ISSN 0005-1179 (print), ISSN 1608-3032 (online), Automation and Remote Control, 2023, Vol. 84, No. 12, pp. 1428–1436.

THEMATIC ISSUE

Probabilistic Assessment of a Pentapeptide Composition Influence on Its Stability

A. I. Mikhalskii[*,a], **J. A. Novoseltseva**[*,b], **A. A. Anashkina**[**,c], **and A. N. Nekrasov**[***,d]

[*] *Trapeznikov Institute of Control Sciences, Russian Academy of Sciences, Moscow, Russia*
[**] *Engelgardt Institute of Molecular Biology, Russian Academy of Sciences, Moscow, Russia*
[***] *Shemyakin–Ovchinnikov Institute of Bioorganic Chemistry, Russian Academy of Sciences, Moscow, Russia*
e-mail: [a] *ipuran@yandex.ru,* [b] *novoselc.janna@yandex.ru,* [c] *a_anastasya@inbox.ru,* [d] *a_nnekrasov@mail.ru*

Received May 31, 2023
Revised September 12, 2023
Accepted September 30, 2023

Abstract—The influence of the arrangement of amino acid residues in a pentapeptide on its stability is being studied. A forecast of pentapeptide stability is made using the gradient boosting method, which allows one to evaluate the influence of each feature on the stability of the pentapeptide. Combinations of amino acid arrangements in the pentapeptide have been identified that make a significant contribution to its stability. It has been shown that the use of such combinations reduces the amount of data required to obtain a reliable prediction of pentapeptide stability.

Keywords: amino acid residue, pentapeptide, gradient boosting, prediction, information sufficiency

DOI: 10.25728/arcRAS.2024.26.65.001

1. INTRODUCTION

The problem of predicting the spatial structure of proteins is one of the priority tasks in the field of mathematical and biological modeling, leading to practical application—the design of new proteins with useful medical properties. Currently, there is a tool for predicting the tertiary structure of a protein from its amino acid sequence, AlphaFold [1], which has shown incredible accuracy of structure prediction, comparable to the accuracy of X-ray diffraction analysis on CASP [2]. However, this tool is based on a deep neural network and the principles behind the styling remain unexplored. Understanding which amino acids and in what combination contribute to increasing the stability of a protein fragment will allow us to create a method for designing protein structure. The goal of the work is to, based on experimental data, identify potential markers of pentapeptides stability (combinations and positions of amino acids in the molecule).

The study of the entropy characteristics of fragments of protein sequences showed that for five sequentially located residues a reduced level of information entropy is observed and, therefore, blocks of this particular size must be considered as elementary units of the sequence. This approximation made it possible to develop a method that reveals the hierarchical structure in protein sequences — the method of analyzing information structure (ANIS method) [3]. Analysis of the conformational stability of pentapeptides by the molecular dynamics method showed that all pentapeptides can be divided into three types [4]: conformationally stable (located in the predominant topology for more than 80% of the simulation time), triggered (having two predominant topologies, in each of which

the peptide was present for at least 40% of the simulation time) and labile. Molecular dynamics is a technique in which the time evolution of a system of interacting atoms or particles is tracked by integrating their equations of motion. Classical mechanics is used to describe atoms or particles and their motion. The law of particle motion is found using analytical mechanics, and the forces of interatomic interaction are represented in the form of classical potential forces (as the gradient of the potential energy of the system).

2. DATA

The work used 44 860 pentapeptides, the sequences of which were created according to a certain rule, and 4885 previously studied pentapeptides from real proteins, the stability of which was determined by molecular dynamics modeling. In the resulting set of 49 745 pentapeptides, only 1705 pentapeptides turned out to be stable, which was about 3.43% of the total number.

During the study, all data were randomly divided into "training", "control" and "validation" samples in proportions of 0.66, 0.17, 0.17, maintaining the original balance of classes. The training and control samples were used during the training phase. The training sample was also used at the stage of interpretation of the results.

