pp. 1–12 (2020)

Application of mutual information estimation for predicting the structural stability of pentapeptides

A. I. Mikhalskii[*], I. V. Petrov[†], V. V. Tsurko[*], A. A. Anashkina[‡], and A. N. Nekrasov[§]

Abstract — A novel non-parametric method for mutual information estimation is presented. The method is suited for informative feature selection in classification and regression problems. Performance of the method is demonstrated on problem of stable short peptide classification.

Keywords: Mutual information, data mining, feature selection, pentapeptide stability prediction

MSC 2010: 92-04, 62-07, 62G07

Large amounts of experimental data used in modern scientific research and applied problems are heterogeneous. The heterogeneity is related to differences in data sources and methods of their recording, noise and information redundancy due to the lack of a model indicating factors uniquely related to the studied target indicators. Such problems are also characterized by a high dimension of the feature space. For example, when studying a cell or tissue sample by sequencing a new generation, it is also possible to obtain information about the expression of almost all protein-coding genes and also short and long non-coding RNA. The typical size of the data set to be studied is negligible in comparison with the number of features. The number of features is measured in thousands, and the number of samples is hundreds at best [4].

The presence of a large number of features not only reduces the effectiveness of solving the problem in terms of required computing resources and impairs data visualization and interpretation, but also reduces the generation ability of the used algorithm. In this case, the algorithm for solving the problem adapts to specific data used in its tuning, but produces big errors when checking on independent material.

The following methods are used to improve the efficiency and reliability of data analysis: methods for selecting features, highlighting patterns inherent to data, and

[*] V. A. Trapeznikov Institute of Control Sciences, RAS, Moscow 117997, Russia.
E-mail: ipuran@yandex.ru

[†] LLC T2 Mobile, Big Data Office, Moscow 108881, Russia

[‡] Shemyakin–Ovchinnikov Institute of Bioorganic Chemistry, RAS, Moscow 117997, Russia

[§] Engelgardt Institute of Molecular Biology, RAS, Moscow 119991, Russia

The work was supported by the RFBR (project No. 20–04–01085).

clustering data into homogeneous groups. As the result, the dimension of the problem decreases, which is equivalent to an increase in the number of observations in experimental data, the effect of random perturbations on the result is reduced, and statistical reliability of the result increases.

Dimension reduction algorithms include RELIEF [6], FOCUS [1], methods transforming the feature space, for example, the principal component analysis [5], independent component method [2], the method for contrasting distributions [16]. The principle of selecting informative features according to the value of the correlation coefficient with the target variable is widely used as well.

The value of the correlation coefficient may be close to zero in the case of using it as a criterion for filtering out features under the presence of a significantly nonlinear relationship between values of features and the target variable. This may lead to incorrect judgments about the significance of the feature and reduces the quality of the algorithm. In this paper, we propose to use the mutual information estimation over experimental data for select features, which, in contrast to the correlation coefficient, allows us to highlight nonlinear dependencies. The results of application of this approach in a real problem of selecting features are presented for the case of predicting the stability in a pentapeptide conformation.

One of key problems in modern biology is the study of physical, chemical, and functional properties of protein molecules and their design. Prediction what spatial structure the protein takes in the process of its folding is important, in particular, for development of drugs affecting the functioning of biological systems. Previously, the entropy of protein sequences was studied in [9]. It was shown that the pentapeptide is the optimal size of a structural block. Using a description of proteins with the use of structural blocks in terms of five amino acid residues, an algorithm revealing the internal hierarchy in the protein sequence was developed [10, 12].

In [11], conformation-stable pentapeptides were found by methods of molecular dynamics and the assumption was made that such peptides play an important role in the formation of native spatial structure by proteins. Such structurally stable peptides can initiate the formation of elements of the secondary structure of proteins and thus ensure correct prefolding conformation of polypeptide chain.

