ISSN 0006-3509, Biophysics, 2019, Vol. 64, No. 2, pp. 182–194. © Pleiades Publishing, Inc., 2019.
Russian Text © The Author(s), 2019, published in Biofizika, 2019, Vol. 64, No. 2, pp. 256–271.

MOLECULAR BIOPHYSICS

The Conformational Stability/Lability of Peptide Fragments in the Sequence Context of Amino Acids

I. Yu. Torshin[a], *, A. V. Batyanovskii[b], L. A. Uroshlev[c], V. G. Tumanyan[d], I. D. Volotovskii[b], and N. G. Esipova[d]

[a] *Dorodnicyn Computing Center, Russian Academy of Sciences, Moscow, 119333 Russia*
[b] *Institute of Biophysics and Cell Engineering, National Academy of Sciences of Belarus, Minsk, 220072 Belarus*
[c] *Vavilov Institute of General Genetics, Russian Academy of Sciences, Moscow, 119991 Russia*
[d] *Engelhardt Institute of Molecular Biology, Russian Academy of Sciences, Moscow, 119991 Russia*
**e-mail: tiy135@yahoo.com*

Received November 14, 2018; revised December 11, 2018; accepted December 14, 2018

Abstract—Criteria for evaluating the conformational stability/lability of peptide fragments referred to fragments of protein structures are formulated. Using the proposed criteria, a statistical analysis of tetrapeptide fragments (their conformations and sequences) was performed in a sample of 25 121 protein chain structures from the PDB protein databank. As a result of the analysis, it was shown that tetrapeptide fragments significantly differ in the degree of the conformational stability/lability from the point of view of the proposed statistical criteria. The results of tetrapeptide denaturing molecular dynamics simulations were used as an independent approach to estimate the stability/lability of peptide fragments. A correlation between the estimates of conformational lability obtained on the basis of a statistical analysis of the ensembles of peptide conformations observed in experimentally determined protein structures and the estimates of conformational lability/stability calculated on the basis of molecular dynamics trajectories is demonstrated. Subgroups of more "conformationally stable peptides," characterized mainly by the α-helical conformation, were obtained. Consensus tetrapeptides characterized by the lowest conformational lability (the highest conformational stability) were determined using complex criteria. Peptides with increased conformational lability were described. Thus, among all combinatorially possible tetrapeptides, the tetrapeptides that are characterized by certainty about their conformational state play a significant role. The relationship between the degree of conformational certainty of the peptide and its involvement in the primary structure of the protein was characterized. It was found that the role of the conformationally stable peptides in the formation of the protein structure was considerable, since they constitute, on average, approximately 10% of the amino-acid sequence. Using real soluble peptides as examples, the possibility of assessing the conformational stability of any preset amino-acid sequence on the basis of the developed criteria of the conformational lability of tetrapeptides was shown.

Keywords: conformationally stable/labile protein segments, local protein structure, conformational analysis, statistical analysis, molecular dynamics

DOI: 10.1134/S0006350919020180

The problem of determining the structure of a protein from its sequence remains one of the key challenges of modern biology. The results of previous studies indicate the extreme complexity of this problem; it seems reasonable to formulate certain subproblems within the general problem, in particular, to investigate the "sequence–structure" relationship as a function of a peptide length. As follows from a priori considerations, the structural certainty (or "conformational stability") should increase with sequence elongation. In fact, the difference in energy (free energy) of the conformations of a short peptide is significantly smaller than that of a long peptide. In principle, a situation is possible where the structural certainty becomes complete; it occurs only when a certain length of the amino-acid sequence is reached (although even individual amino-acid residues tend to concentrate in certain areas of the Ramachandran map [1]).

The question arises as to whether there is a correspondence between the sequence and the structure at the level of sequences shorter than the native protein chain. One particularly promising research problem is to analyze the local structures that correspond to short segments of a polypeptide chain (containing two, three, or four amino acids) and to reveal possible con-

Abbreviations: MD, molecular dynamics.

sistencies of the relationship between the conformation of such peptide segments and their amino-acid sequences. It is of interest to study such short segments, because in this case all possible sequences can be listed. In the most general form, the problem can be considered as a classification of objects in two groups of dissimilar character descriptions: the amino-acid sequence and the spatial conformation.

This research problem can be solved, first, by studying the common conformations for similar amino-acid sequences; second, on the basis of classification of the observed conformations, it is possible to attempt to identify a set of amino-acid sequences that are characteristic of a conformation of a particular type. The approach of systematization of oligopeptides with a certain sequence according to the structure or conformation (by the criterion of the values of dihedral angles in the main chain) taken by them was used for the analysis and prediction of the secondary structure of the protein [2–4] and extrapolated to the common-type conformations. Studies [5, 6] are examples of a successful variant of correlation of conformations with certain amino-acid sequences.

In an alternative methodology [6], fragments are systematized at a structural level according to the character of their structural similarity with distinguishing structural clusters. A particular amino-acid sequence is affiliated with each structural cluster and the structure of the canonical fragment can ultimately be established only on the basis of sequence information. Study [7] is an example of a combined approach, when multidirectional relationships between the conformation and the sequence are taken into account.

As a working hypothesis, we assume the existence of sufficiently short fragments of a polypeptide chain that are characterized by a certain preferable conformation. We define a conformationally stable oligopeptide as a segment of a polypeptide chain with a certain amino-acid residue sequence that exhibits distinct conformational preferences in a sufficiently large sample of protein structures. Obviously, this qualitative definition requires quantitative criteria for estimating the "explicitness" of conformational preferences.

The search for such "conformationally stable" tetrapeptides was started in [8]. The conformationally stable peptides were distinguished by the constancy of the conformation of the same sequence in different proteins. In particular, the oligopeptides for which more than 65% of the observed conformational states were similar were regarded as conformationally stable. The degree of conformational similarity within this basic conformational type also had to be sufficiently large (the difference in two dihedral angles should not exceed 10°). As a result, a list comprising 900 α-helical and approximately 50 non-α-helical conformationally stable peptides was obtained [9].

The selection criteria for the "conformationally stable" peptides that were used in [9] need additional argument in terms of the selection of numerical values of criteria. One of the approaches to refining the numerical criteria for the selection of conformationally stable peptides can be based on assessing the statistics of the distribution of the selected parameters. Another approach that is used in this work consists in introducing a number of more general parameters of conformational stability (or, more conveniently, lability) for the investigation of the phenomenon of the occurrence of "conformationally stable" peptides in protein structures. Parameters of this type were introduced in this study and used for the analysis of a representative sample of proteins from the PDB database.

MATERIALS AND METHODS

Data sample. The structures of proteins were taken from the Protein Data Bank (PDB) protein database [10] release 2017. Using the VAST resource [11], files with a resolution not lower than 2.0 Å and a pairwise sequence identity of not more than 50% were selected from the PDB database. As a result, a sample comprising the structures of 25121 protein chains was formed. For each of the chains of the amino-acid residues from known coordinates of each non-hydrogen atom of the main chain, the values of angles φ and ψ were calculated and the division into tetrapeptides was performed. The analysis included only the tetrapeptide fragments for which all eight dihedral angles were known.

Methods for estimating the conformational lability and stability of peptide fragments on the basis of analysis of the sets of observed conformations of peptides with a certain sequence. Let P be an m-peptide ($m = 2$, 3, 4, ...) with a given amino-acid sequence. We describe an arbitrary conformation of the m-peptide P by setting a set of angles φ and ψ ordered in the form of vector $\vec{p}_k \in R^{2m}$ with dimension $2m$, $\vec{p}_k = (p_1^k, ..., p_i^k, ..., p_{2m}^k)$, $p_i \in R$, where the components p_i with odd values of index i correspond to angles φ and the components p_i with even i values corresponded to angles ψ.