2.1. Data Encodings

In the original data set, each pentapeptide is encoded by a sequence of five letters representing the amino acid residues included in the pentapeptide. The order of the letters corresponds to the sequence of amino acid residues in the pentapeptide molecule. For formal numerical analysis of the data, the five-letter representation was encoded using three different representations. Binary encoding (One Hot Encoding), continuous string representation (n-gram), and discontinuous string representation (broken n-gram) were considered. Each of the considered coding methods allows one to evaluate contributions in its own way and make judgments about the influence of certain combinations of amino acid residues on the stability of the pentapeptide.

2.2. Binary Encoding (OHE)

One hot encoding is an encoding in which the presence of each amino acid at its position is determined by the position of one in the vector, the remaining coordinates of which are equal to zero. The number of vector elements is 20 — the number of amino acid types. As a result, each pentapeptide is encoded by a matrix of 20 rows and 5 columns. The column corresponds to the amino acid position in the pentapeptide molecule, and the row corresponds to the amino acid. For example, when classifying amino acids by the first letter of the name, the pentapeptide DKLNV will be encoded by a matrix in which the first column in the third row is 1, the remaining elements are equal to zero, the second column in the ninth row is 1, the remaining elements are equal to zero, etc. In the calculations, each pentapeptide is represented as a vector in 100-dimensional space.

2.3. Continuous String Representation (n-gram)

An n-gram is a continuous string representation of the amino acid sequence in a pentapetide. Depending on the number of letters included in the string, n-grams of order 1, 2 or more are distinguished. For example, the peptide DKLNV is encoded by five n-grams of order 1 D, K, L, N, V, four n-grams of order 2 DK, KL, LN, NV, three n-grams of order 3 DKL , KLN, LNV. In the analysis carried out, n-grams from 1 to 3 were used. As with OHE encoding, the entire set of n-grams encoding pentapeptides is represented in the form of a table consisting of zeros and ones. In each column of the table, a certain row contains one, and the remaining elements are zeros.

2.4. Discontinuous String Representation (Broken n-gram)

A broken n-gram is a generalization of a continuous n-gram and is a string representation of the sequence of amino acids in a pentapeptide, in which there is a gap of one to three characters between groups of amino acids. When forming a broken n-gram, the amino acids included in the n-gram are indicated, the position of the first amino acid from the n-gram in the pentapeptide molecule is indicated and the number of positions between each of the amino acids included in the n-gram. For example, for the pentapeptide DKLNV there is a second-order broken n-gram 12DN, where 1 is the position of the first amino acid, 2 is the number of positions between amino acids, DN is the list of amino acids included in the broken n-gram. For this pentapeptide there are only six broken n-grams of order 2, namely 11DL, 21KN, 31LV, 12DN, 22KV, 13DV. The study considered broken n-grams of order 2 only.

3. CLASSIFICATION ALGORITHM

To classify pentapeptides into stable and unstable, we used the gradient boosted decision trees algorithm [5]. The algorithm is built on the principle that a relatively weak machine learning algorithm can be strengthened by the same algorithm, which will "refine" the predictions of the previous algorithm based on its errors. When applying this principle to random forest classification, the first row of trees is trained on real data, predicting the class label for each object. The second row of trees is trained on the same data, but giving more weight to objects where the first row of trees made mistakes and correcting them. The trees of the third row are trained by correcting the errors of the trees of the second row, etc. Currently, gradient boosting over decision trees is one of the most popular machine learning algorithms, because at low training costs it provides high accuracy and protection from overfitting due to the fact that a random forest of decision trees is used. In this case, the features and subsample are mixed to construct a new tree. In addition, the obtained result is easy to interpret.

Quality control of training was carried out using the metric F_1, specified by the formula

$$F_1 = 2\frac{precision * recall}{precision + recall}.$$

In this case, one class is considered as a class of "positive objects", for example, a class of stable pentapeptides, and the other is a class of "negative objects". The *precision* metric determines the proportion of correctly identified positive objects among all objects classified as positive. The *recall* metric determines the proportion of correctly identified positive objects among all positive objects. Metric F_1 used to assess the quality of classification in the case of data in which the classes are significantly unbalanced.

