Natural Classification of Nucleotide Sequences

E.E. Vityaev*[1], V.S. Kostin[2], N.L. Podkolodny[3], and N.A. Kolchanov[4]

[1]Institute of Mathematics SB RAS, Novosibirsk, Russia
[2]Institute of Economics and Industrial Engineering SB RAS, Novosibirsk, Russia
[3]Institute of Computer Mathematics and Mathematical Geophysics SB RAS, Novosibirsk, Russia
[4]Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia
E-mail: vityaev@bionet.nsc.ru
*Corresponding author

Keywords: Bioinformatics, Knowledge Discovery and Data Mining, Machine Learning, eukaryotic promoter recognition, transcription factor binding sites

Resume

Motivation: A principally new approach to constructing classifications of nucleotide sequences on the basis of the "natural" classification concept is proposed in the paper. The "natural" classification is based on the following principle: objects of one class should obey one group of rules, and objects of different classes should obey different groups of rules. Based on this principle, a method for constructing a classification, an algorithm, and a GeneNatClass software system have been developed.

Results: A method for constructing the classification, algorithm, and the GeneNatClass software system have been developed, which allows identification of "natural" classes of subsequences, that is, motives.

Availability: Scientific Discovery Website: http://www.math.nsc.ru/LBRT/logic/vityaev

Introduction

Numerous principles of constructing classifications are currently known. The classifications are based on the hypothesis of compactness and various measures of closeness in a certain space, on resemblance of standards, on supertargets, on various criteria of classification quality and quality functionals, on separation of distribution mixtures, etc. (Classification and Clustering, 1977).

Nevertheless, these approaches rarely yield stable and law-like results. Therefore, they should be used carefully, with clear understanding of restrictions in conclusions that can be drawn on the basis of these classifications.

In contrast to the above-listed classifications, the objective of a "natural" classification is the insight into the object domain and identification of those latent relations, notions, and regular features that are important for constructing the theory of the object domain and possessing a predictive force. In this meaning, "a natural classification is only the one that reflects the law of nature" and ensures the following (Zabrodin, 1981; Vityaev and Kostin, 1992):

- Prediction of the maximum of object properties, based on the object place in the classification;

- Maximum of general statements about each class;
- Retaining of the structure of classes with variation of classification features; and
- Objectivity, reliability, and predictive force.

A constructional criterion of a "natural" classification was proposed in (Vityaev, 1983): "Objects should be divided into classes with accordance with the rules satisfied by the objects. More exactly, objects of one class should obey one group of rules, and objects of different classes should obey different groups of rules. Objects of one class should also possess some integrity, which is understood as mutual agreement of rules of each group in terms of mutual prediction of object properties. In addition, groups of rules may have common features that establish relations between features of objects from different classes".

Sets of rules of each class reveal a regular structure of objects of the class. A regular structure exactly reflects the idealization process (Vityaev and Kostin, 1992); therefore, the structure itself is called an ideal object of a class, and the classification procedure is called idealization.

The method of "natural" classification (Vityaev and Kostin, 1992) may be divided into the following stages:

(1) Mapping of raw data and formation of a learning sample.

   (a) Formalization of various types of relations important for the description of chosen objects from the viewpoint of an expert.

   (b) Constructing a feature space of objects. Constructing higher-order variables from other primitive variables.

(2) Data cleaning and preprocessing. Constructing data samples.

(3) Finding rules.

   (a) Specification of a system of nested Rule Types.

   (b) Generation of various statistical hypotheses on the basis of Rule Types and their verification on the learning sample; search for rules relevant for recognition of various types of objects.

(4) Search for all regular structures (ideal objects) of classes.

Methods and algorithms

Nucleotide sequences are used as initial data. To construct a learning sample, one has to define a specification of objects and their properties. A set of features measured on these objects determines various relations between nucleotides, their positions, sections of sequences or full nucleotide sequences, etc.

In the general case, the set of features whose values are determined for particular objects may change. Formally, these data can be represented in the form of an XML description or a set of relational tables.