We assume that in the vector space R^{2m} a scalar product is determined; i.e., for arbitrary $\vec{p}_1, \vec{p}_2 \in R^{2m}$ values, the scalar $\langle \vec{p}_1, \vec{p}_2 \rangle = \sum_{i=1}^{2m} p_i^1 p_i^2$ is defined. Accordingly, for any vector pair $\vec{p}_1, \vec{p}_2 \in R^{2m}$, the distance $\mathrm{dist}(\vec{p}_1, \vec{p}_2) = \sqrt{\langle \vec{p}_1 - \vec{p}_2, \vec{p}_1 - \vec{p}_2 \rangle}$ and the angle $\mathrm{ang}(\vec{p}_1, \vec{p}_2) = \arccos\left(\frac{\langle \vec{p}_1, \vec{p}_2 \rangle}{\sqrt{\langle \vec{p}_1, \vec{p}_1 \rangle \langle \vec{p}_2, \vec{p}_2 \rangle}}\right)$ is defined.

For further convenience, we introduce some additional notations. We preset a finite set of numbers $A = \{a_1, a_2, ... a_i, ... a_n\}$, $a_i \in R$. We call $\hat{\phi}(x)$ the *operator of formation of the empirical distribution function* for the numbers in set A, $\hat{\phi}(x)A =$

$\sup|\{B \subseteq A|\forall a \in B : a \le x\}|/|A|, x \in R$. For brevity, we write $\hat{\phi}(x)A$ also as $\hat{\phi}A$.

We call $\hat{z}$ *the operator of the formation of the set of values* of set A (e.g., $A \in R$); $\hat{z}A = B \subseteq A|\forall a \in A : a \in B, \forall a, b \in B : a \ne b$, and we call $\hat{\imath}^+$ *the operator of ordering the set in the ascending order*; $\hat{\imath}^+A = (a_{I(1)}, a_{I(2)}, \ldots| a_{I(1)} \le a_{I(2)} \le \ldots a_{I(j)} \ldots \le a_{I(n)}, \forall a_i \in A\ \exists j \in N[1\ldots n] : i = I(j))$, where $I(j) : N[1...n] \to N[1...n]$ is a suitable *renumbering function* (enumerator) for elements of A. Then, the jth element of the ordered set $\hat{\imath}^+A$ can be denoted as $\hat{\imath}^+(j)A = a_{I(j)}$.

We then define $\hat{\mu}$, the operator for calculating the mathematical expectation of the value $x \in A$ by the empirical distribution functions $\hat{\phi}A$ as $\hat{\mu}\hat{\phi}A = \frac{1}{|\hat{z}A|}\sum_{j=1}^{|\hat{z}A|} x_j(\hat{\phi}(x_j)A - \hat{\phi}(x_{j-1})A)$, where $x_j = \hat{\imath}^+(j)\hat{z}A$, and the arbitrary value $x_0 < \inf(A), x_0 \in R$. We define the *operator* $\hat{\sigma}$, *which calculates the standard deviation* $x \in A$ by $\hat{\phi}A$ as $\hat{\sigma}\hat{\phi}A = \sqrt{\frac{1}{|\hat{z}A|}\sum_{j=1}^{|\hat{z}A|}(x_j - \hat{\mu}\hat{\phi}A)^2(\hat{\phi}(x_{j-1})A - \hat{\phi}(x_j)A)}$.

We set n_1 conformations for the m-peptide P with a fixed amino-acid sequence, with the vector $\vec{p}_k \in R^{2m}$, $k = 1, \ldots, n_1$ corresponding to each of these conformations. We call the set of vectors $C(P) = \{\vec{p}_k\}$, $k = 1, \ldots, n_1$ the *set of observed conformations* of peptide P. We let $A_1 \subset R$ be a set of pairwise distances between the elements of set C(P): $A_1(C(P)) = \{\text{dist}(a, b)|a, b \in C(P), a \ne b\}$ and the set $A_2 \subset R$ be the set of pairwise values of angles between the elements of set C(P):

$$A_2(C(P)) = \{\text{ang}(a,b)|\ a,\ \ b \in C(P),\ \ a \ne b\}.$$

We set the alphabet $B = \{b_1, \ldots, b_{|B|}\}$, to describe the secondary structure of a single amino acid. One example is the alphabet B = {"α," "β," "*C*"}, where "α" corresponds to α-helix, "β" corresponds to β-strand, and "*C*" corresponds to the unstructured fragments (coil). Let us define the function $\Phi : R^2 \to B$, which determines the correspondence of a symbol from alphabet B to the values of the angle pair φ and ψ (such a function can be determined, for example, in accordance with the DSSR standard).

For the given peptide P, a set of letters of the secondary structure $L(C(P)) = \{\Phi((p_i^k, p_{i+1}^k)|\ i = 2j + 1, j = 0, \ldots, m - 1, k = 1, \ldots, n_1)$ unambiguously corresponds to the set of the observed conformations $C(P) = \{\vec{p}_k\}$. Then, for an arbitrary letter $b \in B$, we define the frequency of occurrence ν_b as $\nu_b(L(C(P)) = |\{a \in L(C(P)a = b\}|/|L(C(P)|$. Thus, for the alphabet B = {"α," "β," "*C*"}, the frequencies of occurrence of respective letters ν_α, ν_β, and ν_C are determined for each peptide P.

In this study, we investigate the conformational stability of peptide fragments. However, on the basis of the sets of vectors $C(P) = \{\vec{p}_k\}$ for the m-peptide P with a preset amino-acid sequence, it is practically much more convenient to estimate the conformational lability of peptides. Then, after ordering the considered m-peptide fragments in accordance with the numerical values of the conformational lability estimates, the upper part of such a list will correspond to the more "conformationally stable" m-peptides, whereas the lower part of this list will correspond to the more "conformationally labile" m-peptides.

Using the notations introduced above, we formulate several approaches to estimate the conformational lability of the peptide fragments represented in the protein structures. In this study, five estimates of the conformational lability of an arbitrary m-peptide P were used:

$s_1(P) = \hat{\mu}\hat{\phi}A_1(C(P))$, i.e., the mean distance between the elements of set C(P);

$s_2(P) = \hat{\sigma}\hat{\phi}A_1(C(P))$, i.e., the standard deviation of the distance between the elements of set C(P);

$s_3(P) = \hat{\mu}\hat{\phi}A_2(C(P))$, i.e., the mean angle between the elements of set C(P);

$s_4(P) = \hat{\sigma}\hat{\phi}A_2(C(P))$, i.e., the standard deviation of the angle between the elements of set C(P);

$s_5(P) = 1 - \max(\nu_\alpha(L(C(P))), \nu_\beta(L(C(P))), \nu_C(L(C(P))))$, i.e., unity minus the maximum ν_α, ν_β, ν_C frequency for peptide P.

In this study, we also used the *composite estimates of the conformational lability* of m-peptides, i.e., the estimates representing the compositions of the above-formulated estimates s_1–s_5. We consider an example of constructing a composition of two estimates x, y. We assume that there are correlations between the estimates x, y that are described by the functions $y = f(x)$ and $x = f^{-1}(y)$. The specific form of functions $f(x)$ and $f^{-1}(y)$ is established as a result of corresponding regression analyzes. Then, the compositions of estimates x and y are introduced by correlation transformations, i.e., as $s_{xy} = x + f^{-1}(y)$ or as $s_{yx} = y + f(x)$. Estimates s_{125}, s_{345}, etc., are obtained similarly.

A method to assess the conformational lability and stability of peptide fragments on the basis of denaturing molecular dynamics simulations. The procedure for denaturing molecular dynamics (MD) simulation in the variant that was described in detail in [12] can be used to assess the conformation stability of a system during heating. To perform MD simulation, we used the ECMMS software package [13] with the UFF force field [14]. The coordinates of the hydrogen atoms were calculated proceeding form the standard geometry. The "denaturing" molecular dynamics sim-

Table 1. Correlation analysis of the conformational lability estimates s_1–s_5

Estimate	s_1	s_2	s_3	s_4	s_5
s_1	1.00	0.67	0.93	0.72	0.63
s_2		1.00	0.75	0.88	0.17
s_3			1.00	0.70	0.62
s_4				1.00	0.20
s_5					1.00

The table summarizes the correlation coefficient values. Since the correlation matrix is symmetrical, only the values for $n(n + 1)/2$ independent elements of the matrix are shown.

ulation was performed using a modified jump algorithm [15] in combination with a thermostat (400, 600, and 1000 K) without solvent model (in order to increase the denaturing effect) and without obvious potential to describe the hydrogen bonds.

