



Quantifying Steric and Hydrophobic Influence of Non-Standard Amino Acids in Proteins That Undergo Post-Translational Modifications

Luiz F. O. Rocha

Department of Physics and Chemistry, Faculty of Pharmaceutical Sciences of Ribeirão Preto, University of São Paulo, Ribeirão Preto, São Paulo, Brazil

Email address:

luiz@fcfrp.usp.br

To cite this article:

Luiz F. O. Rocha. Quantifying Steric and Hydrophobic Influence of Non-Standard Amino Acids in Proteins That Undergo Post-Translational Modifications. *Biochemistry and Molecular Biology*. Vol. 2, No. 2, 2017, pp. 12-24. doi: 10.11648/j.bmb.20170202.11

Received: February 18, 2017; **Accepted:** March 1, 2017; **Published:** March 23, 2017

Abstract: Non-standard amino acids in protein post-translational modifications aid in a wide variety of biological functions and processes, furnishing expansion from the genome to the proteome. First, from structural examinations in unmodified proteins with only standard amino acids, this work empirically obtains numeric relations that reveal how instruction transfers occur between native-state structures. Next, from these relations, the influence of non-standard amino acids inside post-translationally modified proteins is quantified by successfully predicting the contents of large and hydrophobic residues in helices and β -strands for 210 inspections performed. This suggests a twofold molecular mechanism by the fundamental biophysicochemical properties (residue volume and hydrophobicity), and concludes that the utilized non-standard amino acids have limited global influence at the residue level. Our prediction method provides a better underlying understanding of molecular interactions and mechanisms, and is particularly promising in terms of surveying further modified proteins.

Keywords: Protein Physicochemical Property, Protein Synthesis, Proteome Diversification, Residue Content Prediction, Translation

1. Introduction

After biosynthesis from genetic instructions, many peptide and protein species undergo post-translational modifications on the endoplasmic reticulum in order to exert their specialized biological purposes, control varied cellular processes, modulate chemical reactions, and interact with other molecules. These modifications may aid to alter chemical properties, folding events, macromolecular stabilities, activity states, and subcellular locations [1–2] of peptides and proteins. Therefore, the proteome of a living organism is two to three orders of magnitude more diversified than its encrypting genome [3], thus making such modifications a relevant topic for computational biochemistry, molecular biophysics and molecular-cellular biology of macromolecules. There are several post-translational modifications, including those made through additions of non-standard amino acids to one or more standard amino acid residues to redirect the protein chain in the proper direction and, thus influencing or assisting its

charge, hydrophobicity, conformation, stability and function [4].

Among the additions of non-standard amino acids, the phosphorylation (insertion of a phosphate group to an amino acid side chain) [5], acetylation (and amination) by inclusion of a small acetyl radical ACE (and an amino univalent radical NH_2) at the extremities of the primary sequence, and hydroxylation of proline and lysine are very common. Some native proteins greatly increase their resistance against conformational degradation by acetylation/amination of the extreme amino/carboxy positions, since many proteases—specially the aminopeptidases and carboxypeptidases—require either an amino or a carboxy terminal to selectively act [6–7]. Acetylation and amination may also be used to eliminate the influence of charged groups at terminal portions [8–9]. The post-translational succinylation (or insertion of a group succinyl (SIN) to ends of a chain) in the fragmentation of biopolymers can both solubilize the peptide derivatives and block the action of enzymes [10], as well as inhibiting hydrolysis [11].

In modified proteins with the presence of covalent post-translational alterations, selective uses of non-standard compounds or amino acids—such as uncommon residues, cofactors, and prosthetic groups—furnish proteome expansion and diversification [3]. Non-standard compounds usually have specific aims, as shown by the following cases: the alpha-aminobutyric acid (ABA), a parent of the alanine, is used for selective replacement of half-cystines [12–13]; 3-amino-alanine (DNP) may alter the affinity of the amino acids involved in potassium channel binding [14]; D-proline (DPR) can nucleate β -turns of appropriate stereochemistry and type II conformations [15]; alloseucine (ILL), a stereoisomer of isoleucine, orients the side-chain into a trans χ_1 dihedral angle [16]; norleucine (NLE) sometimes produces only local perturbations reflecting in an increased folding rate [17]; and the pyroglutamic acid (PCA), a parent of the glutamic acid, in the N-terminal tail plays useful functional roles, and reduces the susceptibility to aminopeptidases [18–19]. In addition to the above-mentioned post-translational alterations, several other chemical compounds are added to protein chains [20].

With the main purpose of quantifying the global steric and hydrophobic influence and reaching a better underlying understanding of non-standard amino acids in modified proteins, this study makes use of: (i) an empirical approach taking data directly from experimentally derived proteins; (ii) a quantitative formulation for the proteins selected through numerical relations and prediction rules from the relationship between primary and secondary structures; (iii) validations of the utilized relations and rules; and (iv) computer algorithms to facilitate and automatically operate data processing for the items (i)–(iii). The remainder of this study is arranged as follows. In the Materials and Methods section, we establish the benchmark dataset, binary codes for amino acids, and the accuracy scale for residue content predictions. In the Results and Discussion section, we obtain numerical rules from unmodified proteins, and use these rules for inspecting mechanisms and influence of non-standard residues in two target subgroups of post-translationally modified proteins. This study is concluded in the Conclusions and Future Developments section.

2. Materials and Methods

2.1. The Benchmark Dataset

The proteins for our benchmark dataset are carefully selected under the following conditions: in a specific extension with equal number of beads or residues, the modified and unmodified (only with standard or L- α -amino acids) exemplars should exist in a quantity greater than or approximately to two tens each (un)modified exemplar; and in each utilized extension, the chosen exemplars must be non-redundant with either low-similarity residue sequences (less than 25% identity) or with differences in helices or strands of at least four residues. Taking the above restrictive conditions, and among several sequence extensions that were

extensively investigated, we opt for a systematic analysis of 35-residue modified (target subgroup I) and unmodified (template group) proteins deposited in the Protein Data Base (PDB) [21]. The target subgroups I and II with modified proteins from 35 to 40 residues and their non-standard amino acids will be displayed in the Table 1 of Supplementary Data Appendices.

Proteins with 35 residues have diversified residue dispositions, varied structures and functions, and are in many biological sources. The 35-residue unmodified proteins consist of 39 exemplars, while the modified ones include 16 exemplars and residue sequences with at least one of the nine non-standard chemical compounds (ABA, ACE, DNP, DPR, IIL, NH2, NLE, PCA, and SIN), whose skillful roles [6–19] were previously described in the Section 1. Structures of proteins with equal number N of residues are usually compared by their compactness utilizing the radius of gyration R_G defined as:

$$R_G^2 = (2/(N(N-1))) \sum_{k < l} \sum_l r_{k,l}^2 \quad (1)$$

where $r_{k,l}$ is the Euclidean distance between mass centers of residues “k” and “l” from the atomic coordinates laid in the PDB library.

2.2. Amino Acid Types, Primary and Secondary Structures, and Prediction Accuracy

Protein polymers employ a rich repertoire of covalently-linked residues (non-standard and 20 standard amino acids) that cover a wide range of shapes, sizes in many atomic and molecular interactions. This residue diversity consequently determines the broad variety of biophysicochemical properties that are fundamental in ascertaining macromolecular structures and functional activities [22–23]. Among such properties, the volume and hydrophobicity have been considered as two primary components [24–25]. The residue volumes (steric contributions) and hydrophobicities (hydrophobic effect and interactions) are dominant throughout the selection and maintenance of three-dimensional folded configurations under varied physiological conditions and environmental contexts [26–27].

The amino acids are denoted only by their volumes or sizes [28–29] and hydrophobicities [30], via coarse-grained binary codes, large-small (LS) and hydrophobic-polar (HP), respectively. The standard amino acids and by association the non-standard ones, are large-hydrophobic (F, H, I, L, M, V, W, Y; IIL, NLE), large-polar (E, K, Q, R), small-hydrophobic (A, C, P, T; ABA, DPR, PCA), and small-polar (D, G, N, S; ACE, DNP, NH2, SIN). These reduced codes, LS and HP, compress the contained information in biomolecular structures without great loss of direction, capturing many of their essential and basic features [22]. The codification LS expresses the steric constraints and packing efficiency [23–24]; the label HP embodies the intra-chain and medium-chain interactions [25, 30–31]. In the 55 analyzed 35-residue proteins, the standard amino acids are much more frequent than the specific non-standard compounds, having 1899 and

26 residues, respectively. Among the 20 standard amino acids, large and hydrophobic sub-components dominate, both independently having 12 units and, in consequence, these sub-components are taken into account for the outputs shown below.