In total, $20^5 = 3{,}200{,}000$ different sequences of pentapeptides are theoretically possible. A molecular dynamic experiment requires very significant time and calculation resources, therefore, the preliminary search for candidates for further molecular dynamics study is relevant and in demand. In this paper we consider the problem of predicting conformation-stable pentapeptides out of 3,200,000 possible.

1. Materials and methods

1.1. Sequence of pentapeptides

In the present paper we use the results of molecular dynamic modelling for 49745 pentapeptide sequences.

The algorithm for choosing sequences of pentapeptides was based on the assumption that two or three interactions in the pentapeptide are sufficient for form-

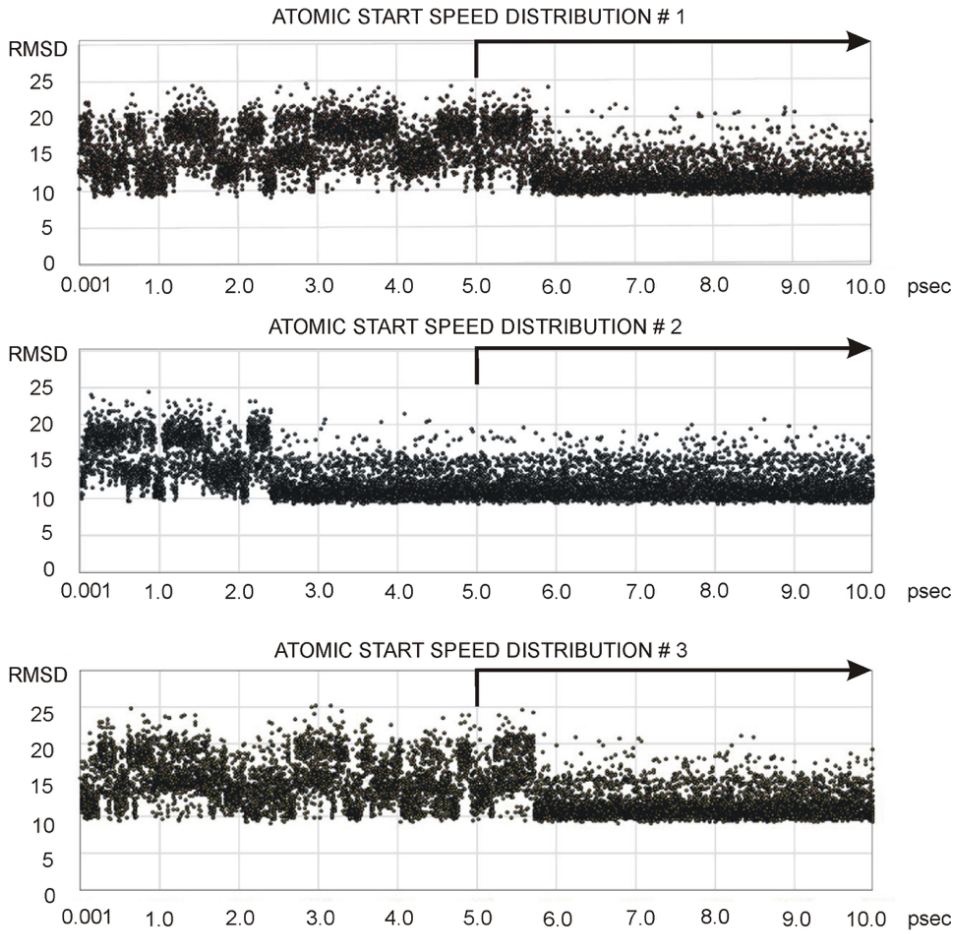

Figure 1. Graphs of RMSD of torsion angles (f, y) of the polypeptide backbone relative to the center of cluster 5 for the IAFAE peptide [11] obtained in the process of molecular-dynamical (MD) modelling. MD modelling was carried out for different start velocities of atoms and the same temperature of 300K. The first 5000 points (5 picoseconds, i.e., 50% of MD-simulation time) correspond to the peptide relaxation time from the initial conformal state. The arrow on graphs indicates the part of MD-trajectory whose conformation was used in the clusterization of spatial structures of pentapeptide.

ation of structural stability. Therefore, three substitutions were introduced into the alanine pentapeptide, one at the central position and two more for other positions. Using this algorithm, 44860 sequences of pentapeptides were made. These peptides were studied by the method of molecular dynamics. All technical details can be found in [11].