The set of Cartesian coordinates of the initial conformation of the test peptide P was selected on the basis of a certain median conformation, which was calculated on the basis of values of angles φ and ψ. In terms of the notations introduced above, the set of observed conformations C (P) = $\{\vec{p}_k\}$, $\vec{p}_k \in R^{2m}$, $k = 1, ..., n_1$ corresponds to peptide P. We call a vector $\vec{a} \in$ C(P), for which the sum of the distances to other elements of C(P) is minimal (i.e., $\vec{a} = \{\vec{p}_k \in \text{C(P)} | \sum_{j=1...n_1} \text{dist}(\vec{p}_k, \vec{p}_j) \to \min\}$) the median conformation of peptide P.

On the basis of the median conformation $\vec{a}$, an appropriate chain was taken from the PDB database and the Cartesian coordinates corresponding to $\vec{a}$ were used as the initial coordinates of the simulated peptide. The initial velocities were set in accordance with the Maxwell distribution for a given temperature. For each peptide fragment, 100-ps trajectories were calculated.

The conformational lability of the test peptide was characterized by comparing the conformation of the peptide at a given simulation time and a similar conformation ($t = 0$). The differences in the conformations were estimated by the root-mean-square deviation of atoms in the main chain ("*rmsd*") and the correlation coefficient "*rmap*" between the values of pairwise distance matrix elements. The *rmap* coefficient characterizes the similarity between the structure at a given time and the initial conformation of the peptide (with the accuracy of the translation and rotation). As a result of MD simulations, *rmsd*(t) and *rmap*(t) curves for each peptide were obtained.

The values designated as AUC_{rms} and AUC_{rmap} were used as integrated measures of conformational lability/stability, calculated on the basis of MD trajectories of peptides. $AUC_{\text{rms}} = \int_{t=0}^{t_{MD}} rms(t)d(t)$ is the area under the *rmsd*(t) curve; it characterizes the degree of deviation of the atoms of the main chain of a peptide fragment from the initial conformation during the simulation time t_{MD} (100 ps), i.e., the conformational lability of the peptide fragment. Accordingly, $AUC_{\text{rmap}} = \int_{t=0}^{t_{MD}} rmap(t)d(t)$ is the area under the *rmap*(t) curve and, conversely, characterized the conformation stability of the peptide (higher values correspond to greater stability). For each peptide, ten MD trajectories with different initial velocities were calculated, on the basis of which the mean values of AUC_{rms} and AUC_{rmap} for a given peptide and the standard deviations of these values (σAUC_{rms} and σAUC_{rmap}, respectively) were calculated.

RESULTS AND DISCUSSION

Calculation of the conformational lability estimates s_1–s_5 for 135000 tetrapeptides, which were found in the PDB database (at a combinatorially possible number of tetrapeptides of 1.6×10^5) showed a significant correlation between these estimates (Table 1, Fig. 1). The strongest linear correlations were found between the estimates s_1 and s_3, s_2 and s_4. Pronounced power correlations were found between the estimates s_1 and s_2, s_3 and s_4 (i.e., between the mean value and its standard deviation), which are described by expressions of the form $y = A\sqrt{x}$. The existence of such correlations between the mean value (μ) and the standard deviation (σ) is evident as a result of asymptotic approximation of the hypergeometric distribution [16], at which $\sigma \approx \sqrt{\mu}$.

The above results of correlation analysis of the conformational lability estimates s_1–s_5, which were proposed in this study, showed that the estimates s_1, s_2, and s_5 were least correlated with each other (see Table 1) and, with certain reservations, can be considered independent. Accordingly, in further analysis, we primarily used these three estimates, as well as with the composite estimate s_{125} (see the Materials and Methods section), which was obtained on the basis of regression equations (examples of the latter are shown in Fig. 1).

The conformations of molecules (including proteins) are usually compared using a measure of similarity between the conformations of two fragments of protein structures, such as the root-mean-square deviation (RMSD) [17]. In the case of protein structures, the RMSD is usually calculated on the basis of the Cartesian coordinates of all atoms (except the hydrogen atoms) and/or the coordinates of only the main-chain atoms. At the same time, in the analysis of structures of protein molecules, the so-called "internal" or "natural" coordinates (i.e., the sets of lengths

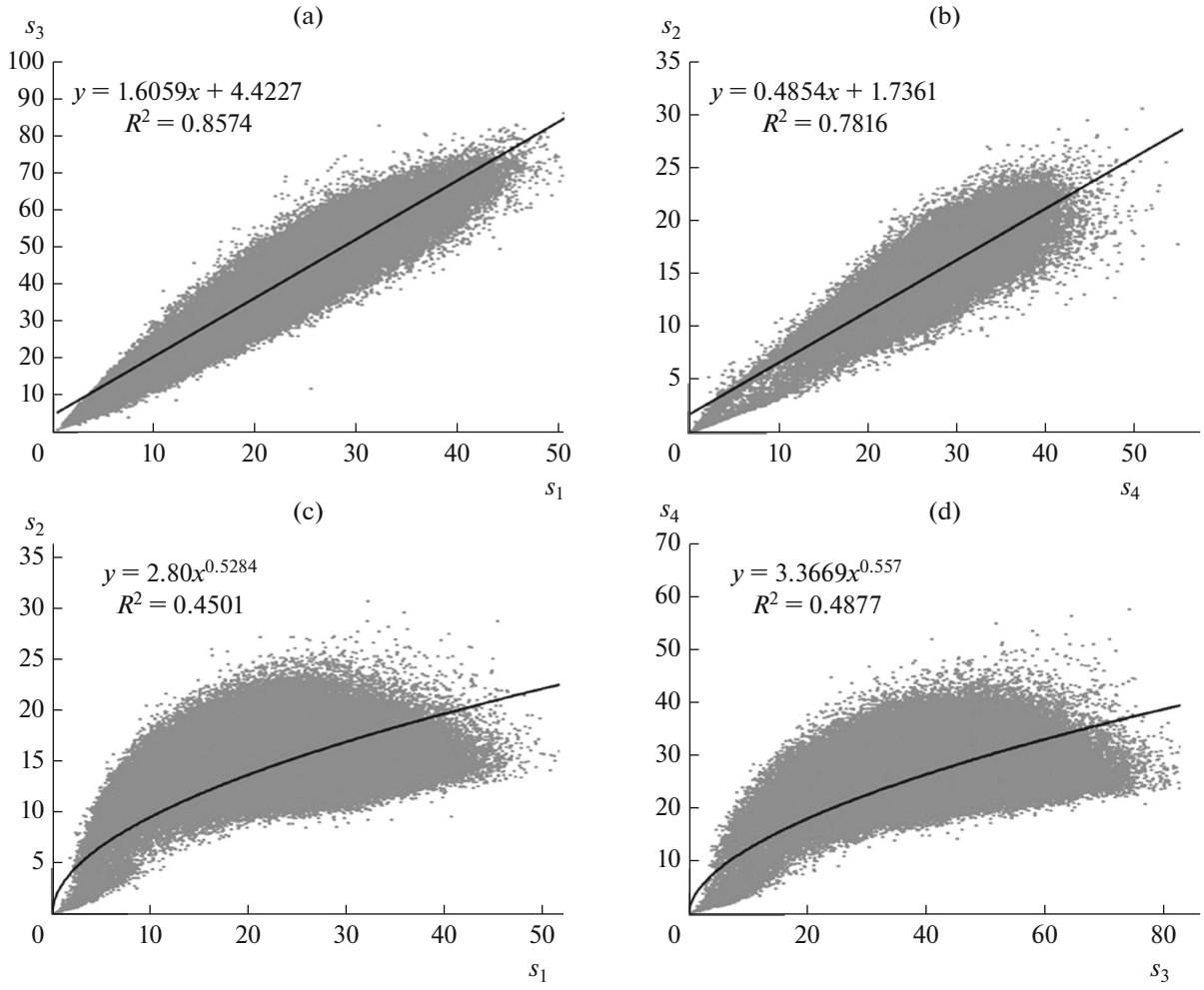

Fig. 1. Examples of correlations between the conformational lability estimates. Each point on the diagram corresponds to a tetrapeptide with a specific amino-acid sequence. The squared correlation coefficient (R^2) and the equations derived as a result of regression analysis are shown.

of bonds, as well as valence and dihedral angles, which are unambiguously converted to Cartesian coordinates, and vice versa) are also used [18].