A one-dimensional residue sequence can be suitably represented by its total number of large (n_L) and hydrophobic (n_H) residues in the primary structure of native chain. Each n_i may be associated with no, one or many proteins in the dataset, whose the subscript character “i” stands for the large or bulky (L) and hydrophobic or apolar (H) residues in the primary and secondary levels.

Omnipresent steric and hydrophobic interactions represented by binary codes (large and hydrophobic residues) are suitable key tools to observe the global biophysicochemical influence of the non-standard amino acids in two periodic structural motifs (helices, and β -sheets formed by strands). Other resolution levels (e.g., more letter codes or atomic approaches) should be utilized for sharper measurements of non-standard amino acids. Furthermore, we only evaluate overall lengths of motifs more than five residues ($L_j > 5$) to provide additional security to our comparative structural studies, where the character “j” accounts for the helices (h) and strands (e). For unmodified proteins of the template group, the actual contents (designated by $t_{i,j}$) of large and hydrophobic sub-components in secondary motifs give rise to percentage fractions $p_{i,j}$ given by:

$$p_{i,j} = (t_{i,j}/L_j)100 \quad (2)$$

where $p_{i,j}$ varies from 0 (when L_j does not own large and hydrophobic residues, $t_{i,j}=0$) to 100% (in case of L_j uniquely owned by these residues, $t_{i,j}=L_j$).

In the modified proteins of the target subgroups I and II, the efficiency of the predictions for the estimated contents $t_{i,j}$ of large and hydrophobic residues in periodic motifs is measured by observing the dissimilarity with the actual values through:

$$\Delta t_{i,j} = |\text{actual } t_{i,j} - \text{estimated } t_{i,j}| \quad (3)$$

where $\Delta t_{i,j}$ is given in absolute value, at residue level, and can range from zero (at best) to L_j (at worst). More specifically, the prediction accuracy is considered excellent (when $\Delta t_{i,j} \leq 0.5$, that is $\Delta t_{i,j} \approx 0.0$), good ($0.5 < \Delta t_{i,j} \leq 1.5$, $\Delta t_{i,j} \approx 1.0$), acceptable ($\Delta t_{i,j} \approx 2.0$ or 3.0 provided that $\Delta t_{i,j} - 0.1L_j \leq 1.0$, where $0.1L_j$ is 10% of L_j), and bad ($\Delta t_{i,j} \geq 2.0$ as long as $\Delta t_{i,j} - 0.1L_j > 1.0$).

3. Results and Discussion

3.1. Prelusive Measures in the Template Group and Target Subgroup I

Modified and unmodified 35-residue proteins are amply diversified structurally (Figure 1a–c), as seen by their varied compactness (R_G (1)), and overall lengths of 3_{10} -, α - and π -helices (L_h), and strands (L_e) inside parallel and antiparallel

β -sheets. In the inset of Figure 1c, a fraction $p_{L,e}$ (2) of large sub-components in strands is displayed, whereas $L_e > 5$. Total length $L_j \leq 5$ has few stabilizing interactions, and structural variations that are often aggravated by conformational changes [33], helical distortions [34] and configurational instability [35].

The functional native conformations have distributed compactness degree from very compact ($R_G \leq 9.0$ Å, Figure 1a) to rather tight, to the swollen or less densely packed ($R_G > 11.0$ Å); and whose cutoff threshold values of 9.0 and 11.0 Å are strategically used for more precise linear adjustments in the following subsection. The modified and unmodified protein conformations are proportionally compatible, as in R_G and L_h (Figure 1a–b), since they spread out in almost all of their possible values, as in L_e (Figure 1c) usually with $L_e \leq 15$. Unmodified proteins do not undergo the delicate and complex array of post-translational modifications, and have greater quantity than those modified ones with 39 and 16 exemplars, respectively. Because of this, the unmodified proteins are examined first.

3.2. Relationship Between Primary and Secondary Structures in the Template Group

According to the compactness and cutoff values (9.0 and 11.0 Å) of R_G (1), the 39 unmodified protein chains are categorized into 11 highly compact, 10 reasonably tight, and 18 less packed chains. For every chain with $L_j > 5$, one separately computes the extent to which the conformational compactness degree is related to contributions of the large and hydrophobic residues from the primary sequence (n_i) to its secondary structural elements ($p_{i,j}$)—although, $p_{i,j}$ (2) does not seem to be connected with n_i . The amounts of $p_{i,j}$ in relation to n_i (Figure 2) are plotted for 29 data points of helices and 11 of strands in their respective graphs, totalizing 80 structural inspections, and adjusted by linear fits, whose general equations are expressed as:

$$p_{i,j} = mn_i + b, \text{ and } R \quad (4)$$

where m , b and R are the slope, intercept, and linear correlation coefficient.

The 80 data points of 40 unmodified samples are sufficiently scattered and give rise to their corresponding linear regressions ((5)–(9) in Figure 2), expressing the noticeable efficacy of the sub-components (large and hydrophobic) of the residues in functional conformations. The fractions $p_{i,j}$ are slightly dependent on the residue sub-component types, considering that the hydrophobic residues (7)–(8) have more sloped regression lines than those large ones (5)–(6), as expressed by their greater slopes, m (and negative intercepts, b), other than at $p_{H,e}$ (9) of less closely packed configurations. Most unmodified samples have both points around the straight lines, indicating an efficient and simultaneous use of both sub-components from primary to secondary structures by means of a double effective molecular mechanism.

Some special samples possess a residue sub-component

that is more selective and compensatory than the other by a single effective mechanism. For instance, the once underlined chains 1PXQ and 1ROO are farthest from the linear fit for

$p_{L,h}$ (5), but at the same time they are near to $p_{H,h}$ (7), as designated by arrows, and therefore using a single

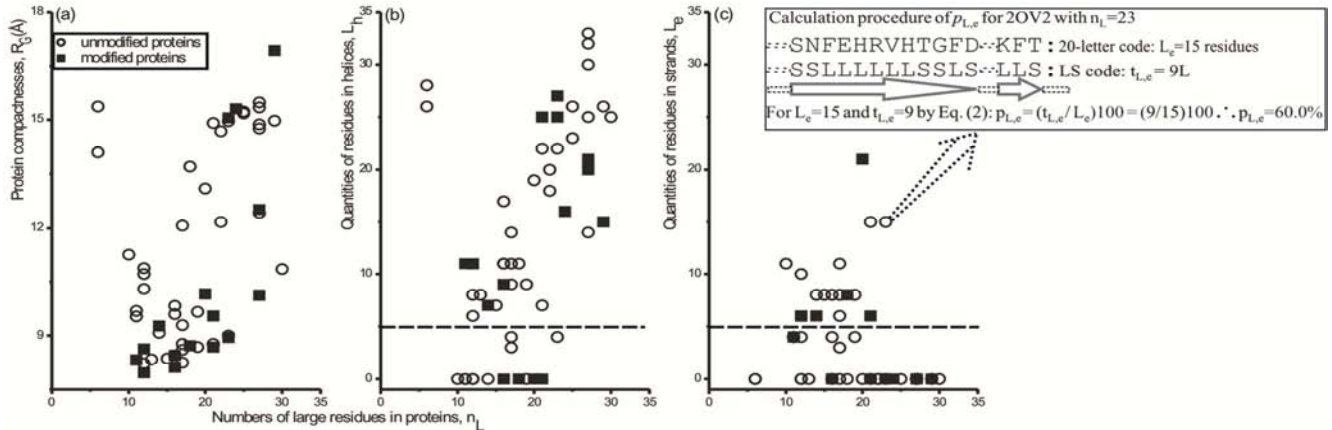


Figure 1. Compactness R_G (a), and total length of helix L_h (b) and strand L_e (c) vs. number n_L of large residues in primary structure of 35-residue proteins. The baselines (b–c) bring out $L_j > 5$ to estimate $p_{i,j}$; and in the inset (c), the twenty-letter and large-small (LS) residue sequence and calculation of $p_{L,e}$ for the transferase 2OV2 are shown. The program Promotif [32] was used for specific assignments of L_j that associated with the binary codes (LS and HP) furnish actual $t_{i,j}$. All the PDB and Promotif data utilized here are freely disposable at: <http://www.pdb.org> and <http://www.ebi.ac.uk/pdbsum/>.