In modern implementations of the molecular dynamics method, particle trajectories correspond to equations of motion that differ from Newtonian ones by small terms having a stochastic character. As a result, a bundle of trajectories is formed on the Newtonian ‘leg’ instead of a unique solution to the Cauchy problem. Vari-

ous modified versions of the method are widely used in the study of both macromolecular structures and monatomic systems, which distorts further the solution to Newton's equations.

In this paper, the 'stability' of a pentapeptide is understood as the presence of a predominant conformations in a molecular dynamic experiment. We are interested in the ensemble of equilibrium conformations of the peptide, but not the technique they were obtained. The applicability of the molecular dynamics method for study of structure and behavior of bipolymers were discussed in detail in a number of books and papers [3, 4, 5, 7, 8, 9, 11, 10, 12]. In practice the conformational stability of the biomolecular structure can be estimated from an experiment by equilibrium molecular dynamics excluding from consideration the conformations during the balancing process and comparing the residence times of molecule in different conformations. Figure 1 shows three different trajectories of the same IAFAE peptide calculated by molecular modelling at different starting velocities of atoms and the same temperature of MD simulation equal to 300K. Only the conformations observed after the first 5 picoseconds of simulation were subject to clusterization. This area is marked with an arrow on the presented trajectories.

We consider a pentapeptide to be conformation-stable if the largest cluster of conformations includes more than 80% of all conformations. Supplementing this set of 44860 pentapeptides with previously studied pentapeptides from real proteins, we obtain a sample of 49745 pentapeptides, and 1705 of them are structurally stable (3.43% of their total number).

1.2. Mutual information and its estimation

Mutual information serves as a measure of relationship between random variables and has a wide range of applications. In information theory, it is taken as a measure of the information contained in a random variable Y relative to a random variable X, it measures the amount of information contained in the received message about the transmitted message. The mutual information is a statistics used in testing statistical hypotheses. The mutual information is used in the analysis of empirical data aimed to find dependencies between variables and classification factors, to construct regression dependencies. Advantages of mutual information in comparison to the correlation coefficient are in its universality not limited by linear dependencies.

An example of the usage of mutual information estimate was presented in [14] for revealing the interconnections between biological processes and stimulating actions, the selection of features with the use of mutual information in the diagnosis of acute pulmonary embolism was considered in [15], the application of mutual information when choosing the model of optimal forecast of stochastic systems was given in [3].

Formally, the mutual information between random variables X and Y having the joint distribution $P(x,y)$ is defined by the relation

$$I(X,Y) = \int \ln \frac{\mathrm{d}P(x,y)}{\mathrm{d}P(x)\mathrm{d}P(y)} \mathrm{d}P(x,y).$$

If the distribution of random variables has the density $p(x,y)$, then the mutual information is representable in the form

$$I(X,Y) = \iint p(x,y) \ln \frac{p(x,y)}{p(x)p(y)} \mathrm{d}x\mathrm{d}y.$$

The definition implies that the mutual information equals zero for independent random variables. There are various ways to estimate the value of mutual information for a set of experimental data $((x_1,y_1),\ldots,(x_n,y_n))$. A simple method for evaluating mutual information consists in replacing the integration over the distribution $P(x,y)$ by averaging over sample values

$$\hat{I}(X,Y) = \frac{1}{km} \sum_{i=1}^{k} \sum_{j=1}^{m} p(x_i,y_j) \ln \frac{p(x_i,y_j)}{p(x_i)p(y_j)}$$

where $p(x_i)$, $p(y_j)$, and $p(x_i,y_j)$ are estimates of densities from experimental data.