In the case of protein molecules, one physically substantiated and widely used model of internal coordinates is the application of dihedral angles φ and ψ, based on whose values Ramachandran maps are constructed [19].

Given the high degree of planarity of the peptide bond plane (ω angle values are in a narrow range of $180° \pm 5°$), the use of only the set of angles φ and ψ as internal coordinates of a peptide fragment is an approximation that is quite acceptable for the purposes of this study. Thus, the conformational lability estimate s_1 proposed in this work is based on the calculation of the RMSD (however, with the use of the sets of angles φ and ψ rather than the sets of Cartesian coordinates). Other measures of conformational lability (s_2–s_4) are based on the standard statistical approaches [20] and are the mean values and their standard deviations (see the Materials and Methods section). The estimate s_5 was used in previous studies [8, 9] to assess the conformational lability/stability of *n*-peptides.

Thus, the correlation analysis indicated the existence of correlations between the conformational lability estimates s_1–s_5. This finding suggests that as a result of ordering the set of all tetrapeptides into a list in the ascending order of values of any of these estimates, the same peptides may reach the top of the list.

We consider, for example, 10% (13500) of the least conformationally labile peptides from the list ordered in accordance with estimate s_1, and 10% of the peptides from the list ordered in accordance with estimate s_2. At the intersection of the two sets, it was found that 3349 tetrapeptides (i.e., 2.09% of all tetrapeptides) were present in both lists, i.e., were "conformationally stable" according to two estimates at once, s_1 and s_2. Similarly, sets of peptides that are "conforma-

Table 2. Examples of the most conformationally stable (or least conformationally labile) tetrapeptide fragments

Peptide	n	s_1	s_2	s_3	s_4	s_5	S.S.
WEYC	5	21.067	22.34	31.86	35.24	0	Hhhh
MSWV	6	28.401	23.22	45.02	38.61	0	Hhhh
LYYC	7	11.726	10.34	14.26	15.48	0	Bbbb
MAVQ	7	1.696	0.518	4.985	1.562	0	Hhhh
VYYC	15	3.98	1.458	4.448	1.851	0.014	Bbbb
CLAM	7	30.756	22.4	50.72	34.79	0.036	Hhhh
VMTI	7	6.761	2.632	7.47	3.019	0.036	Bbbb
CFIT	6	6.886	2.056	6.968	1.676	0.042	Bbbb
EMMS	6	4.514	1.984	13.000	6.164	0.042	Hhhh
VYYC	35	3.98	1.458	4.448	1.851	0.014	Bbbb
FHWG	9	24.577	16.62	27.86	19.56	0.028	Bbbb
IVCN	12	18.433	15.74	22.49	19.77	0.042	Bbbb
TYYC	23	7.904	6.28	8.737	8.529	0.054	Bbbb
YIYV	14	7.411	2.40	8.44	2.67	0.054	Bbbb
DYYC	13	4.792	2.243	5.369	2.797	0.058	Bbbb
IMIT	12	14.908	11.14	18.53	15.67	0.063	Bbbb
MPTF	12	8.695	8.204	11.36	11.64	0.063	Bbbb

Peptides with the lowest values of the s_5 estimate, ordered by s_5, are shown; n is the number of occurrences of a given peptide fragment in the studied protein sample. S.S., secondary structure (h, α-helix, b, β-strand).

tionally stable" in accordance with other conformational lability estimates (s_3, s_4, and s_5) can be obtained.

At the intersection of such sets of tetrapeptides, which were obtained using each of the five estimates s_1–s_5, we obtained a list of 1034 tetrapeptides, each of which was among the 10% less conformationally labile peptides, as estimated using any of the s_1–s_5 estimates. Table 2 shows examples of some of these least conformationally labile (or most conformationally stable) tetrapeptide fragments obtained using the s_1–s_5 estimates.

The examples summarized in Table 2 show that conformationally stable (conformation-non-labile) peptides, in accordance with the conclusions made in the previous studies, may correspond to the α-helical conformations ("hhhh"), while some of them correspond to the β-strand conformations ("bbbb"). It was of interest to determine the degree to which each of these types of secondary structure is preferred for the "conformationally stable" peptides.

For this purpose, we calculated the empirical distribution functions for the conformational lability estimates s_1, s_5 and the composite estimate s_{125} for certain types of secondary structure (Figs. 2–4).

The analysis of the empirical distribution functions of the conformational lability estimates s_1, s_5, and s_{125} for certain types of the secondary structure showed that for all the studied estimates, the α-helical conformations were most stable. This difference between the empirical distribution functions for the α-helices and other types of the secondary structure was significant according to the Kolmogorov–Smirnov test.

As an example, the mean value of the s_1 estimate was $s_1 = 24 \pm 5$ for the α-helical conformations and $s_1 = 27 \pm 4$ for β-strands and unstructured fragments (coil), which is a statistically significant difference ($p < 0.001$). Similar results were obtained for the estimates s_5 (0.45 ± 0.06 for α-helix and 0.48 ± 0.05 for β-strands) and s_{125} (124 ± 12 for α-helix and 135 ± 14 for β-strands). In other words, the α-helical con-

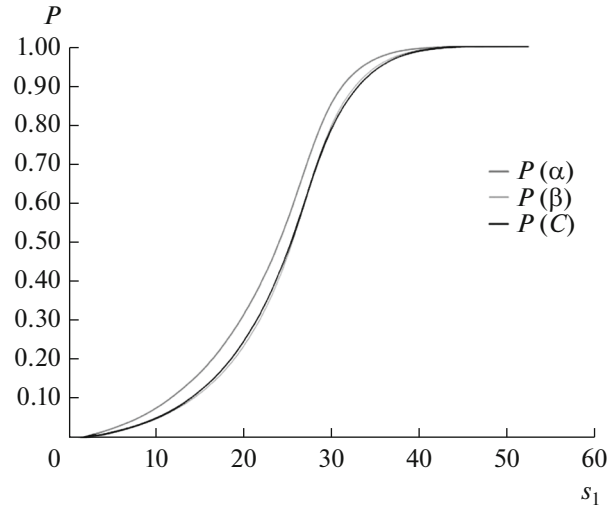

Fig. 2. An empirical function of the distribution of the conformational lability estimate s_1 values for certain secondary-structure types.

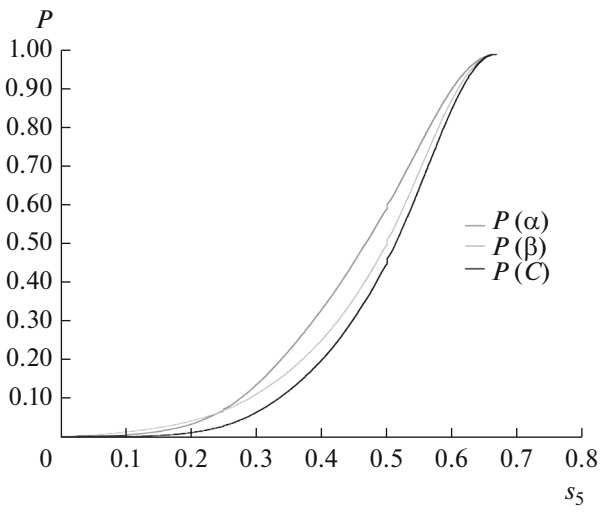

Fig. 3. An empirical function of the distribution of the conformational lability estimate s_5 values for certain secondary-structure types.