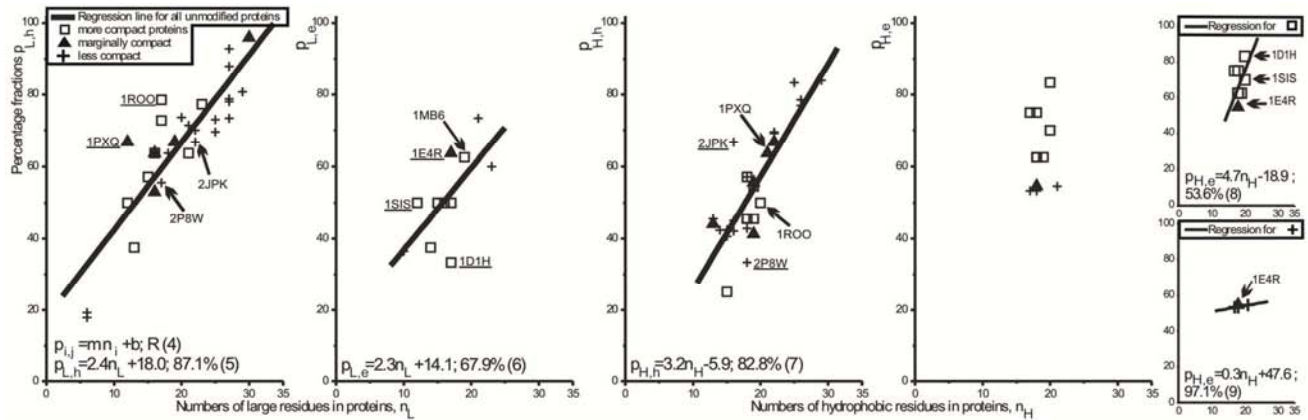


Figure 2. Percentage fraction $p_{i,j}$ of large and hydrophobic residues in helix and strand in function of the number n_i of these residues in primary structure, and linear relations ($p_{i,j}$ vs. n_i) (5)–(9), for 40 unmodified samples. In 39 unmodified proteins of the template group, we find 5, 30 and 4 cases with both, one and no helix/strand ($L_j > 5$) originating 40 samples and, in consequence, 80 data points for $p_{i,j}$ altogether. The underlined proteins are classified as: antimicrobial (1PXQ, 2JPK), defensin (1E4R), potassium channel inhibitor (1ROO), toxin (1D1H, 1SIS, 1MB6), and translation (2P8W).

compensation because only their hydrophobic residues are very efficacious or selective. Similar conducts, like 1PXQ and 1ROO, are observed in other samples: 1D1H, 1E4R, 1SIS, at $p_{L,e}$ (6); 2JPK, 2P8W, $p_{H,h}$ (7); 1MB6, $p_{H,e}$ (8). These samples present particular structural features [36–39], such as reasonable conformational flexibility, relatively short secondary structure motifs ($L_j < 10$), and motifs with residues into segments close to flexible N- or C-terminal domains.

Different from the first three cases ($p_{L,h}$ and $p_{L,e}$) (5)–(7)), the fractions $p_{H,e}$ of hydrophobic residues in strands are dependent on the packing density, whereas more compact globules ($R_G \leq 9.0$ Å) assume an upward-sloping regression line ((8) alike to (5)–(7)), and in a different way the less dense forms ($R_G > 11.0$ Å) are better described by an almost horizontal line (9). The dependence of $p_{H,e}$ with the conformational packing may concurrently result from long-range interactions into hydrophobic exclusions, and non-local bonds into strands forming β -pleated sheets. The partially

tight sample 1E4R ($9.0 \text{ Å} < R_G \leq 11.0 \text{ Å}$) adjusts reasonably well in both linear relations (8)–(9), maintaining fairly unchanged these relations.

The linear relations $p_{i,j}$ (5)–(9) are validated partitioning our benchmark dataset into separated subsamples similar to cross-validation tests in statistical analyses [40]. In different partitions, the $p_{i,j}$ relations remain almost unaltered and, therefore, they should be considered reliable and well constituted. For example, in three sub-collections of the present dataset (compact non-toxins [41], $R_G \leq 11.0$ Å; unmodified samples, Figure 2; together modified and unmodified samples [42]) for helices ($p_{L,h}$ (5)) with 24, 29 and 41 data points, the slopes m are equal to 2.6, 2.4 and 2.5; for $p_{H,h}$ (7), m are 3.2, 3.2 and 3.2; and for strands ($p_{L,e}$ (6)) with 4, 11 and 17 points, m are 2.3, 2.3 and 2.6, respectively. Eqs. (5)–(9) are re-validated by predictions in protein samples of the target subgroups I and II, as shown in the following subsections.

3.3. Influence of Non-Standard Amino Acid Residues in the Target Subgroup I

The sequence–structure relationship between p_{ij} and n_i articulated by five crucial rules (5)–(9) from unmodified samples may or may not be amenable to validation, when applied to modified samples with non-standard amino acids. In addition, whenever the length L_j is previously known, one can predict the most probable contents (or estimated t_{ij}) of large and hydrophobic residues in secondary structure topologies from (2) and (4), according to the estimation equations given as:

$$t_{ij} = L_j (mn_i + b)/100 \quad (10)$$

Utilizing the previous numerical expressions (5)–(10), residue content predictions are made for 12 tubular helices and 6 extended strands in their corresponding graphs, totalizing 36 estimated t_{ij} that are compared with their 36 actual t_{ij} from Promotif (Figure 3). For hydrophobic interplays in β -strands ($p_{H,e}$) of slightly compact samples, the linear relation from very tight globular conformations (8) is arbitrarily employed.

In the current detailed case study, one observes strategic acting of the steric and hydrophobic interactions by means of molecular mechanisms given by: (a) in 26 out of 36 estimates (Figure 3), 13 modified samples make use of a double

effective mechanism with both residue sub-components being simultaneously used in the native conformations, as seen by narrow proximity of actual and estimated t_{ij} (both $\Delta t_{ij} \leq 1.5$ from (3)); (b) seemingly, in eight estimates, four once underlined samples (1BDE, 1C4E, 1RH4 and 1WY3) employ a single mechanism with a looser sub-component (only one $\Delta t_{ij} \leq 1.5$) in the helical or β -sheet topologies, mainly regarding the hydrophobic selectivity; and (c) in two remaining estimates, one double-underlined sample (1JY4) uses both $\Delta t_{ij} \geq 2.0$. Before attributing such unexpected outputs ($\Delta t_{ij} \geq 2.0$ in items (b)–(c)) to occurrence of non-standard amino acids, a deeper survey for reevaluation of these outputs (Figure 4) is necessary.

Two probable reasons for 6 apparently undesirable predictions t_{ij} in 5 underlined samples (items (b)–(c) above, Figure 4) are sizable extents of L_j (first term in (10)), greater or approximated to 10; or the performance of linear relations ($mn_i + b$, second term in (10)). Some values of estimated t_{ij} are related to $L_j \geq 10$ and allow differences $\Delta t_{ij} - 0.1L_j \leq 1.0$ that occur for four predictions in 1JY4 ($t_{L,e}$), 1BDE, 1RH4, 1WY3 ($t_{H,h}$), with Δt_{ij} equal to 2.4, 2.3, 2.2, 1.7, resulting from L_j equal to 21, 27, 16, 25, and so with $\Delta t_{ij} - 0.1L_j$ equal to 0.3, -0.4, 0.6, -0.8, respectively. Hence, when $L_j \geq 10$, the strict accuracy requirement of $\Delta t_{ij} \approx 0.0$ or 1.0 is re-evaluated, and $\Delta t_{ij} \approx 2.0$ is considered plausible—that is, in the vicinity of one-tenth of L_j , as long as $\Delta t_{ij} - 0.1L_j \leq 1.0$.

