International Journal of Biochemistry, Biophysics & Molecular Biology
2018; 3(2): 19-29
http://www.sciencepublishinggroup.com/j/ijbbmb
doi: 10.11648/j.ijbbmb.20180302.11
ISSN: 2575-5889 (Print); ISSN: 2575-5862 (Online)

Structural, Conformational and Interactional Investigation of Proteins with Related Sequences and Multiple Structures

Lessandra Eller, Luiz 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 (L. Rocha)

*Corresponding author

To cite this article:
Lessandra Eller, Luiz Rocha. Structural, Conformational and Interactional Investigation of Proteins with Related Sequences and Multiple Structures.*International Journal of Biochemistry, Biophysics & Molecular Biology*. Vol. 3, No. 2, 2018, pp. 19-29.
doi: 10.11648/j.ijbbmb.20180302.11

Received: April 18, 2018; **Accepted**: May 8, 2018; **Published**: June 1, 2018

Abstract: Homologous proteins are special macromolecules with related primary sequences and multiple native structures and together with sequence-unrelated nonhomologous ones both constitute the protein amazing universe. Here is made a thorough sample selection, and employed quantitative predictions to analyze structures, conformations, steric and hydrophobic interactions and underlying molecular mechanisms in proteins via two coarse-grained (hydrophobic-polar, large-small) models. First, five empirical relations from nonhomologous samples are determined correlating large and hydrophobic residue sequences from primary to helix and β-sheet structures of functional conformations. When applied to homologous proteins, such empirical relations allow precisely surveying the interaction performance, identifying four types of molecular mechanisms, and computing the stability level in conformation ensembles. 1764 structural inspections capture essential features and furnish structural-interactional insights for homologous proteins, as well as suggest a fruitful way for better understanding conformational variability in biomolecular processes such as protein evolution, dynamics, folding and design.

Keywords: Coarse-Grained Model, Conformational Ensemble, Homologous Protein, Molecular Sequence Data, Structural Homology, Sequence-Structure Alignment

1. Introduction

Proteins are specialized molecular machines vital for the existence and proper maintenance of all living organisms. They execute their crucial biological roles by means of an almost endless variety of functions that depend on their three-dimensional (3D) variform native structures constituted by secondary structure elements (mainly helices and β-sheets) and encoded by the amino acid sequences. However, the gap between the sequence and structure knowledge is inherently complex requiring a sum of many different driving forces and interactions, and involving a multitude of spatial and temporal scales, such that to predict unknown structures from the amino acid sequences alone still remain unsolved. Despite of this long-standing conundrum, many endeavors [1–6] have been done by researchers to reduce the protein sequence-structure gap since examining underlying principles and properties until advancing in applicative purposes, such as better understanding the biological and chemical activities of cells/organs, structure-based discovery of specific inhibitors, and to predict protein structures for rational structure based drug design in therapeutic insights, in the development of medicine, and to treat human diseases.

One of the simplest ways of contemplating the extent of the sequence-structure gap is comparing proteins [7–10] by means of alignments between sequences and structures that can be summarized in four broad subsets: (1) alignment whose low residue sequence identity below 25% reveals unrelated proteins; (2) alignment with low sequence identity in distantly related proteins that have changed their sequences by evolution process and are generally clustered into common fold; (3) considerably high sequence identity (>25%) in proteins that usually have both structural and evolutionary relatedness and are assorted into a same family, in addition it is often assumed that such proteins also possess similar structures; (4) alignment with very high sequence

identity (≳98%) in sequence-similar and structure-dissimilar protein chains. The identity threshold in 25% can assume different values depending of the study method and approach utilized.

The first above subset is commonly utilized through pre-stated filters in advanced search interfaces of macromolecule databases to remove redundant structures of third subset and to assemble protein structure library. Nonetheless, such sequence-based criterion for similarity may be harmful, because there are many proteins with high sequence identity but different structures of the fourth subset, so leading to loss of relevant structural and functional information [11]. The second and third subsets, on the other hand, are employed in template-based methods [12–14] of threading (or fold recognition) and homology modeling, respectively, to construct a model for a query or target structure utilizing a known template structure. The fourth subset represents special proteins with equal or very similar sequences but having reasonably dissimilar structures and this case will be more thoroughly evaluated here.