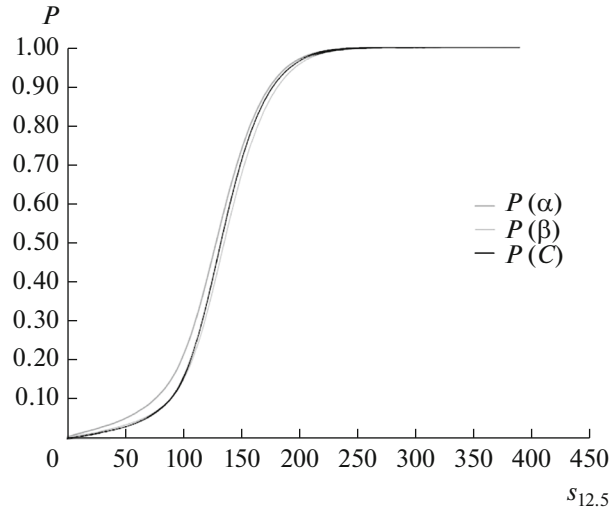

Fig. 4. An empirical function of the distribution of the composite conformational lability estimate s_5 values for certain secondary-structure types.

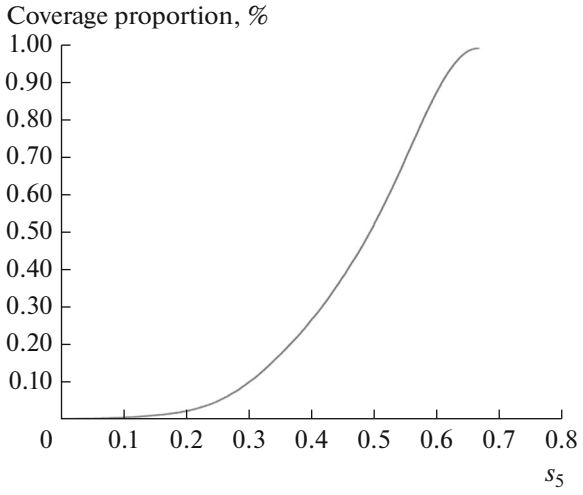

Fig. 5. The proportion of the composition of the amino-acid sequences of proteins by tetrapeptides with a given level of conformational lability s_5.

formation is less conformationally labile (i.e., more conformationally stable) than the other types of tetrapeptide conformations. In accordance with the s_1 estimate, the conformational lability of β-strands and the unstructured fragments is, on average, the same. The s_5 estimate, which is based on the analysis of the frequencies of occurrence of the secondary-structure types (see the Materials and Methods section) showed a somewhat lower conformational lability of β-strands compared to the unstructured fragments.

We analyzed the percentage of composition of the amino-acid sequences by the tetrapeptides with a given level of conformational lability according to the s_5 estimate (Fig. 5). Obviously, the tetrapeptides with a low conformational lability ($s_5 < 0.10$, which means that over 90% of the observed conformations belong to the same secondary-structure type) compose less than 1% (0.07%) of an arbitrary amino-acid sequence. The peptides with a somewhat larger conformational lability ($s_5 < 0.20$) composed 1.0% of an arbitrary amino-acid sequence and the peptides with $s_5 < 0.35$ already composed 11% of the sequence length. Thus, the sufficiently conformationally stable tetrapeptides (i.e., with $s_5 < 0.35$) may account, on average, for a significant proportion of an arbitrary amino-acid sequence.

It should be noted that the conformational certainty of a peptide fragment increases with segment length. It was found earlier that only 29 of 8000 (i.e., 0.36%) combinatorially possible tripeptide sequences have the preferred conformations [8]. Using similar "conformational stability" criteria, in this study we identified ~3300 of 160000 (2.1%) combinatorially possible tetrapeptide sequences as relatively stable. The increase in the proportion of the "conformationally stable" peptides with increasing length of the peptide fragment is consistent with the fact that the conformational certainty of a peptide fragment increases with an increase in the fragment length from three to four residues.

Interestingly, the so-called "chameleon peptides" (polypeptide chain segments which, at the same sequence, are components of the α helix in some proteins and β-strands in others) [21] can be regarded as an antipode of the conformationally stable peptides. In fact, a comparative analysis of the conformational lability estimates s_5–s_5 showed that the "chameleon peptides" are characterized by significantly higher values of the conformational lability estimates than all other peptides, including the "conformationally stable" ones ($s_5 < 0.35$, Table 3).

What is the physical nature of the existence of such "conformationally stable" peptides? It can be assumed that the differences in the conformational stability/lability of peptides (i.e., in the values of s_1–s_5 estimates) are due to the differences in the "flexibility" or mobility of the main-chain atoms, which, in turn, lead to differences in the conformational lability, which is calculated on the basis of the observed sets of peptide conformations.

To test this hypothesis, we performed MD simulations of individual peptides. Using the denaturing MD simulations, we calculated the AUC_{rms} and AUC_{rmap} parameters, which characterize the conformational lability/stability of peptides on the basis of MD trajectories. Using a number of examples, we showed that the more "conformationally stable" peptides have a higher conformational stability according to the denaturing MD simulations compared to the peptides with higher conformational lability estimates.

Table 3. Comparison of estimates of the conformational lability of the "chameleon peptides" from the CHSEQ database [21]

Set of peptide fragments	s_1	s_2	s_3	s_4	s_5
Chameleon peptides	25.67 ± 6.67	15.25 ± 2.54	46.34 ± 11.22	27.98 ± 4.52	0.47 ± 0.11
Non-chameleon peptides	21.91 ± 8.29	14.67 ± 3.97	38.88 ± 14.43	26.49 ± 7.26	0.45 ± 0.13
P	$<10^{-250}$	4×10^{-238}	$<10^{-250}$	$<10^{-250}$	3×10^{-219}
Conformationally stable ($s_5 < 0.35$)	20.46 ± 6.62	13.90 ± 3.75	39.14 ± 11.15	25.60 ± 6.73	0.29 ± 0.06
P	$<10^{-250}$	$<10^{-250}$	$<10^{-250}$	$<10^{-250}$	$<10^{-250}$

P, statistical significance according to the Kolmogorov–Smirnov test.

Table 4. Tetrapeptides for which molecular dynamics simulation was performed

Peptide	N	s_1	s_2	s_3	s_4	s_5	s_{125}	AUC_{rmsd}	σAUC_{rmsd}	AUC_{rmap}	σAUC_{rmap}
More "conformationally stable" tetrapeptides											
HEAA	25	19.59	14.90	29.07	20.48	0.26	47.6	0.11	0.02	0.096	0.017
MELI	21	20.00	12.86	31.87	21.13	0.34	49.7	0.13	0.04	0.098	0.019
NIQK	14	19.65	13.81	32.34	21.25	0.28	46.2	0.14	0.03	0.099	0.021
EAAV	126	15.54	14.34	24.52	22.89	0.33	48.7	0.13	0.02	0.092	0.024
Less "conformationally stable" tetrapeptides											
GGGG	137	50.64	12.85	89.57	22.44	0.55	101	0.21	0.06	0.093	0.021
EEAI	108	27.07	15.88	26.45	31.05	0.39	71	0.25	0.07	0.094	0.019
VVAV	126	24.04	15.91	36.16	30.18	0.40	69	0.92	0.19	0.085	0.022
AIKE	99	23.90	16.37	29.91	30.34	0.38	69	0.18	0.03	0.093	0.016

AUC_{rmsd} (Å ns) and AUC_{rmap} (ns) are the conformational lability measures calculated on the basis of MD trajectories of the peptides; σAUC_{rms} and σAUC_{rmap} are the standard deviations (see the Materials and Methods section). The more "conformationally stable" tetrapeptides were selected according to the criteria $s_1 \le 20$, $s_2 \le 15$, and $s_5 \le 0.35$.

For MD simulations, we selected examples of "stable" tetrapeptides on the basis of the results obtained using the cluster analysis of the conformations of 135000 tetrapeptides from the list of the combinatorially possible tetrapeptides. In this case, the criterion for selection of tetrapeptides was the proximity to the selected peptide to the "centers" of the obtained clusters rather than the greatest possible frequency of occurrence. Additional criteria for selecting the "conformationally stable" peptides were the following threshold values of the conformational lability estimates: $s_1 \le 20$, $s_2 \le 15$, and $s_5 \le 0.35$. These criteria are, of course, purely empirical and were selected on the basis of expert analysis of the curves shown in Figs. 2–5. As a result, we selected four representative "conformationally stable" tetrapeptides, which represented four respective clusters: HEAA, MELI, NIQK, and EAAV tetrapeptides.