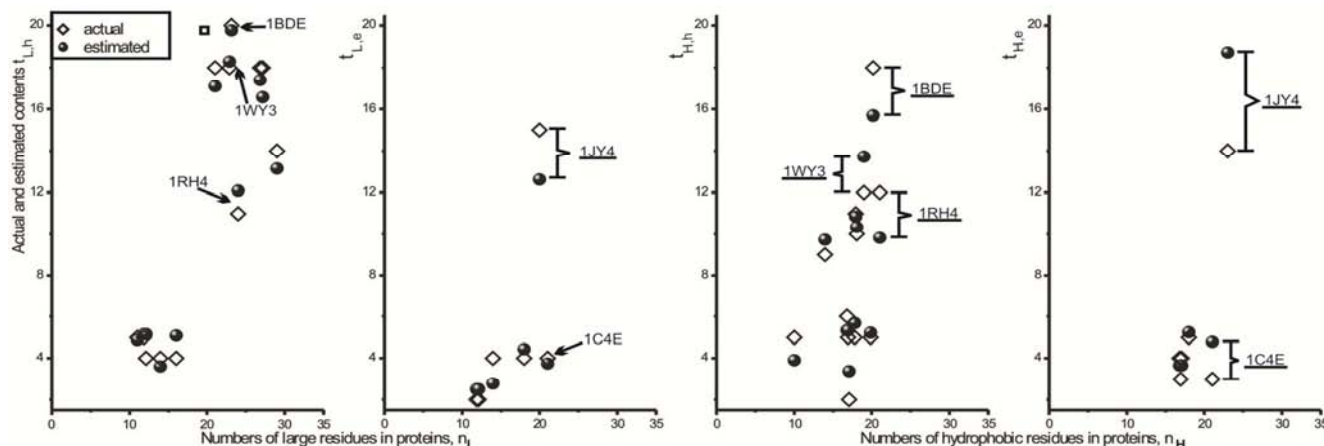


Figure 3. Actual and estimated content t_{ij} of large and hydrophobic residues in helices and strands for 18 modified samples. The 16 post-translationally modified proteins in the target subgroup I have 3, 12 and 1 cases with both, one and no helix/strand resulting in 18 samples and, consequently, 36 actual and estimated t_{ij} . The underlined proteins are sorted as: AIDS (1BDE), coiled coil (1RH4), de novo (1JY4), plant (1C4E), and structural (1WY3).

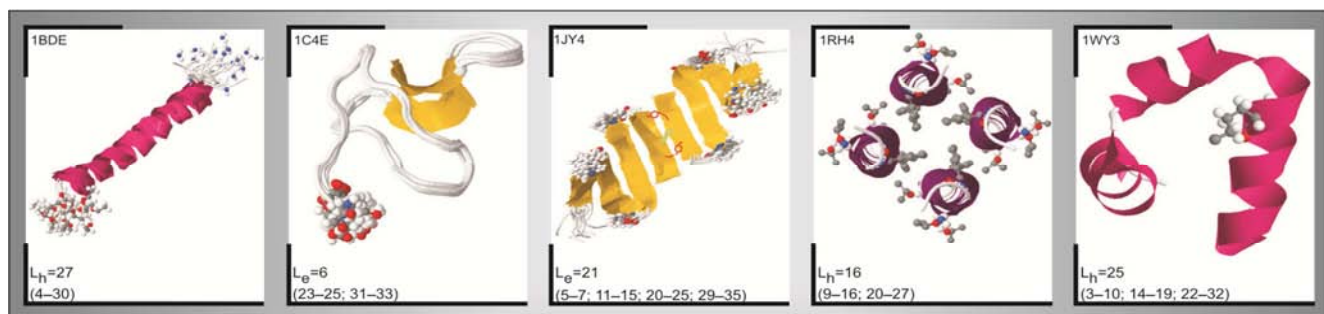


Figure 4. Three-dimensional native configuration, length L_j and residue number range in helix or strand of five underlined proteins in Figure 3. Biological assemblies of 1JY4 and 1RH4 with quaternary structures are shown, and non-standard amino acid residues are seen in a ball and stick model. Native configurations were created from the Jmol viewer [43].

In the two remaining predictions (1C4E and 1JY4) with seemingly inconvenient estimations, both estimated $t_{H,e}$ are connected with the appropriate application of the linear relations (mm_i+b) from less densely folded structures (9) instead of the relation (8) arbitrarily utilized. Thus, the highly tight 1C4E ($L_e=6$) reduces its $\Delta t_{H,e}$ from 1.8 (Figure 3) to 0.3 (now). The marginally compact 1JY4 ($L_e=21$) also compresses its $\Delta t_{H,e}$ from a bad value equal to 4.7 (Figure 3) to 2.4 (now), with $\Delta t_{H,e}-0.1L_e$ equal to 0.3. Therefore, all underlined samples of the items (b)–(c) above should be considered employing a double effective mechanism; and furthermore, 1C4E and 1JY4 reveal that the almost horizontal straight line for $p_{H,e}$ (9) can be predominant, but not exclusive in less dense forms.

1JY4 is alone with both $\Delta t_{i,e} \approx 2.0$, which may happen due to some interactional and structural particularities [15, 44]. Since 1JY4 is a precursor peptide (two-fold symmetric), arising from a 70-residue multistranded polypeptide; it also has hydrophobic interactions crossing its dimer interface, residues Y with I, and it contains one covalent disulfide bond stabilizing antiparallel β -strands in the boundary of two four-stranded β -sheets, as well as utilizing three non-standard residues DPR with ability to nucleate β -turns, at the positions 9, 17 and 27. The single mechanism of unmodified samples (Figure 2) was not detected in modified samples, because in those former ones, we had another greatness and a rougher metric ($p_{i,j}$ of 0–100%), in addition to another goal that was to quantify the sequence–structure relationship (5)–(9). In the subgroup I, only the double mechanism was identified. New inspections and outputs for additional modified–samples (subgroup II) are shown next.

Thus far, 35-residue proteins in the template group and target subgroup I with a total of 116 structural inspections (Figures 2–3) were selected and analyzed. However, the linear estimation equations (2), (4)–(10) are applicable to other post-translationally modified proteins of N residues. For new sequence and structure surveys, we only alter the slopes m (multiplying them by 35) and normalize the numbers n_i (dividing them by N) in (5)–(9) for 89 further modified proteins in the target subgroup II with 29 non-standard amino acids that are not encoded by genetic codes. This subgroup contains other 35-residue proteins as well as all those non-redundant ones from 36 to 40 residues currently available in the PDB database.

3.4. Non-Standard Residues and Molecular Mechanisms in the Target Subgroup II

Five (ACE, NH2, NLE, PCA and SIN) among nine non-standard amino acids are present in the modified proteins of the subgroups I and II. Other 24 non-standard amino acids [20] employed here occur exclusively in the subgroups II. For more details on the 33 non-standard amino acids and the proteins of the subgroups I and II, see Table 1 in Supplementary Data Appendices.

In the 89 modified proteins of the subgroup II, there are 7, 73 and 9 proteins with both, one and no secondary elements

originating to 87 samples being formed by 71 helix and 16 strand samples. 71 samples with helices summing 142 $\Delta t_{i,h}$ (Figure 5a) are firstly inspected. Furthermore, a careful analysis is made when a sample has one $\Delta t_{i,h} \geq 2.0$ (Figure 5b) or both $\Delta t_{i,h} \geq 2.0$ (Figure 5c).

Among 71 samples with helical structures, 49 of them have both points $\Delta t_{i,h} \leq 1.5$ inside translucent rectangles (Figure 5a), and thus utilizing a double effective mechanism, in contrast to other 22 underlined samples (17 single-underlined and 5 double-underlined ones, with one and both $\Delta t_{i,h} \geq 2.0$, respectively). In order to more thoroughly investigate these 22 special cases, the biophysicochemical compositions $\Delta t_{i,h}$ and the lengths L_h through $\Delta t_{i,h}-0.1L_h$ (Figure 5b–c) are observed and shown that 16 samples (12 with one $\Delta t_{i,h} \leq 1.5$ and another $\Delta t_{i,h}-0.1L_h \leq 1.0$ (Figure 5a–b), and 4 with both $\Delta t_{i,h}-0.1L_h \leq 1.0$ (Figure 5c)) utilize the same mechanism of the 49 former samples summing 65 samples. Six bold-faced samples (numbers 8, 29, 32, 55, 56 (Figure 5b), and 30 (Figure 5c) with one $\Delta t_{i,h}-0.1L_h > 1.0$) employ a single mechanism.

Adopting similar procedures of the previous helical structures (Figure 5) for 16 modified samples with strand structures, 32 dissimilarities $\Delta t_{i,e}$ (Figure 6a) are surveyed particularly when a sample possesses one $\Delta t_{i,e} \geq 2.0$ (Figure 6b) or both $\Delta t_{i,e} \geq 2.0$ (Figure 6c). For the hydrophobic residues in strands, we show the best predictions using both $p_{H,e}$ (8) and (9) in each estimated $t_{H,e}$ (10).