Such estimates can be taken from the histogram estimation, i.e., as the portion of sample elements falling into the cell with the number i, j of rectangular grid consisting of k cells in the variable x and m cells in the variable y. Another technique consists in the use of parametric, kernel estimators for the densities $p(x)$, $p(y)$, and $p(x,y)$ [7, 8].

The drawback of both approaches is the requirement for a large set of experimental data for obtaining accurate density estimates.

1.3. Evaluation of mutual information by solving an integral equation

Mutual information can be written as the mathematical expectation of the difference of entropies

$$I(x,y) = \mathbb{E}_x \left[H(y) - H(y|x) \right]$$

where $H(y) = -\sum_{t=\{0,1\}} p(y=t) \log_2 p(y=t)$ is the entropy of the random variable y, $H(y|x)$ is the conditional entropy.

The expression $I(x,y)$ can be rewritten in the form

$$I(x,y) = \sum_{t=\{0,1\}} \int p(x,y=t) \log_2 \frac{p(x,y=t)}{p(x)p(y=t)} \mathrm{d}x.$$

A method of direct evaluation of the mutual information in the problem of binary classification was proposed in [17], the variable y takes the values 0 or 1 in this method. Let $x_1^y,\ldots,x_{\ell_y}^y$ be a sample from the class y, $y=0,1$. Write down the empirical mutual information, the derivation of equation (1.1) is presented in [17],

$$I_e(x,y) = \frac{1}{\ell_0+\ell_1} \sum_{t=\{0,1\}} p(y=t) \left(\sum_{i=1}^{\ell_0} r_t(x_i^0) \log_2 r_t(x_i^0) + \sum_{i=1}^{\ell_1} r_t(x_i^1) \log_2 r_t(x_i^1) \right) \tag{1.1}$$

where $r_t(x) = p(x,\ t)/p(x)$.

By definition, the ratio of densities is the solution to the integral equation

$$\mathscr{A} r_t(x) = \int \mathbb{I}\{x \geqslant u\} r_t(u)\, \mathrm{d}F(x) = F_t(x). \tag{1.2}$$

The solution to integral equation (1.2) can be obtained by approximation of the right-hand side of the equation and the operator:

$$F_t(x) = \frac{1}{\ell_t} \sum_{i=1}^{\ell_t} \mathbb{I}\{x \geqslant x_i^t\}$$

$$\mathscr{A}_e r_t(x) = \frac{1}{\ell_0 + \ell_1} \left(\sum_{i=1}^{\ell_0} \mathbb{I}\{x \geqslant x_i^0\} r_t(x_i^0) + \sum_{i=1}^{\ell_1} \mathbb{I}\{x \geqslant x_i^1\} r_t(x_i^1) \right), \quad t = 0, 1.$$

Equation (1.2) is an ill-conditioned first kind Fredholm equation. It is solved by regularization, the metric is defined through a special V-matrix [17] which preserves geometrical properties of the sample.

1.4. Evaluation of mutual information with the use of a quadratic functional

An estimation of mutual information without preliminary estimation of distribution densities was proposed in [14]. The ratio of densities $w(x,y) = p(x,y)/(p(x)p(y))$ is sought by minimizing the functional

$$J(\hat{w}) = \frac{1}{n^2} \sum_{i,j=1}^{n} \left(\hat{w}(x_i, y_j) - 1\right)^2$$

in the space of estimators of the form $\hat{w}(x,y) = \sum_{j=1}^{m} a_j \varphi_j(x,y)$, where $\varphi_j(x,y)$ are nonnegative independent functions. m is the parameter of the algorithm. As functions $\varphi_j(x,y)$ it is proposed to take some kernel functions, for example, Gaussians whose centers are taken from a random set of experimental points.

