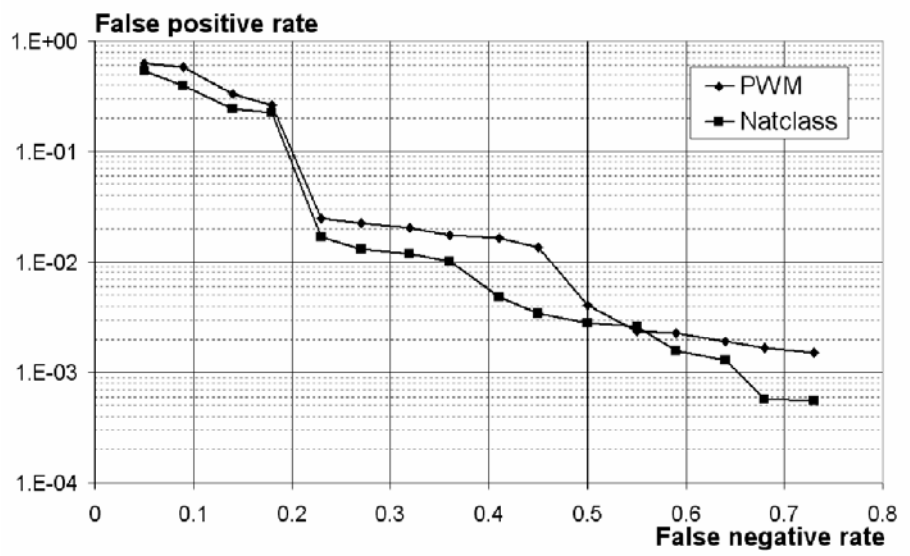


Fig. 5. Recognition accuracy of Natclass system and PWM for EGR1 binding sites, estimated according to the jackknife procedure. Correlations between the false positive/negative rates for two programs.

PWM and consensus-based methods involve an explicit assumption concerning the independent contribution of each nucleotide position to the binding affinity, producing the cumulative effect to the binding strength. A number of works [2, 15, 1, 33] indicate that nucleotides of TFBSs cannot be treated independently. This assumption is invalid and contradicts the processes underlying the biological model. Unlike PWM and consensus methods, the Natclass system reveals the mutual interdependences between the nucleotides or their negations in the specific positions. The clustering and detailed analysis of the regularities found allows to reveal the essential positions of DNA-protein interactions.

After the recognition accuracy step we trained Natclass system using the whole data set (22 sequences) and establishing the same set of parameters as during the jackknife iterations. The system revealed 2354 regularities and discovered exactly the one class [G][C][G][G][G][G][G][C][G][G] covering the positive data. As the example of regularities revealed let's consider the following rule:

(1=g & 7=g & 8=c) $\Rightarrow$ {[3] = (-c: 0.96) & [3] = (-a: 0.84) & [3] = (g: 0.56) & [3] = (t: 0.28)}

Here the rule precondition contains the conjunction of nucleotides or their negations in the specific positions. The rule post condition predicts the specific nucleotide positions suggesting the nucleotides associated with the probabilities values. Here the rule predicts the third nucleotide position and it states, that if the sequence under analysis possesses "g" in the first position and "g" in seventh and "c" in eights then in the third position there is "not c" and "not a", but "g" or "t" with the corresponding probabilities.

From the 2354 regularities revealed only 508 describe the unique class. It is obvious, that each rule predicts four possible nucleotides or their negations associated with the special position with probabilities, that is why, totally, we have 508*4 = 2032 regularities for the discovered class of objects. Among them there are 78 regularities, predicting "c" in eighth position, 69 regularities predicting -"c" in second position, 42 - "g" in first position and 28 - "g" in sixth position (table). What is more, among the 78 regularities predicting "c" in eighth position 51 (65,4%) regularities possess

“g” in fifth position in the rule precondition. Table 2 establishes the most frequently observed interdependencies between nucleotides and their negations in specific positions.

Table 2.