Among 16 strand samples, 11 of them have $\Delta t_{i,e} \leq 1.5$ inside translucent rectangles (Figure 6a), the sample of number 14 with one $\Delta t_{i,e} \leq 1.5$ and another $\Delta t_{i,e}-0.1L_e \leq 1.0$ (Figure 6a–b), and the number 5 with both $\Delta t_{i,e}-0.1L_e \leq 1.0$ (Figure 6c) totaling 13 samples make use of a double effective mechanism. Other three bold-faced numbers (6, 7, 15 with $\Delta t_{i,e}-0.1L_e > 1.0$ (Figure 6b)) work with a single mechanism by the hydrophobic selectivity.

In summary, all the 210 residue content predictions (Figures 3, 5–6) for 105 modified samples are successful from the five linear relations (5)–(9) inside the estimation equation (10), and indicate a twofold molecular mechanism by the sub-components (large and/or hydrophobic) of the residues. Specifically, we observe: (a) in 96 out of 105 samples, a double effective mechanism leading to 192 excellent, good and/or acceptable predictions; and (b) in 9 bold-faced samples (six with helices (Figure 5b–c), and three with strands (Figure 6b)), a single mechanism furnishes an excellent, good or acceptable and another bad output for 18 predictions. Dissimilarities $\Delta t_{i,j} \approx 2.0$ or 3.0 (as long as $\Delta t_{i,j}-0.1L_j \leq 1.0$) should be tolerable as molecular fluctuations, whereas proteins are functional macromolecules of intrinsic dynamic nature, have non-ideal secondary arrangements, and can contain non-standard amino acids to aid punctual jobs, beyond other intra- and inter-molecular interactions influencing the steric and hydrophobic driving forces measured by the contents $t_{i,j}$ and dissimilarities $\Delta t_{i,j}$.

The sequence-based predictions and twofold molecular mechanism (items (a)–(b) above) further suggest that the

non-standard amino acids do not significantly modify the secondary structure contents of the standard amino acids into specific native conformations. Non-standard amino acids at the residue level act in reasonable harmony with biological roles exerted by the 20 naturally occurring standard amino acids, whereas they have local working and limited global influence measured by the content comparisons Δt_{ij} that

remain oscillating inside a bearable threshold of $\Delta t_{ij} \leq 1.5$ or of $\Delta t_{ij} - 0.1L_j \leq 1.0$. The existence of a bad or malfunctioning mechanism by both large and hydrophobic sub-components could indicate a more pronounced global influence of the non-standard residues, but this mechanism is hypothetically possible to occur in other proteins not analyzed yet.

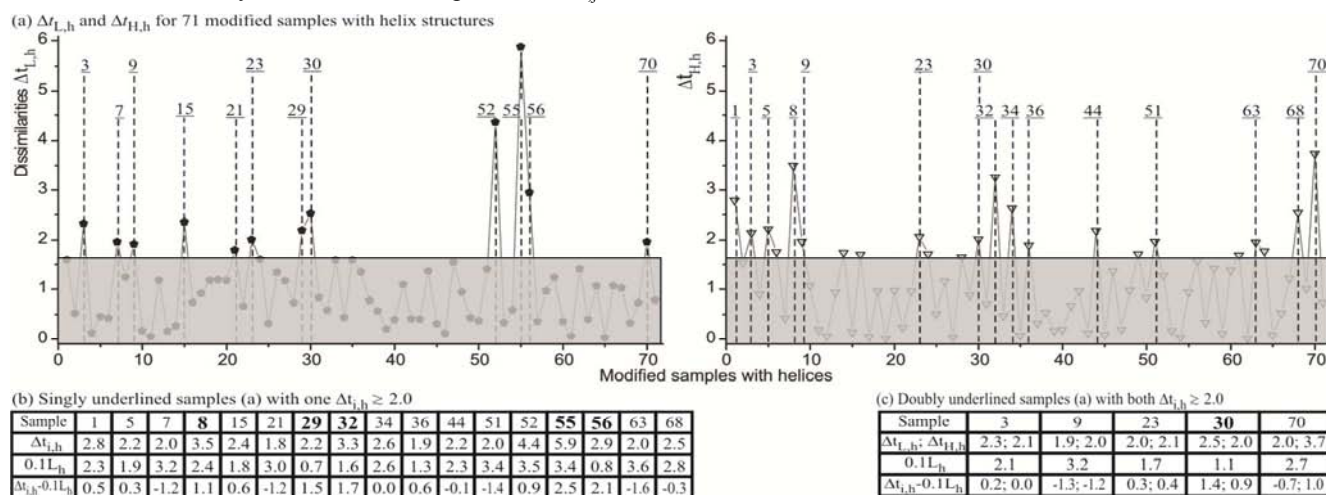


Figure 5. (a) Dissimilarity $\Delta t_{i,h}$ by large and hydrophobic residues in helix of 71 modified samples, (b) single and (c) double underlined samples (a) with $\Delta t_{i,h} \geq 2.0$.

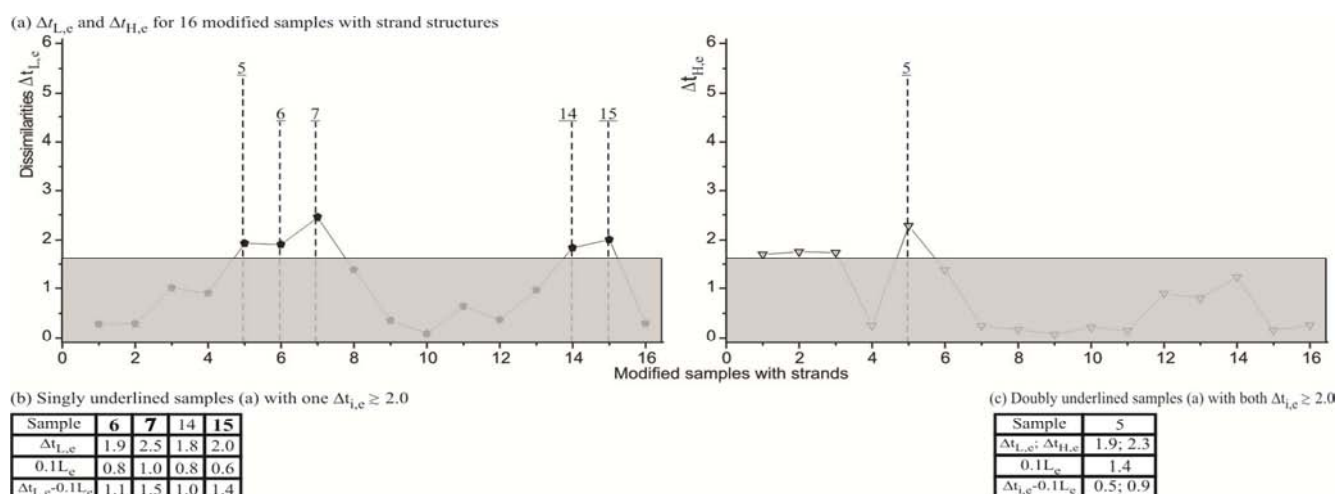


Figure 6. (a) Dissimilarity $\Delta t_{i,e}$ resulting from the steric and hydrophobic selectivity in strands of 16 modified samples, (b) single and (c) double underlined samples (a) with $\Delta t_{i,e} \geq 2.0$.

The current approach based on two coarse-grained (LS and HP residues) models is suitable in quantitative analyses and when atomic or molecular details can be suppressed or uninterested, but it is rather insufficient for sharper or local measurements, or very specific difference between non-standard and standard residues, as frequently occur in other applications utilizing coarse-grained models [45–48]. However, detailed approaches (such as semi-empirical or atomic models) are also limited in many features, since they are very CPU demanding. Furthermore, the complexity of realistic proteins and unknown factors of molecular interactions and cellular environments that detailed approaches strategically use are not yet fully understood [49–

50]. In addition our prognostication expressions (3), (5)–(10) are increasingly being used for other (un)modified proteins of different extensions and functions, with opportune and promising preliminary outputs [41–42]. This study therefore establishes the foundations and protocols of a future web-server software to determine quantitatively as happens the steric and hydrophobic selectivity and molecular mechanism in secondary structural motifs of a specific modified protein.