In this paper we propose to construct the estimation of the ratio of densities by minimizing the functional

$$J_0(\hat{w}) = \frac{1}{2} \iint \left(w(x,y) - \hat{w}(x,y)\right)^2 p(x)\, p(y)\, \mathrm{d}x\, \mathrm{d}y.$$

in a reproducing kernel Hilbert space (RKHS). The problem of choosing the location of kernels and their number disappears in this case. This functional can be rewritten as

$$J_0(\hat{w}) = \frac{1}{2} \iint \hat{w}^2(x,y)\, p(x)\, p(y)\, \mathrm{d}x\, \mathrm{d}y - \iint \hat{w}(x,y)\, p(x,y)\, \mathrm{d}x\, \mathrm{d}y + C$$

where the constant C does not depend on the approximation $\hat{w}$ and will be omitted below. The empirical estimator of the functional $J_0(\hat{w})$ over a sample of pairs $((x_1,y_1),\ldots,(x_n,y_n))$ of experimental data has the form

$$J_e(\hat{w}) = \frac{1}{2n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\hat{w}^2(x_i,y_j) - \frac{1}{n}\sum_{i=1}^{n}\hat{w}(x_i,y_i). \tag{1.3}$$

In order to minimize functional (1.3) with respect to $\hat{w}$, we may assume a certain model dependent on a finite number of parameters. The parameters of the model are determined by minimization of (1.3) in a finite-dimensional space.

Another approach consists in non-parametric evaluation of $\hat{w}$. The following regularized functional is minimized in this case in the infinite-dimensional Hilbert space L:

$$J_e(\hat{w},\lambda) = J_e(\hat{w}) + \frac{\lambda}{2}\|\hat{w}\|_L^2 \tag{1.4}$$

where the parameter $\lambda > 0$, and $\|\cdot\|_L$ denotes the norm in the Hilbert space L. The regularizing term added to functional (1.3) provides the uniqueness of the point of minimum and increases the stability of solution to random perturbations in experimental data. If functional (1.4) is minimized in a reproducing kernel Hilbert space, then by the representer theorem [13] the approximation $\hat{w}$ minimizing the functional $J_e(\hat{w},\lambda)$ under fixed λ is representable in the form

$$\hat{w}(z) = \sum_{i=1}^{n}\alpha_i K(z,z_i) \tag{1.5}$$

where $z=(x,y)$, $z_i=(x_i,y_i)$, and the nonnegative definite function $K(z,t)$ is the kernel corresponding to the scalar product in the space L, the coefficients α_i are determined by minimization of functional (1.4). Any nonnegative definite function can be taken as a kernel, for example,

$$K(z,t) = \exp(-\sigma^{-2}\|x-t\|^2).$$

The coefficients α_i in expression (1.5) are calculated by the formula

$$\alpha^* = (H+\lambda K)^{-1}h$$

where the elements of the matrix H are calculated by the formula

$$H_{lm} = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}K(x_i,y_j,x_l,y_l)K(x_i,y_j,x_m,y_m)$$

and the elements of the matrix K are calculated by the formula $K_{ij}=K(x_i,y_i,x_j,y_j)$. The derivation of formulas is presented in Appendix A.

1.5. Sliding control for selecting optimal parameters

To choose the parameter λ and the window width σ in expression (1.5), an estimator for the value of the functional $J_0(\hat{w})$ for given parameters is required. To compute it, we use the procedure of sliding control, which is the following:

- the experimental data sample is divided into K parts $Z_k, k = 1, \ldots, K$, by Z_{-k} we denote the part of the sample without Z_k;
- for each Z_{-k} we calculate $\hat{w}(x,y)$, and on the set Z_k we calculate the value of the functional
$$\hat{J}_k = \frac{1}{2} \sum_{x,y \in Z_k} \frac{\hat{w}_k^2(x,y)}{n_k^2} - \sum_{(x,y) \in Z_k} \frac{\hat{w}(x,y)}{n_k};$$
- for given parameters λ and σ the value of the functional $J_0(\hat{w})$ is estimated by the mean value $\hat{J} = \frac{1}{K} \sum_k \hat{J}_k$;
- searching over the values of λ and σ, we determine the best values of parameters such that the functional $\hat{J}$ being the unbiased estimator of the functional $J_0(\hat{w})$ takes the minimal value.