The parameters of the classification algorithm were adjusted for each encoding method used using the cross-validation procedure in a high-dimensional space [6] using the *hyperopt* package. Table 1 shows the classification results achieved with the found settings.

Table 1. Results of classification of pentapeptide stability for different encoding methods

Encoding	Metrics		
	precision	*recall*	F_1
OHE	0.39	0.54	0.45
n-grams	0.39	0.41	0.40
broken n-grams	0.32	0.54	0.40

The best quality in terms of F_1 metrics is achieved when using OHE encoding. For n-gram and broken n-gram encodings, the quality is lower. This is explained by the small sample size and large number of features when encoding using n-grams and broken n-grams.

4. PROBABILISTIC ASSESSMENT OF THE SIGNIFICANCE OF AMINO ACID POSITIONS IN A PENTAPEPTIDE

In addition to assessing the quality of classification, it is of great interest to assess the importance of individual features in the stability of pentapeptides. To construct such an estimate using gradient boosting, the SHAP (SHapley Additive exPlanations) algorithm was used in this study [7], which allows you to estimate the probabilistic contribution of each combination of amino acids to the probability of classifying a pentapeptide as stable, taking into account the interaction of factors (amino acids and their positions) with each other. This method calculates the importance of a particular feature by comparing the results obtained with and without that feature. When constructing a classification rule in the form of a tree, the result can be influenced by the order in which the elements of the training set are used. To eliminate such an influence on the assessment of the importance of a feature, elements of the training sample are received for training many times in a random sequence.

The SHAP method was justified in the theory of cooperative games, when game participants can unite in coalitions to achieve the best result. The payoff of each player is equal to his average contribution to the total payoff over all coalitions under a random, equally probable ordering of the participants. This value is called the Shapley index [7] and is calculated by summing over all sets of features that do not include feature i, the weighted effect of using the excluded feature. In this case, the effect of using feature i is understood as the difference in the classification accuracy of a pentapeptide taking into account feature i and without taking it into account. The Shapley index is calculated using the formula

$$\Phi_i = \sum_{S \in F \setminus i} \frac{n_S!\,(n_F - n_S - 1)!}{n_F!} \left(f_{S \bigcup i} - f_S \right),$$

here F denotes the set of all possible sets of features, $F \setminus i$ denotes the set of sets of features that do not include the feature i, S – a set of features without feature i, $S \bigcup i$ – set of features S with the addition of feature i, f_S and $f_{S \bigcup i}$ – classification accuracy when using feature sets S and $S \bigcup i$, respectively, n_F and n_S – number of feature sets in the sets F and S, respectively. The significance of a feature is determined by the absolute value of the corresponding Shapley index.

5. INTERPRETATION OF RESULTS

Below are the results of interpretation using the SHAP method of the results of classifying the stability of pentapeptides by the gradient boosting algorithm using three different encodings.

5.1. Binary Encoding (OHE)

Table 2 presents an example of assessing the influence of the position of amino acids in the DRNAA pentapeptide on its stability. It is important to note that the stability of a pentapeptide is affected not only by the presence of an amino acid at any position, but also by its absence.

In Table 2 rows are ordered in order of decreasing probabilistic contribution of amino acids and their positions to the stability of the pentapeptide. Negative values indicate a negative impact on stability. It follows from the table that the presence of amino acid D in the first position in the fifth position increases the probability that the pentapeptide is stable, and the absence of amino

Table 2. Probabilistic contribution of amino acids and their positions on the stability of the DRNA pentapeptide

Amino acid		position	Probabilistic contribution
presence	absence		
D		1	0.048
R		2	0.018
	A	1	0.010
A		4	0.0040
A		5	–0.0083
N		3	–0.0096

acid A in the first position increases the probability that the pentapeptide is stable by only 1%. The presence of amino acid A at the last position reduces the probability of pentapeptide stability by 0.8%. It is assumed that the features influence the stability of the pentapeptide independently of each other.