Regularity type	Number of regularities
totally	2032
… ⇒ [8] = (c: …)	78
… ⇒ [2] = (c: …)	69
… ⇒ [1] = (g: …)	42
… ⇒ [6] = (g: …)	28
…& 5=g &… ⇒ [8] = (c: …)	51 (65,4%)
…& 4=g &… ⇒ [2] = (c: …)	39 (56,5%)
…& 5=g & 7=g … ⇒ [8] = (c: …)	17 (21,8%)
… & 5=-t &8=c … ⇒ [6] = (g: …)	13 (46,4%)
… & 3=-a &6=g … ⇒ [1] = (g: …)	10 (23,8%)

The sequence [G][C][G][G][G][G][G][C][G][G] of the class maximizes the criterion Γ of regularity interconnection, it corresponds to the consensus sequence of EGR1 and equally with the most frequently observed regularities agrees with the biological data. The Egr-1, also named NGFI-A, krox-24, zif268, Cef5, and Tis8, is belongs to a family of C2H2 zinc finger proteins which recognizing a GC-rich sequence, 5'-GCG(G/T)GGGCGG-3' [3, 4]. The majority of selected sites possess an invariant pattern of guanines on the primary DNA strand at position 1, 3, 6, 7, and 9 [22]. This pattern corresponds to the sequence of guanines found to be contacted by arginine side chains in the NGF-A zinc fingers-DNA X-ray cocrystal structure [17]**.** At position 5 on the primary strand of the consensus site, G was selected with a frequency of about 70% and A was selected with a frequency of about 30%. At positions 2 and 8 of the site, G was never selected by any of the proteins. At position 2 cytosine (or G on the opposite strand) was nearly exclusively selected for by all the proteins. At position 8 there was less selective pressure [22].

14.2. SF1, SRE binding sites analysis.

Totally, the data sets contained 54 sequences of SF1 BSs and 38 of SREBP. We performed the accuracy comparison of the Natclass system and the PWM according to the bootstrap procedure [6] When the optimal sequence lengths for PWM were found to be 13 nucleotides (SF1 BSs) and 18 (SREBP), we prepared positive training sets containing sequences of BSs of the same length. The negative training set consisted of randomly generated sequences with the same frequencies as in the positive set.

The positive training sets were randomly sampled 15 times into the new subsets, each containing 90% of the whole data sets. The PWM and Natclass methods trained on the basis of these subsets were applied to the rest of sequences (control subsets, 10% of the whole data set). For each of the control TFBSs we estimated the false positive (FP) rate relying on the sets of randomly generated sequences of sufficient size (each set of 1 000 000 sequences). Further we ranged the joint set of the control TFBSs according to the corresponding FP rates. Table 3 presents the FP rates with the false negative (FN) rate being equaled to 50%.

Table 3. The data used in the accuracy comparison of the Natclass system and the PWM. False positive rates at the stringent threshold are defined by the false negative rate equal to 50%.

TFBS	# positive sequences	# control negative sequences for each bootstrap iteration	# of regularities belonging to the class	FN rate	FP rate (Natclass)	FP rate (PWM
SF1	54	1 000 000	1670	27	2e-005	6.87e-005
SREBP	38	100 000	789	19	0/110000 < 1E-005	8.32e-004

The score of the sequence was equal to the negative sum of the significance levels of the regularities the sequence satisfied.

The class [T/C][C][A][A][G][G][T/C][C][A][G] was discovered for the SF1 site, where [T/C] means that on the first place there can be one of two nucleotides T or C.

Table 4 present the regularity matrix for the class [T/C][C][A][A][G][G][T/C][C][A][G] and table 5 present the recognition matrix for the class [G][C][G][G][G][G][G][C][G][G].

Table 4. The regularity matrix R([TC][C][A][A][G][G][TC][C][A][G])

A	0.00	0.00	0.00	0.00	17.92	1512.67	0.00	0.00	0.00	0.00
T	0.00	484.95	643.73	481.73	421.14	872.68	0.00	0.00	0.00	0.00
G	154.92	0.00	2.31	61.84	0.00	0.00	0.00	0.00	0.00	0.00
C	0.00	0.00	4.13	9.89	103.06	634.36	0.00	0.00	0.00	0.00

Table 5. The recognition matrix R([G][C][G][G][G][G][G][C][G][G]).