Algorithm for finding rules. To find rules, we use the relational approach to Data Mining methods (Kovalerchuk and Vityaev, 2000) verified by the Discovery system, which allows one to find and test almost all types of hypotheses in the first-order language. A system of nested Rule Types implements a strategy of more and more detailed and exact analysis of the object domain theory. This approach allows one (1) to process data of all types, measured in arbitrary scales (partial-order, grids, titles, orders, log-linear, trees, networks, graphs, etc., and also mixtures of these quantities) and (2) to find law-like rules under conditions of noise and small learning samples.

The simplest rules on symbolic sequences have the form
IF $(Pos_{i1}(\boldsymbol{\alpha}) = \{A|T|G|C\})^{\varepsilon 1}\&...\&(Pos_{ik}(\boldsymbol{\alpha}) = \{A|T|G|C\})^{\varepsilon k}$
THEN $(Pos_{i0}(\boldsymbol{\alpha}) = \{A|T|G|C\})^{\varepsilon 0}$, (1)
where $(Pos_{ij}(\boldsymbol{\alpha}) = \{A|T|G|C\})^{\varepsilon j}$ that one of the values of {A|T|G|C} is located (for $\varepsilon_j = 1$) or is not located (at $\varepsilon_j = 0$) in the position ij of the object **α**. We denote the hypothesis of rule (1) after the condition IF as Premis(**£**).

Statement (1) is a rule if the following conditions are satisfied:
$\text{Prob}(\text{Premis}(\boldsymbol{£})) > 0;$ (2)
$\text{Prob}((Pos_{i0}(\boldsymbol{\alpha})=\{A|T|G|C\})^{\varepsilon 0})/\text{Premis}(\boldsymbol{£}))>\text{Prob}((Pos_{i0}(\boldsymbol{\alpha})=\{A|T|G|C\})^{\varepsilon 0})/\text{SubPremis}(\boldsymbol{£})$.
Here, Prob is the probability of the statement and SubPremis(**£**) means that one or several relations are lacking in the hypothesis. If the second condition of (2) is not satisfied, then, if some relation from the hypothesis Premis(**£**) is deleted, the conditional probability of some subrule **IF** SubPremis(**£**), **THEN** $(Pos_{i0}(\boldsymbol{\alpha}) = \{A|T|G|C\})^{\varepsilon 0})$ is not lower than the conditional probability of the rule itself; hence, this relation can be deleted.

The relational approach to Data Mining methods means the use of a strategy of a more and more exact and detailed analysis of data by means of arbitrarily complicated refinement of the hypothesis being tested. For instance, hypotheses for symbolic sequences may include additional features of belonging of nucleotides to a certain region, specific or admissible distances between nucleotides, properties of nucleotides themselves, etc.

All features of objects are tested as target features of the rule. The hypothesis (Premis) plays the role of a filtering query that chooses those objects from the sample for which all features of the hypothesis have the values indicated in the rule. To measure the rule force, we compare the conditional distribution of target values obtained when all hypothesis features are satisfied and the distribution of target values on all objects. The stronger the rule, the greater the deviation of the conditional distribution of the target values from the initial distribution. One of the simplest methods of measuring this deviation is the statistics χ^2. We use it in the form of the so-called normalized $Z\chi^2$-deviation:

$$Z_{\chi^2_{i_0}} = \sqrt{2\chi^2} - \sqrt{5}$$

$$\chi^2_{i_0} = \sum_{k=1}^{2}\sum_{j_0=1}^{4}\frac{(N_{kj_0} - E_{kj_0})^2}{E_{kj_0}}$$

N is the total number of rows in the table;
N_k is the number of rows in the table where the hypothesis (k=1) is satisfied and the hypothesis (k=2) is not satisfied;
N_{j0} is the number of rows in the table where the values of the target feature $j_0 \in \{ATGC\}$ are observed;
$E_{kj0} = N_k N_{j0}/N$ is the expected value of N_{kj0} under the condition that k and j_0 are independent;
N_{kj0} is the number of rows in the table where the values k and j_0 are observed simultaneously.

The probability inequalities (2) are tested by this $Z\chi^2$-deviation. The rules are sought by gradual increasing of the rule hypothesis by one feature at each step. The extended hypothesis should yield a stronger rule than the short one.