If a similar probabilistic analysis is carried out for a variety of pentapeptides, the overall result can be presented in the form of a diagram of the probabilistic contributions of amino acids and their positions to stability. Figure 1 shows a diagram for the most significant features. Due to the great computational difficulties associated with the need to solve the classification problem for all possible sets of features, calculations were carried out for 1000 randomly selected pentapeptides. In the diagram, a single point corresponds to the result of an analysis of a single pentapeptide. The presence of a feature (the presence of an amino acid at a specified location) is represented by an open symbol, and its absence is represented by a closed symbol.

The figure shows that if the pentapeptide contains amino acid K at the fifth position, it has the greatest positive effect on its stability. The opposite effect is that amino acid A in the first position has a negative effect on stability.

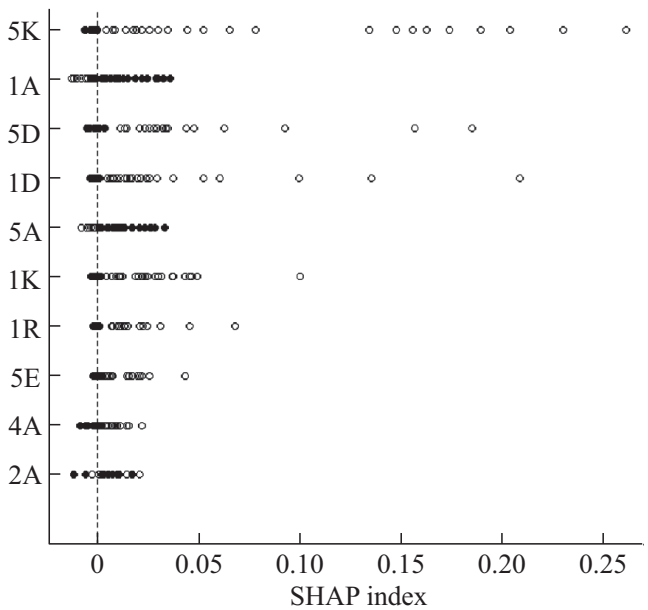

Fig. 1. Diagram of the probabilistic contributions of features to the stability of 1000 randomly selected pentapeptides under OHE encoding, constructed using the SHAP algorithm.

5.2. Continuous String Representation

When encoding using n-grams, the estimate of the probabilistic contribution to the stability of an individual feature turns out to be less than when encoding OHE. This is a consequence of the fact that when using n-grams up to the third order, the number of features is 256 times greater than with ONE encoding. Table 3 shows examples of estimates of the probabilistic contribution to the stability of the DRNAA peptide.

In Table 3 rows are ordered in order of decreasing probabilistic contribution of amino acids and their positions to the stability of the pentapeptide. The table shows that when encoding using n-grams, the joint contribution of amino acids D and R located in the first and second positions to the stability of the pentapeptide is estimated at about 0.5%, whereas with OHE coding the estimate is 6%.

Table 3. Examples of assessing the probabilistic contribution to the stability of the DRNAA pentapeptide when encoded using n-grams

Combination of amino acids		position	Probabilistic contribution
presence	absence		
D		1	0.0030
R		2	0.0022
	K	2	–0.00004
	EK	1	–0.00008
	T	5	–0.000028
	R	5	–0.000029
A		5	–0.0001

Figure 2 shows an example diagram of the probabilistic contributions of amino acids and their positions to the stability of 1000 randomly selected pentapeptides. The figure shows that single combinations of amino acids have the greatest significance; amino acid K in the fifth position has the most powerful positive effect, and amino acid A in the first position has the most negative effect.