A	-71.05	-131.96	-91.07	-204.49	-24.48	-114.76	-89.80	-89.96	-229.11	-145.05
T	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
G	0.00	447.36	0.00	-47.98	-15.30	-2.69	0.00	515.56	0.00	0.00
C	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

14. Discussion

Further we plan to improve the method and use it in cooperation with the ExpertDiscovery method [31]. We can discover complex signals by the ExpertDiscovery system and use them as ordinary properties in the classification system DNANatClass.

Acknowledgments

The work is partially supported by the Russian Foundation for Basic Research 05-07-90185-v, Scientific Schools grant from the President of the Russian Federation 4413.2006.1, INNOVATION PROJECT IT-CP.5/001 "Development of software for computer modeling and design in postgenomic system biology (system biology in silico)".

References

[1] Y. Barash, G. Elidan, F. Friedman, and T. Kaplan. Modeling dependencies in protein-DNA binding sites. RECOMB 2003, 28–37.

[2] Benos PV, Bulyk ML, Stormo GD: Additivity in protein-DNA interactions: how good an approximation is it? Nucleic Acids Res 2002, 30:4442-4451.

[3] Cao, X. M., R. A. Koski, A. Gashler, M. McKiernan, C. F. Morris, R. Gaffney, R. V. Hay, and V. P. Sukhatme. Identification and characterization of the Egr-1 gene product, a DNA-binding zinc finger protein induced by differentiation and growth signals. Mol. Cell. Biol. 10:1931-1939, 1990.

[4] Christy, B. A., L. F. Lau, and D. Nathans. A gene activated in mouse 3T3 cells by serum growth factors encodes a protein with "zinc finger" sequences. Proc. Natl. Acad. Sci. USA 85:7857-7861, 1988.

[5] Classification and Clustering (1977), Ed. By J. Van Ryzin, Academic Press, New York.

[6] B. Efron and G. Gong, A leisurely look at the bootstrap the jackknife and resampling. American Statistician, 37 (1983), 36-48.

[7] Halpern, J.Y. (1990), `An analysis of first-order logic of probability', Artificial Intelligence 46, pp.311-350.

[8] Hempel, C. G. (1965) Aspects of Scientific Explanation, In: C. G. Hempel, Aspects of Scientific Explanation and other Essays in the Philosophy of Science, The Free Press, New York.

[9] Hempel, C. G.: 1968, 'Maximal Specificity and Lawlikeness in Probabilistic Explanation', Philosophy of Science 35, 116–33.

[10] Kozhara V.L. Classification functions. // Theory of classification and data analysis, Novosibirsk, 1982.

[11] Kolchanov, N.A., Ignatieva, E.V., Ananko, E.A., Podkolodnaya, O.A., Stepanenko, I.L., Merkulova, T.I., Pozdnyakov, M.A., Podkolodny, N.L., Naumochkin, A.N., Romashchenko, A.G. Transcription Regulatory Regions Database, (TRRD): its status in 2002. // Nucleic Acid Res., 2002, V. 30, P. 312-317.

[12] Kovalerchuk, B., Vityaev, E. (2000), Data Mining in finance: Advances in Relational and Hybrid Methods, Kluwer Academic Publishers, 308 p.

[13] Krantz, D.H., Luce, R.D., Suppes, P., Tversky, A. (1971, 1989, 1990), Foundations of measurement, Vol. 1,2,3 - NY, London: Acad. press, (1971) 577 p., (1989) 493 p., (1990) 356 p.

[14] V. Levitsky, E. Ignatieva, G. Vasiliev, N. Limova, T. Busygina, T.Merkulova, N. Kolchanov, The SiteGA tool for recognition and context analysis of transcription factor binding sites: significant dinucleotide features besides the canonical consensus exemplified by SF-1 binding site, In: Bioinformatics of Genome Regulation and Structure II. (Eds. N.Kolchanov, R. Hofestaedt ,L.Milanesi), Springer Science+Business Media, Inc. (2006), pp. 31-41.