4. Conclusions and Future Developments

From 80 preliminary structural examinations in unmodified protein samples, the present article succeeded in

to quantify the global influence of non-standard compounds in modified samples through 210 inspections in cylindrical helices and flat β -sheets. The unmodified samples reveal a direct and linear synchronization between the opportune use ($p_{i,j}$) with the objective availability (n_i) of the large and hydrophobic residues by five special linear relations ($p_{i,j}$ vs. n_i in (5)–(9), Figure 2). In most cases both large and hydrophobic sub-components act concurrently by a double effective molecular mechanism. In some cases, one sub-component is more efficient or selective than the other through a single mechanism, but no sample owns $p_{i,j}$ simultaneously very far from the linear fits for both sub-components by means of a malfunctioning mechanism. The linear relations display the following: the simple and strategic rules employed for the transmission of orientation from primary to secondary levels, the interaction specificity in the protein structural organization, and the remarkable efficiency and combination of the residue biophysicochemical properties (volume and hydrophobicity) by folded configurations under varied conditions and contexts [51–53].

The post-translationally modified samples have suitable predictions for all the 210 compositions (as measured by $t_{i,j}$ and $\Delta t_{i,j}$, Figures 3, 5–6), confirm the double and single mechanisms already indicated by unmodified samples, and lead to a new prediction method. Such outputs reveal the balanced dependence between functional conformations with the residue sequences—as also shown by other structure prediction methods [54–56]—and establish the thorough contributions of the steric and hydrophobic selectivity. The successful predictions, besides the absence of a malfunctioning molecular mechanism, indicate that at a residue level the non-standard chemical compounds considered hitherto do not drastically alter the secondary structure contents suggested by the 20 standard amino acids. Consequently, such compounds work locally as additional, complementary and harmonic partners to the standard amino acids and immediately favor the complexity and robustness of the native-state conformations for skilled cellular and

physiological services, as well as the molecular diversifications from the genome to the proteome.

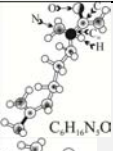

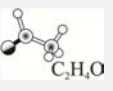

Taken together, our exhaustive quantitative analyses for modified and unmodified proteins are striking due to the high conformational and functional complexities of these proteins, including their residue compositions, compactness degrees, structural classes, and biological duties. These analyses can provide a deeper understanding of steric constraints and hydrophobic/hydrophilic interactions in sequence-based analysis problems, including the protein folding mechanism [57–58], configurational flexibility and stability [59], and sequence design [60]. Here, we worked to make a systematic progressions and protocols of a new computational prediction method for basic insights and knowledge on molecular interactions and mechanisms, and quantification of the global influence by non-standard amino acids. In the future, we shall supply a user-friendly and freely accessible web server for our method, contributing and collaborating with other useful databases, specialized servers, and web resources [20, 61–62] for post-translationally modified proteins.

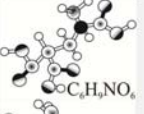


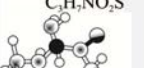
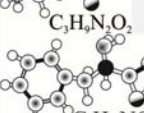
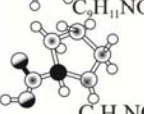
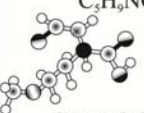
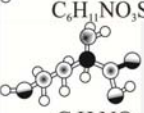
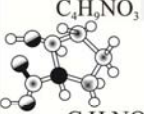
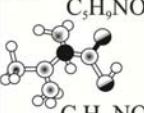
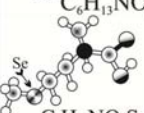


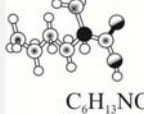

Supplementary Data Appendices

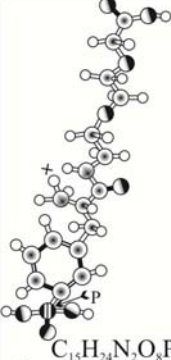
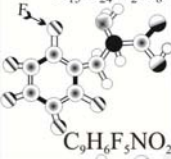
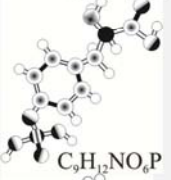
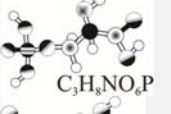


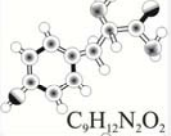
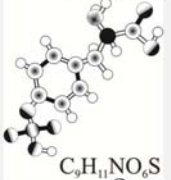
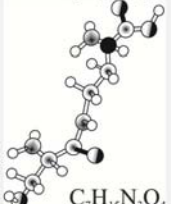
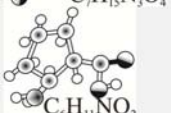
This work has surveyed independently two target subsets (subgroups I and II) of modified proteins with non-standard amino acids. The first previously examined in the subsection 3.3 comprises sixteen 35-residue proteins and residue sequences with at least one non-standard amino acid in 9 frameworks (Table 1). The second inspected in the subsection 3.4 consists of 89 modified proteins with 35 to 40 residues and at least one non-standard amino acid among 29 non-standard compounds (Table 1).

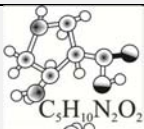
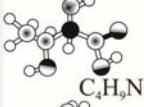
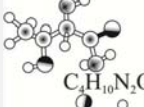
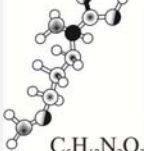
Table 1 has columns with code and name of the 33 non-standard amino acids in alphanumeric order, molecular structure and formula of these amino acids, a parent standard amino acid (if any), binary code for the volume and hydrophobicity, target subgroup (I and/or II), and PDB code for modified proteins, respectively.

Table 1. Information for 33 non-standard amino acids in modified proteins.

Chemical group	Structure/Formula	Parent	Binary code	Subgroup	PDB code
AAR Arginineamide	 <chem>C6H14N4O</chem>	Arg	L, P	II	1ZTO
ABA Alpha-aminobutyric acid	 <chem>C4H9NO2</chem>	Ala	S, H	I	1C2U, 1WZ5
ACE Acetyl group	 <chem>C2H4O</chem>	—	S, P	I, II	1BDE, 1PEH, 1RH4, 1ZWF; 1AIK, 1BB1a, 1BB1b, 1BB1c, 1JEK, 1KD8, 1PONA, 1PONb, 1UNC, 1UND, 2BEQ, 2L1O, 2O6N, 2SIVa, 2SIVb, 2Z2T, 3AHA, 3G1E, 3O3Y, 3R46, 3TQ2, 3VIEa, 3VIEb, 3VTP, 3VTQa, 3VTQc, 3W19, 4HU5, 4I2L, 4P67
B3E (3s)-3-aminohexanedioic acid	 <chem>C6H11NO4</chem>	Glu	L, P	II	3G7A, 3O3Y

Chemical group	Structure/Formula	Parent	Binary code	Subgroup	PDB code
CGU Gamma-carboxy-glutamic acid	 <chem>C6H9NO6</chem>	Glu	L, P	II	2O6N
DAL D-alanine	 <chem>C3H7NO2</chem>	Ala	S, H	II	2MJ9
DCY D-cysteine	 <chem>C3H7NO2S</chem>	Cys	S, H	II	3FIE
DNP 3-amino-alanine	 <chem>C3H7N2O2</chem>	Ala	S, P	I	1BEI
DPN D-phenylalanine	 <chem>C9H11NO2</chem>	Phe	L, H	II	1PXQ,2RME
DPR D-proline	 <chem>C5H9NO2</chem>	Pro	S, H	I	1JY4
FME N-formylmethionine	 <chem>C6H11NO3S</chem>	Met	L, H	II	4IL6
HSE L-homoserine	 <chem>C4H9NO3</chem>	Ser	S, H	II	1CIR
HYP 4-hydroxy-proline	 <chem>C5H9NO3</chem>	Pro	S, H	II	2JTU,2LAQ
ILL Iso-isoleucine	 <chem>C6H13NO2</chem>	Ile	L, H	I	1RH4
MSE Selenomethionine	 <chem>C5H11NO2Se</chem>	Met	L, H	II	1VZJa,1VZJb,3A1G,3PBPc,3PBPI,4H62
NAL Beta-(2-naphthyl)-alanine	 <chem>C13H13NO2</chem>	Ala	L, H	II	2CYU
NH2 Amino group	 <chem>H2N</chem>	—	S, P	I; II	1BDE,1LU8,1PEH,1QUZ,1RH4,1TXM,1V56,1ZDC; 1ABZ, 1BB1a,1BB1b,1BB1c,1BKT,1EIT, 1ICY,1JEK,1J5B,1KTX,1K8V,1LJV,1PONA, 1PONb,1RYG,1SXM,1TZ4,1TZ5,1V90,1V91, 1WFB,2BEQ,2DF0,2D2P,2E2S,2KJ7,2K38, 2K9E,2LA2,2L1O,2L86,2O6N,2RLK,2SIVa, 2SIVb,1SIS,2Z2T,3AHA,3FIE,3G1E,3O3Y, 3R46,3VTP,3VTQa,3VTQc,3VU5,3W19,4HU5, 4P67
NLE Norleucine	 <chem>C6H13NO2</chem>	Leu	L, H	I; II	1WY3;2K9E,2Z2T, 4HU5
PCA Pyroglutamic acid	 <chem>C4H7NO3</chem>	Glu	S, H	I; II	1C4E; 1BIG,1LIR,2AXK 2BMT,2LZY,3Q8J,3R0L,4JTA