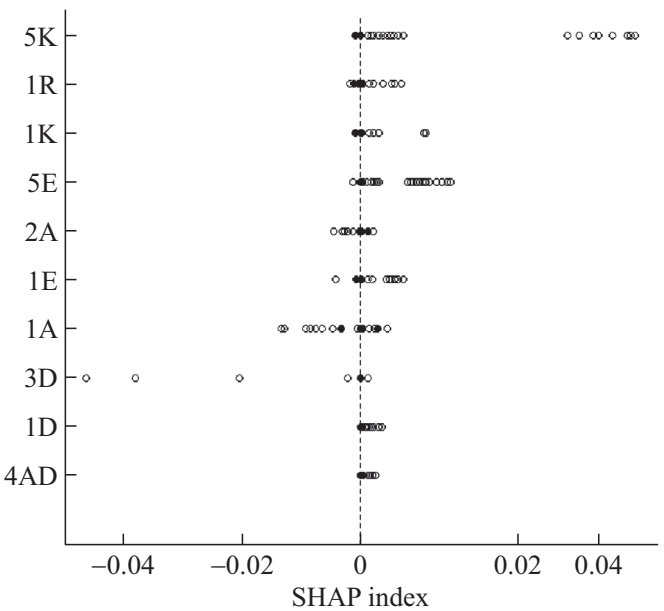

Fig. 2. Diagram of the probabilistic contributions of features to the stability of 1000 randomly selected pentapeptides when encoded using n-grams, constructed using the SHAP algorithm.

5.3. Discontinuous String Representation

Table 4 presents the result of estimating the probabilistic contribution to the stability of an individual feature using the example of the DRNAA pentapeptide when encoded with a discontinuous n-gram. Figure 3 shows an example of a diagram of the probabilistic contributions of amino acids and their positions to the stability of 1000 randomly selected pentapeptides with the same encoding.

Table 4. Examples of assessing the probabilistic contribution to the stability of the DRNAA pentapeptide when encoded with a broken n-grams

Combination of amino acids		position	Probabilistic contribution
presence	absence		
R–A		2	0.0093
R–A		2	0.0041
	A–A	1	0.0031
D–A		1	0.0020
	A–A	1	0.0013
	A–K	2	–0.0014
D–N		1	–0.0030

In Table 4 rows are ordered as the probabilistic contribution of the combination of amino acids and their positions to the stability of the pentapeptide decreases. It follows from the table that the greatest effect on the stability of the DRNAA pentapeptide is exerted by the combination of amino acids R in the second position and A in the fourth or fifth position. The absence of amino acid A in the first position and simultaneously in the fourth or fifth positions also increases the likelihood of stability of the DRNAA pentapeptide, but to a lesser extent.

Figure 3 shows that combinations with amino acid A in the second and K in the fifth position have the greatest significance for stability. The presence of two amino acids A in a pentapeptide with two or three gaps between them, on the contrary, is a sign of its instability.

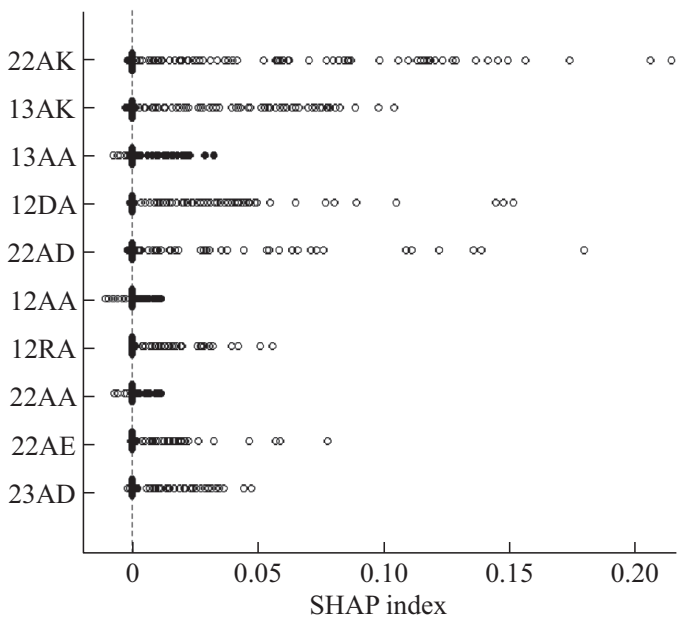

Fig. 3. Diagram of the probabilistic contributions of features to the stability of 1000 randomly selected pentapeptides when encoded with a broken n-gram, constructed using the SHAP algorithm.