The GGGG, EEAI, AIKE, and VVAV peptides were used as control "conformationally labile" peptides. It should be noted that the EEAI and AIKE peptides are fragments of the "chameleon peptide" AKEEAIKE, which is represented in the ShSeq database [21]. The AKEEAIKE peptide takes the α-helical conformation in one set of structures (e.g., 2jws, 2kdl, 2lhs, and 2lhg) and the β-strand conformation in other structures (e.g., 2lhd, 2kdm, 2jwu, and 2lhe). The results of MD simulation of the denaturation of these peptides are summarized in Table 4. Table 5 shows the results of analysis of the statistical differences between the investigated conformational lability measures. Examples of individual *rmsd*(*t*) and *rmap*(*t*) curves are shown in Figs. 6–9.

MD simulations of all peptides were performed for 100 ps under mild denaturing conditions (300 K in vacuo, the model without solvent, that is, without specifying the coordinates of water molecules and without using the corresponding intermolecular potential term, which stabilizes the peptide molecule). For comparative assessment of the conformational stability, we calculated the standard deviations for all atoms of the main chain and the correlation coefficient between the values of pairwise distance matrix elements of the initial structure and the structure at a given simulation time (1 ps, 2 ps, ..., 100 ps).

The analysis of the root-mean-square deviations of the main-chain atom coordinates showed (Fig. 6) that the conformationally-stable HEAA, MELI, and NIQK peptides by a simulation time of ~100 ps entered the corresponding "pools" of the stable con-

formations (RMSD ~1.2 Å), whereas the "unstable" peptides were characterized by significantly higher RMSD values (2.0...2.5 Å for GGGG, 1.8...2.0 Å for EEAI, and 1.5...1.7 Å for AIKE). We note that the *rmsd*(*t*) curves for EAAV and VVAV peptides are not shown in Fig. 6 to improve the clarity of illustration (the *rmsd* values for these peptides are much higher).

The coefficient of correlation between the values of pairwise distance matrix elements characterizes the structural similarity at a given time with the initial starting peptide conformation (with an accuracy of the translation and rotation). The calculations of MD trajectories showed that the structures of the conformationally stable HEAA, MELI, and NIQK peptides were characterized by high values of correlation coefficients along the trajectory (0.96...1.00), which is indicative of their MD-stability. However, the conformationally unstable GGGG peptide was characterized by drastic conformational changes (which was reflected in variations of the correlation coefficient over a much wider range of values, from 0.88 to 0.98) throughout the simulation. After a 25-ps simulation, the "unstable" EEAI and AIKE peptides can be characterized by significantly lower correlation coefficient values (0.94) than the "stable" peptides (see Fig. 7).

In general, the more "conformationally stable" HEAA, MELI, NIQK, and EAAV tetrapeptides differed from the less "conformationally stable" GGGG, EEAI, AIKE, and VVAV tetrapeptides in smaller values of the conformational lability estimates (including the AUC_{rmsd} parameter, which was obtained by denaturing MD (see Table 5)). The differences in the values of the estimates were statistically significant, except for the s_3 estimate. The AUC_{rmap} parameter (which, conversely, estimates the conformational stability of peptides) was significantly higher in the more "conformationally stable" tetrapeptides.

Table 5. Differences in the mean values of the conformational lability/stability estimates between the group of four more "conformationally stable" ($s_1 \le 20$, $s_2 \le 15$, and $s_5 \le 0.35$) and the group of four less "conformationally stable" tetrapeptides

Estimate	More "conformationally stable" tetrapeptides	Less "conformationally stable" tetrapeptides	*P*
s_1	18.70 ± 2.11	31.41 ± 12.90	0.011
s_2	13.98 ± 0.87	15.25 ± 1.62	0.015
s_3	29.45 ± 3.59	45.52 ± 29.64	>0.1
s_4	21.44 ± 1.03	28.50 ± 4.06	0.018
s_5	0.30 ± 0.04	0.43 ± 0.08	0.021
s_{125}	48.12 ± 1.52	77.88 ± 15.71	0.016
AUC_{rmsd}	0.13 ± 0.03	0.39 ± 0.35	0.012
AUC_{rmap}	0.10 ± 0.02	0.09 ± 0.02	0.053

P, statistical significance according to the Kolmogorov–Smirnov test.

It was of interest to assess the degree to which the conformational lability estimates that were obtained on the basis of analysis of the sets of observed conformations of peptides correlated with conformational lability/stability estimates that were calculated on the

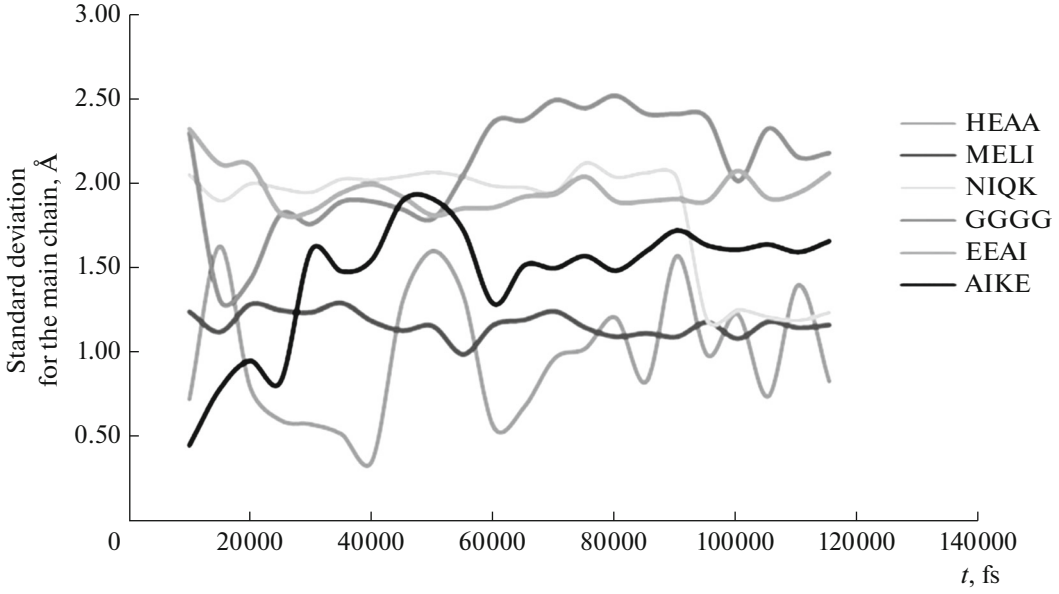

Fig. 6. The dependence of the roof-mean-square deviation of the coordinates of atoms of the polypeptide chain backbone on the simulation time for the "conformationally stable" peptides HEAA, MELI, and NIQK and the "unstable" peptides GGGG, EEAI, and AIKE.

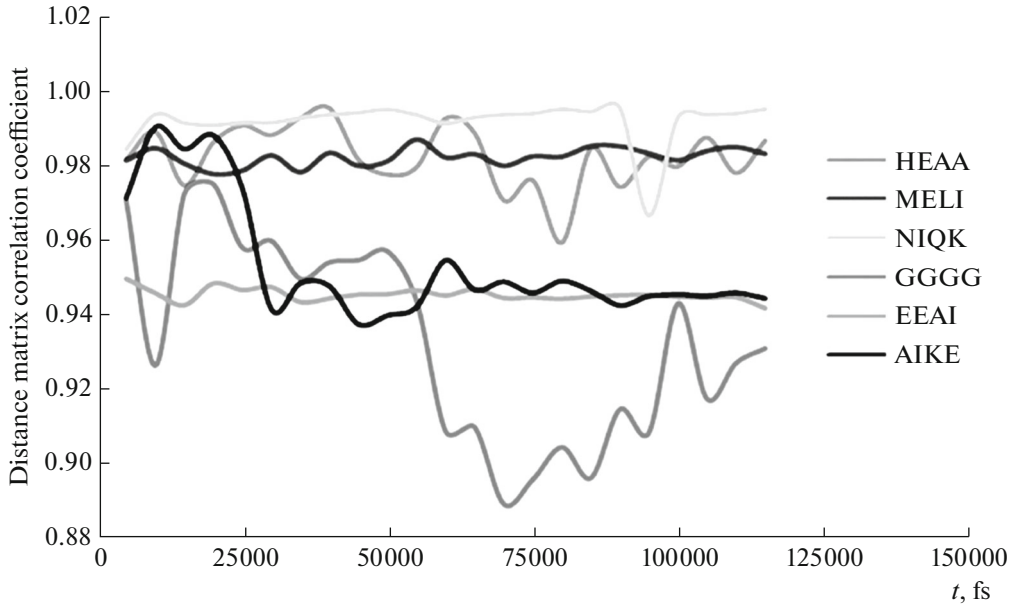

Fig. 7. The dependence of the correlation coefficient for distance matrices on the simulation time for the conformationally stable peptides HEAA, MELI, and NIQK and the "unstable" peptides GGGG, EEAI, and AIKE.

basis of MD trajectories. The results of the correlation analysis are summarized in Table 6.