Chemical group	Structure/Formula	Parent	Binary code	Subgroup	PDB code
PFX (2s)-1-({2-[2-(carboxy-methoxy)ethoxy]ethyl}amino)-1-oxo-3-(4-phosphono-phenyl)pro-pan-2-aminium	 $C_{15}H_{24}N_2O_8P$	—	L, P	II	2K9E
PF5 2,3,4,5,6-pentafluoro-l-phenylalanine	 $C_9H_6F_5NO_2$	Phe	L, H	II	2JM0
PTR O-phospho-tyrosine	 $C_9H_{12}NO_6P$	Tyr	L, P	II	2RSY
SEP Phosphoserine	 $C_3H_8NO_6P$	Ser	L, P	II	2LIC,2LID
SIN Succinic acid	 $C_4H_6O_4$	—	S, P	I; II	1ZWG; 1ABZ
TPO Phosphothreonine	 $C_4H_{10}NO_6P$	Thr	L, P	II	2JOC
TYC L-tyrosinamide	 $C_9H_{12}N_2O_2$	—	L, H	II	2BF9
TYS O-sulfo-l-tyrosine	 $C_9H_{11}NO_6S$	Tyr	L, P	II	2K03
UU4 (2s)-2-amino-4-(l-serylamino)butanoic acid	 $C_7H_{15}N_3O_4$	—	L, P	II	4HU5
XCP (1s,2s)-2-amino-cyclopentanecarboxylic acid	 $C_6H_{11}NO_2$	—	L, H	II	3G7A,3O3Y

Chemical group	Structure/Formula	Parent	Binary code	Subgroup	PDB code
XPC (3s,4r)-4-amino pyrrolidine-3-carboxylic acid	 <chem>C5H10N2O2</chem>	—	L, H	II	3G7A,3O3Y
2TL D-allothreonine	 <chem>C4H9NO3</chem>	Thr	S, H	II	1PXQ
9AT Amidated threonine	 <chem>C5H10N2O2</chem>	—	S, H	II	2Y8T
19W 5-(aminooxy)-l-norvaline-acid	 <chem>C6H12N2O3</chem>	—	L, P	II	4HU5

References

- [1] C. T. Walsh, S. Garneau-Tsodikova, G. J. Gatto Jr., Protein Posttranslational Modifications: The Chemistry of Proteome Diversifications, Angew. Chem. Int. Ed. 44, 2005, 7342–7372.
- [2] Y. Tweedie-Cullen, I. M. Mansuy, Towards a better understanding of nuclear processes based on proteomics, Amino Acids 39, 2010, 1117–1130.
- [3] O. N. Jensen, Modification-specific proteomics: characterization of post-translational modifications by mass spectrometry, Curr. Opin. Chem. Biol. 8, 2004, 33–41.
- [4] J. Seo, K. J. Lee, Post-translational Modifications and Their Biological Functions: Proteomic Analysis and Systematic Approaches, J. Biochem. Mol. Biol. 37, 2004, 35–44.
- [5] E. S. Groban, A. Narayanan, M. P. Jacobson, Conformational Changes in Protein Loops and Helices Induced by Post-Translational Phosphorylation, PLoS Comput. Biol. 2, 2006, 0238–0250.
- [6] J. H. McKerrow, E. Sun, P. J. Rosenthal, The proteases and pathogenicity of parasitic protozoa, Annu. Rev. Microbiol. 47, 1993, 821–853.
- [7] C. Drouet, A. Désormeaux, J. Robillard, D. Ponard, L. Bouillet, L. Martin, G. Kanny, D. A. M. Vautrin, J. L. Bosson, J. L. Quesada, M. L. Trascasa, A. Adam, Metallopeptidase activities in hereditary angioedema: Effect of androgen prophylaxis on plasma aminopeptidase P, J. Allergy Clin. Immunol. 121, 2008, 429–433.
- [8] S. J. Dunne, R. B. Cornell, J. E. Johnson, N. R. Glove, A. S. Tracey, Structure of the membrane binding domain of CTP: Phosphocholine Cytidylyltransferase, Biochemistry 35, 1996, 11975–11984.
- [9] P. Savarin, R. Romi-Lebrun, S. Zinn-Justin, B. Lebrun, T. Nakajima, B. Gilquin, A. Ménez, Structural and functional consequences of the presence of a fourth disulfide bridge in the scorpion short toxins: Solution structure of the potassium channel inhibitor HsTX1, Prot. Scien. 8, 1999, 2672–2685.
- [10] K. D. Hapner, P. E. Wilcox, Fragmentation of bovine chymotrypsinogen A and chymotrypsin A. Specific cleavage at arginine and methionine residues and separation of peptides, including B and C chains of chymotrypsin, Biochemistry 9, 1970, 4470–4480.
- [11] V. Serval, T. Galli, A. Cheramy, J. Glowinski, S. Lavielle, *In vitro* and *in vivo* inhibition of N-acetyl-L-aspartyl-L-glutamate catabolism by N-acylated L-glutamate analogs, J. Pharmacol. Exp. Ther. 260, 1992, 1093–1100.
- [12] M. W. Pennington, M. D. Lanigan, K. Kalman, V. M. Mahnir, H. Rauer, C. T. McVaugh, D. Behm, D. Donaldson, K. G. Chandy, W. R. Kem, R. S. Norton, Role of disulfide bonds in the structure and potassium channel blocking activity of ShK toxin, Biochemistry 38, 1999, 14549–14558.
- [13] L. Carrega, A. Mosbah, G. Ferrat, C. Beeton, N. Andreotti, P. Mansuelle, H. Darbon, M. D. Waard, J. M. Sabatier, The impact of the fourth disulfide bridge in scorpion toxins of the α -KTx6 subfamily, Proteins 61, 2005, 1010–1023.
- [14] K. Kalman, M. W. Pennington, M. D. Lanigan, A. Nguyen, H. Rauer, V. Mahniri, K. Paschetto, W. R. Kem, S. Grissmer, G. A. Gutman, E. P. Christian, M. D. Cahalan, R. S. Norton, K. G. Chandy, ShK-Dap²², a potent Kv1.3-specific immunosuppressive polypeptide, J. Biol. Chem. 273, 1998, 32697–32707.
- [15] J. Venkatraman, G. A. N. Gowda, P. Balaram, Design and construction of an open multistranded β -sheet polypeptide stabilized by a disulfide bridge, J. Am. Chem. Soc. 124, 2002, 4987–4994.
- [16] P. B. Harbury, J. J. Plecs, B. Tidor, T. Alber, P. S. Kim, high-resolution protein design with backbone freedom, Science 282, 1998, 1462–1467.
- [17] T. K. Chiu, J. Kubelka, R. Herbst-Irmer, W. A. Eaton, J. Hofrichter, D. R. Davies, High-resolution x-ray crystal structures of the villin headpiece subdomain, an ultrafast folding protein, Proc. Natl. Acad. Sci. USA 102, 2005, 7517–7522.
- [18] J. I. Fletcher, A. J. Dingley, R. Smith, M. Connor, M. J. Christie, G. F. King, High-resolution solution structure of gurmairin, a sweet-taste-suppressing plant polypeptide, Eur. J. Biochem. 264, 1999, 525–533.