6. CONCLUSION

The article discusses the result of using three different encodings of the pentapeptide structure when predicting its stability through solving the classification problem. Binary encoding (One Hot Encoding), continuous string representation (n-gram), and discontinuous string representation (broken n-gram) were considered. Each of the encodings generates feature spaces of different dimensions: 100 when using the OHE binary encoding, 25 600 when encoding using n-grams no higher than the third order, and 10 400 when using a discontinuous n-gram. This creates varying degrees of data sparsity. The problem of classifying pentapeptides into stable and unstable was solved by the gradient boosting method (LGBM). The study used a set of 49 745 pentapeptides, of which 3.43% were stable. The data was randomly divided into "training", "testing" and "validation" sets in proportions while maintaining the original class balance. After training, the results of testing on the control sample for each of the encodings showed approximately the same value for the quality metric F_1, equal to 0.45 for binary encoding and 0.40 when using different n-grams.

Assessing the importance of features for predicting the stability of pentapeptides highlighted the most important features. Each encoding method has its own characteristics. OHE coding evaluates the importance of the location of a particular amino acid at a particular position. When using n-grams, encoding allows one to evaluate the importance of the combination of amino acids at adjacent positions, and when using broken n-grams, the importance of the arrangement of amino acids at positions distant from each other is assessed. Encoding using broken n-grams makes it possible to highlight the effect of the influence of a combination of amino acids located in different positions of the pentapeptide molecule.

The question of the structural stability of pentapeptides was considered in [8]. In this work, a problem dimensionality reduction method based on the calculation of mutual information between the stability feature and the pentapeptide description was used in the binary OHE encoding. It turned out that dimensionality reduction using mutual information allows one to use the "simple" K-nearest neighbors classification method for stability prediction. At the same time, the quality of the result in terms of the "accuracy" and "completeness" metrics practically coincides with the result of using the "random forest" method, which requires significantly greater computational and time costs. A probabilistic assessment of the effect of pentapeptide composition on its stability was not performed in that study. In this work, the emphasis was placed on assessing the influence of the pentapeptide composition and the results of such an assessment are presented for 1000 randomly selected pentapeptides, which is associated with large requirements for the necessary computing power.

REFERENCES

1. Senior, A.W., Evans, R., Jumper J., et al., Improved Protein Structure Prediction Using Potentials from Deep Learning, *Nature*, 2020, vol. 577, pp. 706–710.

2. Pereira, J., Simpkin, A.J., Hartmann, M.D., et al., High Accuracy Protein Structure Prediction in CASP14, *Proteins Structure Function and Bioinformatics*, 2021, vol. 89, no. 12, pp. 1687–1699. https://doi.org/10.1002/prot.26171

3. Nekrasov, A.N., Kozmin, Yu.P., Kozyrev, S.V., et al., Hierarchical Structure of Protein Sequence, *Int. J. Mol. Sci.*, 2021, vol. 22, no. 15, 8339. https://doi.org/10.3390/ijms22158339

4. Anashkina, A.A., Nekrasov, A.N., Alekseeva, L.G., et al., A Minimum Set of Stable Blocks for Rational Design of Polypeptide Chains, *Biochimie*, 2019, vol. 160, pp. 88–92.

5. Ke, G., Meng, Q., Finley, T., Wang, T., et al., A Highly Efficient Gradient Boosting Decision Tree, *Proc. 31st Conference on Neural Information Processing Systems (NIPS). Long Beach*, 2017, pp. 3149–3157.

6. Bergstra, J., Yamins, D., and Cox, D.D., Making a Science of Model Search: Hyperparameter Optimization in Hundreds of Dimensions for Vision Architectures, *Proc. of the 30th International Conference on Machine Learning (ICML)*, 2013. pp. 115–123.

7. Lundberg, S.M. and Lee, S.I., A Unified Approach to Interpreting Model Predictions, *Proc. 31st Conference on Neural Information Processing Systems (NIPS). Long Beach*, 2017, pp. 4765–4774.

8. Mikhalskii, A.I., Petrov, I.V., Tsurko, V.V., Anashkina, A.A., et al., Application of Mutual Information Estimation for Prediction the Structural Stability of Pentapeptides, *Rus. J. Numer. Anal. Math. Model.*, 2020, vol. 35, no. 5, pp. 263–271.

This paper was recommended for publication by A.A. Galyaev, a member of the Editorial Board