The data in Table 6 show that, as expected, the correlations between the s_1–s_5 estimates and the conformational lability estimate AUC_{rms} were directly proportional, whereas correlations between the s_1–s_5 estimates and the conformational stability estimate AUC_{rmap} were inversely proportional. In general, the correlation coefficient values were fairly low (which is quite natural, because the conformational lability estimates on the basis of analysis of structures and on MD trajectories are based on different principles).

The strongest correlations were found for the estimates obtained as a result of MD simulations and estimates s_2 and s_4. We recall that the "variance" estimates

Table 6. Correlation analysis of the conformational lability estimates s_1–s_{125} and the AUC_{rmsd} and AUC_{rmap} (ns) estimates calculated on the basis of MD trajectories of peptides

Estimate	AUC_{rmsd}	AUC_{rmap}
s_1	0.07	–0.16
s_2	0.42	–0.51
s_3	0.03	–0.11
s_4	0.56	–0.62
s_5	0.27	–0.43
s_{125}	0.27	–0.42

The correlation coefficient values are shown. The negative values correspond to the inverse correlations (i.e., $k < 0$ in the regression formula $kx + b$).

s_2 and s_4 characterize the standard deviation of the distance between the elements of the set C(P) and the standard deviation of the angle between the elements of the set C(P), respectively. In other words, the results of MD simulations are most strongly correlated with the estimates s_2 and s_4, which characterize the scatter of distances and angles, respectively, rather than with the estimates of the mean distances (s_1(P)) or mean angles (s_2(P)) between the elements of C(P) (i.e., the conformations observed in the PDB database). Examples of the discussed correlations are shown in Fig. 8.

Thus, the combined analysis of the data obtained by denaturing molecular dynamics and the conformational lability estimates s_1–s_5 indicates the tendency of the conformationally stable peptides to "retain" a certain "stable" conformation. During the entire period of MD simulation, the conformations of the tested "conformationally stable" peptides did not undergo drastic changes and were characterized by high correlation coefficient values compared to the initial conformation.

Molecular mechanical simulation of the structures of real peptides to assess their conformational stability. Using the data on the tetrapeptide conformations, the developed conformational lability estimates s_1–s_5, and molecular dynamics simulations, we studied examples of known soluble peptides to assess their conformational stability. For the simulated peptides, we calculated the conformational lability estimates as the mean values of the s_1–s_5 estimates of the tetrapeptides present in each of the peptides. The results are summarized in Table 7.

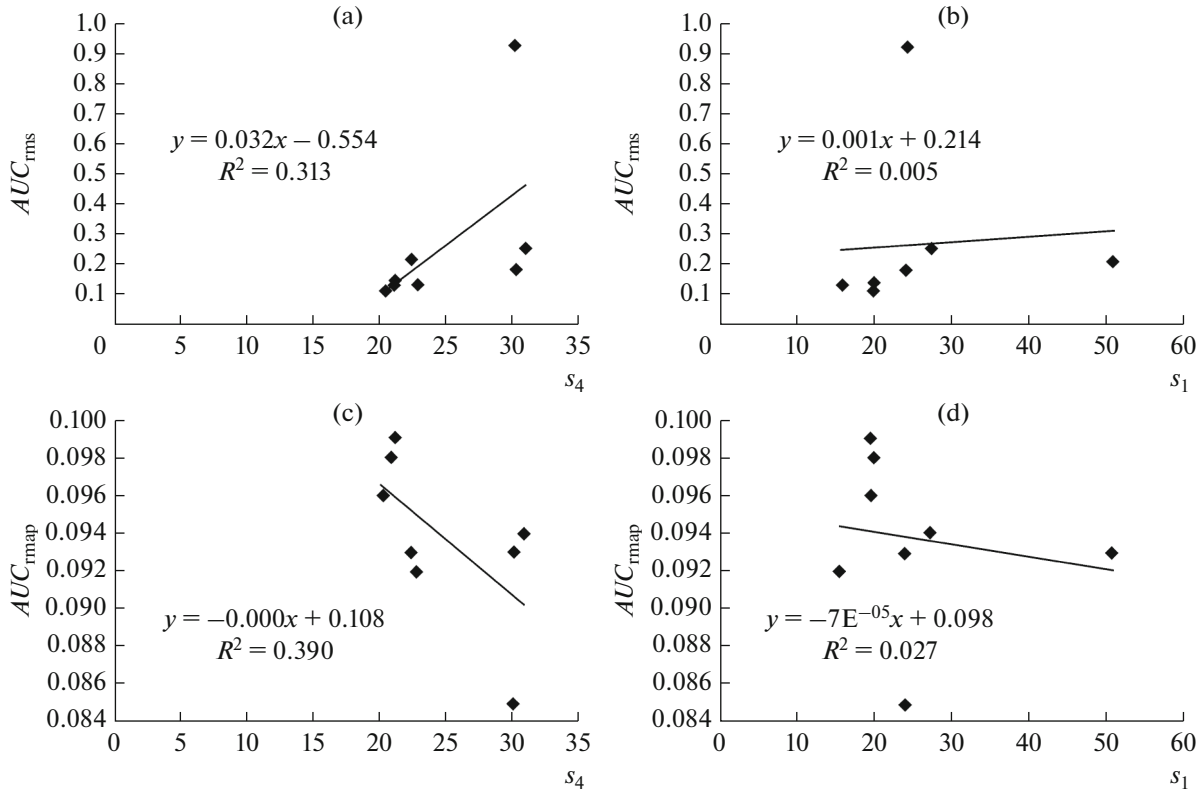

Fig. 8. Examples of correlations between the conformational lability estimates obtained on the basis of analysis of the sets of observed tetrapeptide conformations and the conformational lability/stability estimates calculated on the basis of the MD trajectories for eight peptides.

The MD simulations were performed using the denaturing MD procedure described earlier, at T = 350 K and a trajectory length of 1 ns. Since the spatial structures of the test peptides are not known, each of the peptides was simulated proceeding from two starting conformations (α-helix and β-strand). The examples of *rmsd*(*t*) trajectories are shown in Fig. 9.

Correlation analysis for the conformational lability estimates calculated on the basis of the tetrapeptide composition and the estimates obtained as a result of the MD simulations of five peptides in two starting conformations (Table 8) was performed. According to the results of the analysis, ds_1, ds_2, ds_3, and ds_5 estimates (i.e., the standard deviations of s_1, s_2, s_3, and s_5 estimates) showed significant correlations with the values of standard deviations of MD trajectories, σAUC_{rmsd}, regardless of the starting conformation type.

Thus, the calculations of the tetrapeptide composition and MD simulations indicate that RSWFETWV, SFEDFWK, and RLSKEEI peptides are the most conformationally stable. These results make it possible to plan respective experiments to analyze the conformations of these peptides.

Table 7. Soluble peptides and the predicted values of the conformational lability estimates

Peptide	s_1	ds_1	s_2	ds_2	s_3	ds_3	s_4	ds_4	s_5	ds_5
RSWFETWV	26.2	10.9	11.2	5.0	48.0	15.6	22.0	7.1	0.49	0.13
SFEDFWK	22.8	7.7	14.6	2.4	43.0	12.5	26.3	3.2	0.47	0.11
RLSKEEI	22.7	4.0	14.4	1.4	45.7	6.6	28.1	1.4	0.42	0.07
LSLGLETAGG	32.4	6.7	14.0	1.5	60.5	9.0	25.3	2.9	0.52	0.08
NHRWLGGM	36.7	4.3	13.5	1.4	66.4	8.5	23.2	1.6	0.58	0.05

The estimates in columns s_1–s_5 were calculated as the mean value of s_1–s_5 estimates of the tetrapeptides comprising the peptide. The ds_1–ds_5 values are the standard deviations of s_1–s_5 estimates.