Factors contributing for structural differences in sequence-identical proteins (4° subset above) [11, 15–18] typically include: alternative conformations (e.g. protein crystallized in different spatial groups, alternative fits to the same NMR/crystallographic data); solvent (crystallization conditions with solvent in different pH or salt concentrations); temperature; apo versus ligand-bound forms of a protein; inter- or intra-chain interactions, as those due to different quaternary protein-protein, point mutation, oxidized versus reduced disulfide bridges; and large fragments or domain motions.

Here is explored the sequence-structure correlation and utilized two coarse-grained (HP (hydrophobic-polar) and LS (large-small)) models in a quantitative, empirical approach especially applied to homologous proteins in which one sequence can assume conformational multiplicity and functional diversity. This paper is arranged as follow: Section 2 sets next out the methodology for the selection of nonhomologous and homologous proteins, energetics (molecular interactions), secondary structure elements, and used structural variables. Then, Section 3 presents the initial results selecting samples, and computing in nonhomologous proteins the sequence-structure correlation via five empirical linear relations. In 684 structural-interactional inspections for homologous proteins, the linear relations are used to thoroughly examine the individual as well as mutual action of steric and hydrophobic interactions by four types of molecular mechanisms, quantify the strengths of these interactions, and measure the stability level for protein conformational ensembles. Lastly, the main conclusions are epitomized.

2. Materials and Methods

2.1. Definition of Homologous Proteins and Protein Structure Library

It is previously necessary to define the terms nonhomologous and homologous proteins used in this paper. Experimentally determined macromolecular structures deposited in the Protein Data Bank (PDB) [19] are culled under the following conditions:

(i) non-redundant chains with different primary sequences of at least three residues are included in the set of nonhomologous or sequence-unrelated proteins. Here is always examined primary structure solely consisting of 20 types of naturally occurring amino acids.

(ii) redundant protein chain pairs having none (or 100% sequenceidentity), one or two different residues in the primary sequence together with secondary structure elements (helices and/or β-strands) with less than four different residues should be removed one chain and the other inserted as member of the nonhomologous set.

(iii) parent chain pairs sharing primary sequences with none, one or two different residues, along with dissimilar segments of helices or strands in at least four different residues are both together inserted as part of the set of homologous or sequence-related proteins. The extension to one and two residue differences in primary sequences would allow us to employ our approach to explore mutation-induced fold changes, protein evolution and misfolding [20–21].

As a consequence of the conditions (i)–(iii), a protein pair should be considered as redundant only when both of their sequences and structures are highly similar and as homologous when both proteins are similar sequences and dissimilar structures. In order to select homologous proteins in a given chain length with N residues, an alignment and comparison of residue-per-residue sequences and secondary structure elements are employed together with the condition (iii) for each protein pair (Figure 1). In a helix and/or strand ensemble with n homologous proteins, the total number of protein pair combinations $C_{n,2}$ [22] is obtained by:

$$C_{n,2} = n!/(2!(n-2)!) \quad (1)$$

For instance, if n is equal to 2, 3, 4 (Figure 1), 5, 6 or 7 N-residue proteins, then there are 1, 3, 6, 10, 15, 21 $C_{n,2}$, respectively.

For nonhomologous proteins six chain lengths N that are 30, 40, 50, 60, 70, 100 residues, and twelve N for homologous proteins that are 30, 35, 40, 45, 50, 60, 70, 80, 90, 100, 120 and 200 are utilized. Residue sequence and function information for proteins were downloaded from online PDB archives [19] (at www.pdb.org). The program Promotif [23] (www.ebi.ac.uk/pdbsum) was employed to take specific assignments referring to secondary structures, helices and strands. For example, when explored PDB archives, it is identified that the current homologous proteins have tens of relevant biological roles, including antibiotic, antifreeze, antimicrobial, blood coagulation inhibitor, cell invasion, gene regulation, immune system, molecular motor, nuclear, steroid binding, toxin, transport, and viral. For more detailed information on these homologous proteins, see section of Supplementary Materials.