2. Results

Two classification methods were compared to predict the stability of short proteins: the nearest neighbor method with feature selection based on the value of mutual information (implementation in the R language from the fastknn package with proximity estimation inversely proportional to the distance of a neighbor from the desired point) and the implementation of a random forest in the R language from the ranger package.

With a random choice of pentapeptides from a training sample of 49745 pentapeptides, the probability of detection of a structurally stable one was 3%.

Table 1 presents the result of predicting the stability of pentapeptides on a test sample using the nearest neighbor method. We have chosen the number of neighbors and mutual information resulting in the maximum of F1 on the test sample. The accuracy of detection of stable pentapeptides equal to 0.443 is high taking into account their small presence in the data of molecular dynamic modelling and a small training sample size.

The random forest algorithm suits well for processing large-size data. Each tree in the forest is trained on a small random subset of the entire feature set. The implicit selection of features works by selecting optimal partitions based on an information criterion and due to limited depth of trees. Table 2 presents the results of stability prediction for short proteins by the method of random forest with the dimension parameter $mtry = 20$ of the random subspace showing the highest value of F1 on the test sample.

The comparison of tables shows that the choice of informative features by the value of mutual information allowed us to achieve the quality of classification of

Table 1. Accuracy of the method of k nearest neighbors on the test sample.

Set of features	Accuracy	Completeness	F1	k
All features	0.213	0.200	0.214	2
Informative features	0.443	0.485	0.463	4

Table 2. Accuracy of the method of random forest calculated by the out-of-bag method.

Set of features	Accuracy	Completeness	F1
All features	0.490	0.564	0.525

Table 3. Informative features for the method of nearest neighbors (mutual information) and random forest (permutation importance).

Position in pentapeptide	Method of nearest neighbors	Random forest
Position 1	A, D, E, G, K, P, Q, R	A, D, E, K, P, R
Position 2	D, E, G, K, M	A, D, E, K
Position 3	E, P, Q, R	D, E, K, P, R
Position 4	C, D, E, G, H, P, Q, R	D, E, G
Position 5	C, D, E, G, H, K, P, R, S	D, E, K, P, R

the simple nearest neighbor method close to that of the more complex and resource-intensive random forest algorithm.

To identify the priority pentapeptides for calculations, we classified all 3,200,000 possible pentapeptides except for the training sample. The models were retrained on the full data set.

The method of k nearest neighbors predicted 6.9% of structurally stable pentapeptides from the total number, and the random forest method predicted 4.3%. The second result is closer to the percentage of required desired class in the training sample.

The result is easily explained. The method of k nearest neighbors uses only informative features. Therefore, the transformation translating pentapeptides into the feature space loses its mutual unambiguity. This means that different pentapeptides sometimes get the same feature representation. Due to this fact, the percentage of pentapeptides classified as structurally stable will be higher.

Table 3 shows the informative features selected in the classification methods used in the paper. They are the most important for searching for structurally stable pentapeptides. The results of both methods are consistent. The features selected by the random forest are almost completely included in those selected according to the mutual information.

Each pentapeptide position contains any of 20 canonical amino acid residues. The use of only informative features significantly reduces the dimension of the space in which the classification occurs.

The set predicted by the k nearest neighbor method contains 218,362 pentapeptides. The set predicted by the random forest method contains 137,300 pentapeptides.