- [19] Y. C. Lou, Y. C. Huang, Y. R. Pan, C. Chen, Y. D. Liao, Roles of N-terminal pyroglutamate in maintaining structural integrity and pK_a values of catalytic histidine residues in bullfrog ribonuclease 3, *J. Mol. Biol.* 355, 2006, 409–421.
- [20] E. S. Witze, W. M. Old, K. A. Resing, N. G. Ahn, Mapping protein post-translational modifications with mass spectrometry, *Nat. Methods* 4(10), 2007, 798–806.
- [21] F. C. Bernstein, T. F. Koetzle, G. J. B. Williams, E. F. Meyer Jr., M. D. Brice, J. R. Rodgers, O. Kennard, T. Shimanouchi, M. Tasumi, The protein data bank: A computer-based archival file for macromolecular structures, *Eur. J. Biochem.* 80, 1977, 319–324.
- [22] M. I. Sadowski, D. T. Jones, The sequence–structure relationship and protein function prediction, *Curr. Opin. Struct. Biol.* 19, 2009, 357–362.
- [23] W. Taylor, The classification of amino acid conservation, *J. Theor. Biol.* 119, 1986, 205–218.
- [24] G. E. Schulz, R. H. Schirmer, Noncovalent forces determining protein structure, in: C. R. Cantor (Ed.), *Principles of Protein Structure*, Springer-Verlag, New York, 1990, 27–45.
- [25] R. Srinivasan, G. D. Rose, LINUS: A hierarchic procedure to predict the fold of a protein, *Proteins* 19, 1995, 81–99.
- [26] L. F. O. Rocha, I. R. Silva, A. Caliri, Distinct conformational properties determined by implicit and explicit representation of protein-solvent interactions. An analytical and computer simulation study, *Phys. A* 388, 2009, 4097–4104.
- [27] H. Goodarzi, A. Katanforoush, N. Torabi, H. S. Najafabadi, Solvent accessibility, residue charge and residue volume, the three ingredients of a robust amino acid substitution matrix, *J. Theor. Biol.* 245, 2007, 715–725.
- [28] A. A. Zamyatnin, Protein volume in solution, *Progr. Biophys. Mol. Biol.* 24, 1972, 107–123.
- [29] C. Chothia, Structural invariants in protein folding, *Nature* 254, 1975, 304–308.
- [30] S. Moelbert, E. Emberly, C. Tang, Correlation between sequence hydrophobicity and surface-exposure pattern of database proteins, *Protein Sci.* 13, 2004, 752–762.
- [31] L. F. O. Rocha, M. E. P. Tarragó, A. Caliri, The water factor in the protein-folding problem, *Braz. J. Phys.* 34, 2004, 90–101.
- [32] E. G. Hutchinson, J. M. Thornton, PROMOTIF-A program to identify and analyze structural motifs in proteins, *Protein Sci.* 5, 1996, 212–220.
- [33] N. Bhardwaj, M. Gerstein, Relating protein conformational changes to packing efficiency and disorder, *Prot. Sci.* 18, 2009, 1230–1240.
- [34] R. Sreekanth, S. S. Rajan, The study of helical distortions due to environmental changes: Choice of parameters, *Biophys. Chem.* 125, 2007, 191–200.
- [35] T. Haltia, E. Freire, Forces and factors that contribute to the structural stability of membrane proteins, *Biochim. Biophys. Acta* 1228, 1995, 1–27.
- [36] K. E. Kawulka, T. Sprules, C. M. Diaper, R. M. Whittall, R. T. McKay, P. Mercier, P. Zuber, J. C. Vederas, Structure of subtilisin A, a cyclic antimicrobial peptide from *Bacillus subtilis* with unusual sulfur to α -carbon cross-links: Formation and reduction of α -thio- α -amino acid derivatives, *Biochemistry* 43, 2004, 3385–3395.
- [37] H. Takahashi, J. I. Kim, H. J. Min, K. S. Kenton, J. Swartz, I. Shimada, Solution structure of hanatoxin1, a gating modifier of voltage-dependent K^+ channels: Common surface features of gating modifier toxins, *J. Mol. Biol.* 297, 2000, 771–780.
- [38] D. J. Taylor, J. Nilsson, A. R. Merrill, G. R. Andersen, P. Nissen, J. Frank, Structures of modified eEF2.80S ribosome complexes reveal the role of GTP hydrolysis in translocation, *Embo J.* 26, 2007, 2421–2431.
- [39] K. Peng, Q. Shu, Z. Liu, S. Liang, Function and solution structure of huwentoxin-IV, a potent neuronal tetrodotoxin (TTX)-sensitive sodium channel antagonist from chinese bird spider celenocosmia huwena, *J. Biol. Chem.* 277, 2002, 47564–47571.
- [40] S. Borra, A. D. Ciaccio, Measuring the prediction error. A comparison of cross-validation, bootstrap and covariance penalty methods, *Comput. Statist. Dat. Analys.* 54, 2010, 2976–2989.
- [41] L. F. O. Rocha, Analysis of molecular structures and mechanisms for toxins derived from venomous animals, *Comput. Biol. Chem.* 61, 2016, 8–14.
- [42] L. F. O. Rocha, Toward a better understanding of structural divergences in proteins using different secondary structure assignment methods, *J. Mol. Struct.* 1063, 2014, 242–250.
- [43] R. M. Hanson, Jmol – a paradigm shift in crystallographic visualization, *J. Appl. Cryst.* 43, 2010, 1250–1260.
- [44] J. M. Thornton, Protein structures: The end point of the folding pathway, in: T. E. Creighton (Ed.), *Protein Folding*, W. H. Freeman and Company, New York, 1992, 59–81.
- [45] B. Li, M. Lin, Q. Liu, Y. Li, C. Zhou, Protein folding optimization based on 3D off-lattice model via an improved artificial bee colony algorithm, *J. Mol. Model.* 21, 2015, 261–1–15.
- [46] F. L. Custódio, H. J. C. Barbosa, L. E. Dardenne, A multiple minima genetic algorithm for protein structure prediction, *Appl. Soft Comput.* 15, 2014, 88–99.
- [47] J. Santos, P. Villot, M. Diéguez, Emergent Protein Folding Modeled with Evolved Neural Cellular Automata Using the 3D HP Model, *J. Comp. Biol.* 21, 2014, 823–845.
- [48] A. Irbäck, J. Wessén, Thermodynamics of amyloid formation and the role of intersheet interactions, *J. Chem. Phys.* 143, 2015, 105104–1–9.
- [49] H. I. Ingólfsson, C. A. Lopez, J. J. Uusitalo, D. H. Jong, S. M. Gopal, X. Periole, S. J. Marrink, The power of coarse graining in biomolecular simulations, *WIREs Comput. Mol. Sci.* 4, 2014, 225–248.
- [50] W. Li, H. Yoshii, N. Hori, T. Kameda, S. Takada, Multiscale methods for protein folding simulations, *Methods* 52, 2010, 106–114.
- [51] T. J. Richmond, Solvent accessible surface area and excluded volume in proteins: Analytical equations for overlapping spheres and implications for the hydrophobic effect, *J. Mol. Biol.* 178, 1984, 63–89.
- [52] E. Kussell, J. Shimada, E. I. Shakhnovich, Excluded volume in protein side-chain packing, *J. Mol. Biol.* 311, 2001, 183–193.

- [53] L. R. Pratt, Molecular Theory of hydrophobic effects: "She is too mean to have her name repeated.", *Annu. Rev. Phys. Chem.* 53, 2002, 409–436.
- [54] G. E. Crooks, J. Wolfe, S. E. Brenner, Measurements of protein sequence–structure correlations. *Proteins*, *Proteins* 57, 2004, 804–810.
- [55] K. Sobha, C. Kanakaraju, K. S. K. Yadav, Is protein structure prediction still an enigma?, *Afr. J. Biotechnol.* 7, 2008, 4687–4693.
- [56] C. Benros, A. G. Brevern, S. Hazout, Analyzing the sequence–structure relationship of a library of local structural prototypes, *J. Theor. Biol.* 256, 2009, 215–226.
- [57] A. M. Gutin, V. I. Abkevich, E. I. Shakhnovich, Is burst hydrophobic collapse necessary for protein folding?, *Biochemistry* 34, 1995, 3066–3076.
- [58] B. Nölting, D. A. Agard, How general is the nucleation–condensation mechanism? *Proteins* 73, 2008, 754–764.
- [59] D. R. Livesay, S. Dallakyan, G. G. Wood, D. J. Jacobs, A flexible approach for understanding protein stability, *FEBS Lett.* 576, 2004, 468–476.
- [60] K. Fan, W. Wang, What is the minimum number of letters required to fold a protein?, *J. Mol. Biol.* 328, 2003, 921–926.
- [61] A. Yamaguchi, K. Iida, N. Matsui, S. Tomoda, K. Yura, M. Go, Het-PDB Navi.: A database for protein–small molecule interactions, *J. Biochem.* 135, 2004, 79–84.
- [62] G. J. Kleywegt, Crystallographic refinement of ligand complexes, *Acta Cryst. D* 63, 2007, 94–100.