[15] Man TK, Stormo GD: Non-independence of Mnt repressoroperator interaction determined by a new quantitative multiple fluorescence relative affinity (QuMFRA) assay. Nucleic Acids Res 2001, 29:2471-2478.

[16] Meien S.V., Shreider C.A. Methodological aspects of classification theory. Philosofical questions. v.12, 1976.

[17] Pavletich, N. P., and C. O. Pabo. Zinc finger-DNA recognition: crystal structure of a Zif268-DNA complex at 2.1 A. Science 252:809-817, 1991.

[18] Pfanzagl J. (1971). Theory of measurement (in cooperation with V.Baumann, H.Huber) 2nd ed. Physica-Verlag.

[19] Rutkovskii L. Elementary logic textbook. – Spt., 1884.

[20] Shreider S.A. Systematic, typology, classification // Theory and methodology of biological classification. Moscow, Science, 1983.

[21] Smirnof E.S. Constructions of forms from the taxonomic view // Zool. Jour., v.17(3), 1938, pp. 387-418.

[22] Swirnoff A.H. and J. Milbrandt. DNA-binding specificity of NGFI-A and related Zinc finger Transcription factors. Mol. Cell. Biol. 15:2275-2287, 1995.
[23] Vityaev, E.E. (1992), `Semantic approach to knowledge base development: Semantic probabilistic inference', Computer Systems 146, pp.19-49. (in Russian).
[24] Vityaev E. The logic of prediction. In: Mathematical Logic in Asia. Proceedings of the 9th Asian Logic Conference (August 16-19, 2005, Novosibirsk, Russia), World Scientific, Singapore, 2006, pp.263-276
[25] Vityaev, E.E. Classification as a determination of groups of objects that satisfy different sets of consistent regularities. Comp. Syst., 99:44-50, 1983. (in Russian).
[26] Vityaev, E., Kovalerchuk, B., Empirical Theories Discovery based on the Measurement Theory. Mind and Machine, v.14, #4, 551-573, 2004.
[27] Vityaev E.E., Kostin V.V. et al. Natural classification of nucleotide sequences. // Proc. of the Third International Conference On Bioinformatics of Genome Regulation and Structure (BGRS'2002, Novosibirsk, Russia, July 14-20, 2002), v3, ICG, Novosibirsk, 2002, pp. 197-199
[28] Vityaev, E.E. and Kostin, V.S. Natural Classification as the law of Nature, in: Intelligent systems and Methodology, Proc. Symp. "Intelligent supporting of activity in complex subject domains", Novosibirsk, 7-9 Apr., 1992, part 4, Novosibirsk, 1992, p.107-115 (in Russian).
[29] Vityaev E.E., Lapardin K.A. et al. Natural classification and systematic as the laws of nature // Analysis of structural regularities (Comp. syst. #174), Novosibirsk, 2006, pp. 80-92 (in Russian).
[30] Vityaev EE, Logvinenko AD (1998): Laws discovery on empirical systems. Axiom systems of measurement theory testing. Sociology: methodology, methods, mathematical models (Scientific journal of the Russian Academy of Science) 10:97-121. (in Russian)
[31] E.E. Vityaev, T.I. Shipilov et al. Software for analysis of gene regulatory sequences by knowledge discovery methods. In: Bioinformatics of Genome Regulation and Structure II. (Eds. N.Kolchanov and R. Hofestaedt) Springer Science+Business Media, Inc. 2006, pp. 491-498
[32] Zabrodin, V.Yu. Criteria of naturalness of classifications. NTI, ser. 2., 1981.
[33] Udalova, I.A., Mott, R., Field, D., and Kwiatkowski, D. Quantitative prediction of NF-kB DNA–protein interactions. Proc. Natl. Acad. Sci. USA 99, 2002, 8167–8172.