3. Discussion

Most short peptides have no stable structural states. However, such stable states are observed in a very small part of peptides (1.7–7.2% according to various estimates). Such stable states are observed. These peptides can play the role of centers in protein folding. These peptides are the subject of our research. The molecular dynamic experiment requires very much time and significant computational resources. Given the huge number (3,200,000) of possible pentapeptide sequences, the relevance of development of mathematical methods for predicting the presence of structural stability is evident. The prediction method proposed here allowed us to select pentapeptides for further checking their structural stability by other methods.

Let us try to explain why at this stage of research we used the method of molecular dynamics instead of analyzing pentapeptides from experimentally defined protein structures. In experimental structures, in addition to local interactions, the interactions with residuals distant in sequence, but close in space play an important role. Such interactions provide a common topology for laying the polypeptide chain. Local interactions form the initial points of folding, but in the process of further laying the conformations of these sections may change to ensure low overall energy. In addition, each experimental method for determining the structure has an error due to the specific features of the method. The study of occurrence of structurally stable pentapeptides in experimental structures of proteins is the goal of our further research.

References

1. H. Almuallim and T. G. Dietterich, Learning with many irrelevant features. *Proc. 9th National Conf. on Artificial Intelligence*, AAAI Press, 1991, pp. 547–552.
2. P. Comon, Independent component analysis. A new concept. *Signal Processing* **36** (1994), 287–314.
3. D. Darmon, Information-theoretic model selection for optimal prediction of stochastic dynamical systems from data. *Phys. Review E* **97** (2018), No. 3, 032206.
4. L. Ein-Dor, O. Zuk, and E. Domany, Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer. *Proc. Natl. Acad. Sci. USA* **103** (2006), No. 15, 5923–5928.
5. I. T. Jolliffe, *Principal Component Analysis*. Springer–Verlag, New York, 1986.
6. I. Kononenko, Estimating attributes: analysis and extensions of RELIEF. *Proc. 7th Europ. Conf. on Machine Learning*, 1994.
7. A. Kraskov, H. Stoogbauer, and P. Grassberger, Estimating mutual information. *Phys. Review E* **69** (2004), No. 6, 066138.
8. O. F. Lange and H. Grubmuller, Generalized correlation for biomolecular dynamics. *Proteins* **62** (2006), 1053–1061.
9. A. N. Nekrasov, Entropy of protein sequences: an integral approach. *J. Biomolecular Struct. Dynam.* **20** (2002), 87–92.
10. A. N. Nekrasov, Analysis of the information structure of protein sequences: a new method for analyzing the domain organization of proteins. *J. Biomolecular Struct. Dynam.* **21** (2004), No. 5, 615–623.

11. A. N. Nekrasov, L. G. Alekseeva, R. A. Pogosyan, D. A. Dolgikh, M. P. Kirpichnikov, A. G. de Brevern, and A. A. Anashkina, A minimum set of stable blocks for rational design of polypeptide chains. *Biochimie* **160** (2019), 88–92.

12. A. N. Nekrasov, A. A. Anashkina, and A. A. Zinchenko, A new paradigm of protein structural organization. *Theoretical Approaches to BioInformation Systems* (2014), 1–22.

13. B. Scholkopf, R. Herbrich, and A. J. Smola, A generalized representer theorem. *LNAI* (2001), 416–426.

14. T. Suzuki, M. Sugiyama, T. Kanamori, and J. Sese, Mutual information estimation reveals global associations between stimuli and biological processes. *BMC Bioinformatics* **10** (2009), 552.

15. G. D. Tourassi, E. D. Frederick, M. K. Markey, and C. E. Jr. Floyd, Application of the mutual information criterion for feature selection in computer-aided diagnosis. *Medical Physics* **28** (2001), No. 12, 2394–2402.

16. V. Tsurko and A. Michalskii, Contrasting method for selection of informative features using empirical data. *Avtomatika i Telemekhanika* **12** (2016), 136–154 (in Russian).