Construction of ideal objects. The next stage of analysis of nucleotide sequences is the construction of ideal images of real-world sequences. If the objects of a class possess some integrity, it is manifested in the structure of regular relations unifying the parts/features of an object into a single unit. It is the structure of regular relations that determines the unification of the parts/features of an object into a single unit.

The idealization procedure reduces to the following. Using all rules, we supplement the description of a real object by additional values of features that are predicted with a high probability by the remaining set of features and already included features and delete those features that fall outside the overall ensemble. This procedure is continued until all the necessary values are included and all random values are sorted out. This procedure is regulated by the criterion of mutual agreement of rules, which should be rigorously increasing at each step.

In terms of software, the idealization process is implemented as follows: for a certain real sequence, a matrix M is generated, which contains the number of rows equal to the number of nucleotides in the sequence and four columns (one for A, T, G, and C). The entire set of rules R is considered. Each rule applied to the sequence adds its four predictions in the form of $Z\chi^2$-deviations (for A, T, G, and C) into four cells of the row corresponding to the target feature of the rule. If the sequence contains one or more of these four values, the total criterion of self-consistency acquires a contribution equal to the sum of the predictions ($Z\chi^2$-deviations) of these values (and these values only). If the sequence contains the value with the negative sign, the corresponding contribution is also taken with the minus sign. The occurrence with the negative sign means that the sequence should not contain the corresponding nucleotide. Zero indicates that this nucleotide does not enter this sequence, but the absence of this nucleotide is not required.

We determine the sequences that are **ideal objects of classes.** For this purpose, we introduce a criterion of mutual agreement of the rules on prediction on these objects:

$$\Gamma(\mathrm{M}) = \sum_{R} \sum_{j_0=1}^{4} Z_{\chi^2_{i_0}} \delta(i_0, j_0)$$

where R is the set of rules and $\delta(i_0,j_0) = 1$ (-1) if the state of the nucleotide $j_0 = \{A|T|G|C\}$ in the current position i_0 of the sequence coincides (does not coincide in the case of –1) with the values in the sequence itself.
DEFINITION (Vityaev, 1983). **An ideal object** of a class is the set of nucleotides <{A|T|G|C}{A|T|G|C}…{A|T|G|C}> for which the criterion Γ(M) has a local maximum (the value of this criterion rigorously decreases in the case of deletion or addition of an arbitrary value in this set). The record {A|T|G|C} means that one or several nucleotides from those indicated in the braces can be treated as an ideal object.

Conclusions

A software system GeneNatClass has been developed. The system implements the above approach to constructing “natural” classifications. The system developed has been tested on problems of classification of splicing sites and transcription factor binding sites; the efficiency of the system has been demonstrated.

Acknowledgments

The authors are grateful to I.B. Rogozin for providing splicing site data. This work was partly supported by the Russian Foundation for Basic Research (Grant Nos. 01-07-90376, 00-04-49229, and 02-07-90355) and Siberian Branch of the Russian Academy of Sciences (Integration Project No. 65).

References

Vityaev, E.E. and Kostin, V.S. (1992). 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).

Vityaev, E.E. (1983). Classification as a determination of groups of objects that satisfy different sets of consistent regularities. *Computational Systems,* **99**:44-50 (in Russian).

Zabrodin, V.Yu. (1981) Criteria of naturalness of classifications. NTI, ser. 2.

Classification and Clustering, Ed. By J. Van Ryzin, Academic Press, New York, 1977.

Kovalerchuk, B. and Vityaev E. (2000*). Data Mining in Finance: Advances in Relational and Hybrid methods.* Kluwer Academic Publishers, p.308.

Vityaev, E.E., Orlov, Yu.L., Vishnevsky, O.V., Kovalerchuk, B.K., Belenok, A.S., Podkolodny, A.S., and Kolchanov, N.A. (2001). Computer System "Gene Discovery" for Functional Annotation of DNA Sequences. *Proc. Workshop Machine Learning as Experimental Philosophy of Science, ECML/PKDD 2001* (Eds., K.B. Korb, H. Bensusan), Freiburg, Germany, 3-7 September, 2001, Freiburg University, 1-11.