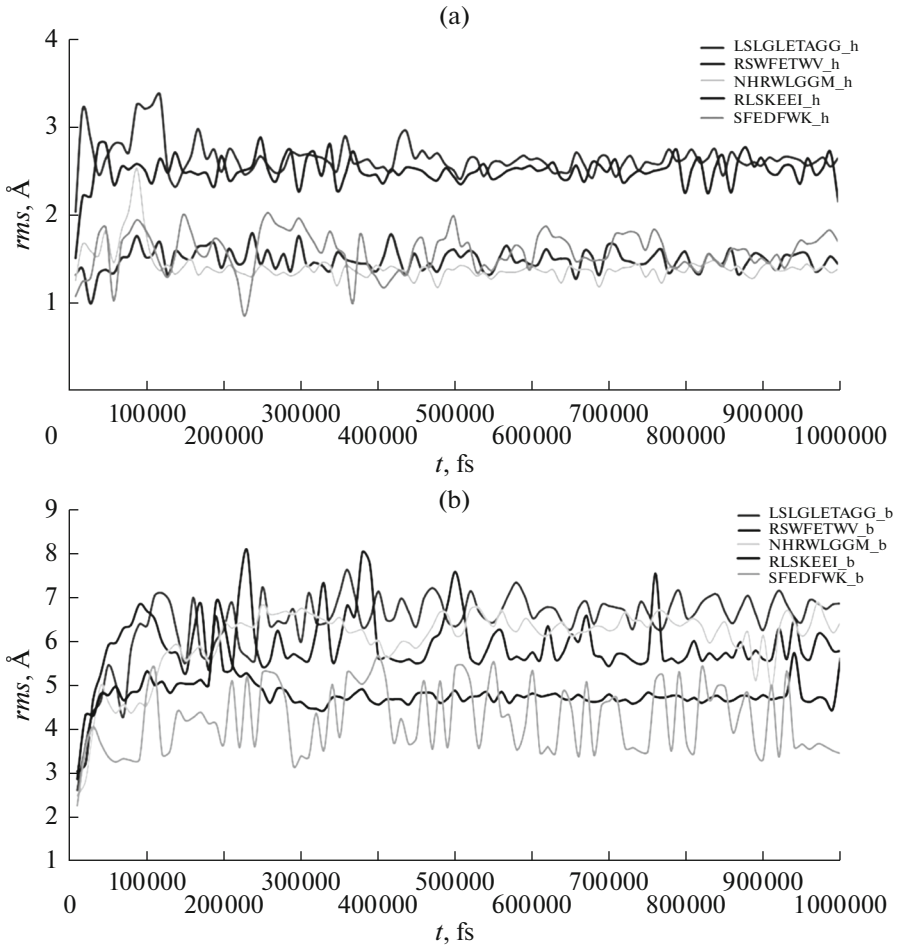

Fig. 9. Molecular dynamics trajectories for the soluble peptides used as examples: (a) initial conformation is α-helix; (b) initial conformation is β-strand.

CONCLUSIONS

In this study, we proposed and tested several methods for assessing the conformational stability/lability of peptide fragments that compose protein structures. The proposed formulas for estimating the conformational lability of the peptide fragments by statistical analysis of structures allowed us to show that the tetrapeptide fragments significantly differ in the conformational stability/lability estimates. As a result, the subgroups of more "conformationally stable" peptides were distinguished. The latter have a predominantly α-helical conformation. A database (list) of 1034 tetrapeptides that are characterized by the lowest confor-

Table 8. Results of the correlation analysis between the conformational lability estimates calculated on the basis of the tetrapeptide composition and the estimates obtained as a result of MD simulation of the peptides in two starting conformations

Estimate	Initial conformation: β-strand		Initial conformation: α-helix	
	AUC_{rmsd}	σAUC_{rmsd}	AUC_{rmsd}	σAUC_{rmsd}
ds_1	0.10	0.72	–0.37	0.23
ds_2	0.07	0.62	–0.50	0.10
ds_3	–0.02	0.81	–0.61	0.27
ds_5	–0.22	0.85	–0.29	0.51

mational lability according to all proposed estimates was formed. As a result of calculations performed using denaturing MD simulations, qualitative and quantitative estimates of the conformational lability or "flexibility" of some tetrapeptides were obtained, which are consistent with the estimates obtained by analyzing the conformation sets of peptides observed in the experimentally determined protein structures. The peptides that are characterized by a sufficiently high degree of conformational stability constitute a substantial proportion of the combinatorially possible peptides (tetrapeptides). They may play a major role in protein structure formation, because they account for approximately 10% of the amino-acid sequence. Using real soluble peptides as examples, we demonstrated the possibility of assessing the conformational stability of an arbitrary amino-acid sequence on the basis of the obtained criteria of the conformational lability of tetrapeptides.

FUNDING

This study was supported by the Russian Foundation for Basic Research (project nos. 16-54-00219-Bel and 18-54-00037-Bel) and the Belarusian National Foundation for Basic Research (project no. B18R-268).

COMPLIANCE WITH ETHICAL STANDARDS

The authors declare that they have no conflict of interest. This article does not contain any studies involving animals or human participants performed by any of the authors.

REFERENCES

1. N. V. Kalmankar, C. Ramakrishnan, and P. Balaram, Proteins **82**, 1101 (2014).
2. A. Figureau, M. A. Soto, and J. Toha, Protein Eng. **16** (2), 103 (2003).
3. P. K. Vlasov, A. V. Vlasova, V. G. Tumanyan, and N. G. Esipova, Proteins **61** (4), 763 (2005).
4. C. Bystroff, K. T. Simons, K. F. Han, and D. Baker, Curr. Opin. Biotechnol. **7** (4), 417 (1996).
5. C. Bystroff and D. Baker, J. Mol. Biol. **281**, 565 (1998).
6. C. G. Hunter and S. Subramaniam, Proteins **50** (4), 572 (2003).
7. O. Sunder, I. Sommer, and T. Lengauer, BMC Bioinform. **7**, 14 (2006).
8. A. V. Batyanovskii and P. K. Vlasov, Biophysics (Moscow) **53** (4), 264 (2008).
9. A. V. Batyanovskii, I. D. Volotovsky, V. A. Namiot, et al., Biophysics (Moscow) **60** (3), 348 (2015).
10. H. M. Berman, J. Westbrook, Z. Feng, et al., Nucleic Acids Res. **8**, 235 (2000).
11. J. F. Gibrat, T. Madej, and S. H. Bryant, Curr. Opin. Struct. Biol. **6**, 377 (1996).
12. I. Yu. Torshin, N. G. Esipova, and V. G. Tumanyan, J. Biomol. Struct. Dyn. **32** (2), 198 (2014).
13. I. Y. Torshin, Sci. World J. **4**, 228 (2004).
14. A. R. Rappé and W. A. Goddard, J. Phys. Chem. **95** (8), 3358 (1991).
15. M. A. Cuendet and van W. F. Gunsteren, J. Chem. Phys. **127** (18), 184102 (2007).
16. K. K. Talukdar and W. D. Lawing, J. Acoust. Soc. Am. **89** (3), 1193 (1991).
17. *A Dictionary of Physics*, 6th ed. (Oxford Univ. Press, Oxford, 2009).
18. J. Parsons, J. B. Holmes, J. M. Rojas, et al., J. Comput. Chem. **26**, 1063 (2005).
19. G. N. Ramachandran, C. Ramakrishnan, and V. Sasisekharan, J. Mol. Biol. **7**, 95 (1963).
20. *Quadratic Deviation (Encyclopedia of Mathematics),* Ed. by M. Hazewinkel (Springer, 2001).
21. L. Wenlin, L. Kinch, A. Karplus, and N. Grishin, Protein Sci. **24**, 1075 (2015).

Translated by M. Batrukova