17. V. Vapnik and R. Izmailov, Statistical inference problems and their rigorous solutions. *Statistical Learning and Data Sciences LNAI* (2015), No. 9047, 33–75.

Appendix A. Nonparametric estimation of mutual information

Substituting representation (1.5) into functional $J_e(\hat{w},\lambda)$, we get

$$J_e(\hat{w},\lambda) = \frac{1}{2n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left(\sum_{l=1}^{n}\alpha_l K(x_i,y_j,x_l,y_l)\right)^2 - \frac{1}{n}\sum_{i=1}^{n}\sum_{l=1}^{n}\alpha_l K(x_i,y_i,x_l,y_l) + \frac{\lambda}{2}\left\|\sum_{l=1}^{n}\alpha_l K(x_i,y_i,x_l,y_l)\right\|_L^2 + C.$$

The first summand is transformed to the form

$$\begin{aligned}
&\frac{1}{2n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left(\sum_{l=1}^{n}\alpha_l K(x_i,y_j,x_l,y_l)\right)^2 \\
&\quad= \frac{1}{2n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\sum_{m=1}^{n}\alpha_l K(x_i,y_j,x_l,y_l)\alpha_m K(x_i,y_j,x_m,y_m) \\
&\quad= \frac{1}{2n^2}\sum_{l=1}^{n}\sum_{m=1}^{n}\alpha_l\alpha_m\sum_{i=1}^{n}\sum_{j=1}^{n}K(x_i,y_j,x_l,y_l)K(x_i,y_j,x_m,y_m) = \frac{1}{2}\sum_{l=1}^{n}\sum_{m=1}^{n}\alpha_l\alpha_m H_{lm}
\end{aligned}$$

where $H_{lm} = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}K(x_i,y_j,x_l,y_l)K(x_i,y_j,x_m,y_m)$.

The second summand is transformed to the form

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{l=1}^{n}\alpha_l K(x_i,y_i,x_l,y_l) = \frac{1}{n}\sum_{l=1}^{n}\alpha_l\sum_{i=1}^{n}K(x_i,y_i,x_l,y_l) = \sum_{l=1}^{n}\alpha_l h_l$$

where $h_l = \frac{1}{n}\sum_{i=1}^{n}K(x_i,y_i,x_l,y_l)$.

Calculate the last summand

$$\begin{aligned}
&\frac{\lambda}{2}\left\|\sum_{l=1}^{n}\alpha_l K(x_i,y_i,x_l,y_l)\right\|_L^2 \\
&\quad=\frac{\lambda}{2}\Big\langle\sum_{l=1}^{n}\alpha_l K(x_i,y_i,x_l,y_l),\sum_{m=1}^{n}\alpha_m K(x_i,y_i,x_m,y_m)\Big\rangle \\
&\quad=\frac{\lambda}{2}\sum_{l=1}^{n}\sum_{m=1}^{n}\alpha_l\alpha_m\big\langle K(x_i,y_i,x_l,y_l),K(x_i,y_i,x_m,y_m)\big\rangle \\
&\quad=\frac{\lambda}{2}\sum_{l=1}^{n}\sum_{m=1}^{n}\alpha_l\alpha_m K(x_l,y_l,x_m,y_m).
\end{aligned}$$

The calculation uses the property of the scalar product in the Hilbert space with the reproducing kernel $K(z,t)$, namely, $\big\langle K(z,u),K(t,u)\big\rangle = K(z,t)$. Denoting the matrix with the elements $K_{ij}=K(x_i,y_i,x_j,y_j)$, by K, we finally obtain the expression

$$J_e(\alpha,\lambda)=\frac{1}{2}\alpha^T H\alpha-\alpha^T h+\frac{\lambda}{2}\,\alpha^T K\,\alpha+C.$$

The minimum of the later functional is attained at the vector

$$\alpha^*=(H+\lambda K)^{-1}h.$$