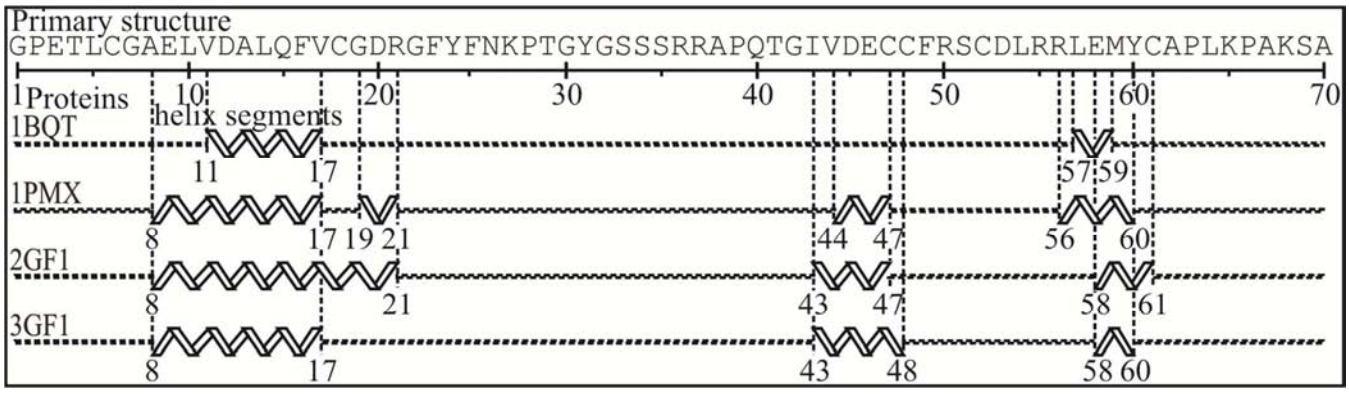

Figure 1. *Example of sequence-structure alignment and comparison to select four 70-residue homologous proteins with 100% sequence identity and different helical segments. The four growth factor/hormone proteins (PDB identifications: 1BQT, 1PMX, 2GF1, 3GF1) provide six pair combinations $C_{4,2}$ (1), whose helix differences (≥4 residues) inside parentheses are 1BQT (12, 15, 11 residues with the following proteins), 1PMX (5, 7), 2GF1 (6), 3GF1.*

2.2. Molecular Interactions

Proteins make use of a rich repertory of amino acid residues by means of strategic physico-chemical properties in their molecular and cell activities. Among these properties, the volume and hydrophobicity have been recognized indispensable in the selection and maintenance of native conformations and biological functions [3, 24–26], and they refer to steric and hydrophobic interactions. Here, the volume and hydrophobicity of 20 natural amino acids (single letters) are assigned by binary codes [27–29] – large-small LS and hydrophobic-polar HP – in the following subgroups: large-hydrophobic (F, H, I, L, M, V, W, Y), large-polar (E, K, Q, R), small-hydrophobic (A, C, P, T), and small-polar (D, G, N, S). These residue-level codes of four-letter sequences LS and HP capture essential features and information for proteins especially when confronted results of both, with the first referring the steric hindrance and macromolecular packing, and the second contemplating the hydrophobic interaction and effect. The large and hydrophobic sub-components are predominant, both detachedly with 12 among 20 amino acids, and therefore they are taken into account for the results below.

2.3. Computation of Residue Sequences and Secondary Structures

In the primary structures, the residue sequences can be properly expressed by total number of large (N_L) and hydrophobic (N_H) residues. Sequences with N_i may have none, one, two or many associated proteins, where the subscript index "i" accounts for the large (L) and hydrophobic (H) residues in both primary and secondary structural levels. In two periodic secondary structure elements (helices and β-sheets constituted by strands), here is not considered very short overall lengths L_j (only elements with $L_j>6$ residues) to have more reliable measures, since proteins are dynamically diffusive besides subjected to environmental perturbations [18, 30], where the index "j" stands for the (3_{10}, α, π)-helices (h) and β-strands (e). Furthermore, turns and coils have less accurate regions than helix and strand regions; hence, the formers will not be inspected here.

2.4. Sequence-Structure Variables and Their Accuracies

For proteins, the total numbers $t_{i,j}$ of large and hydrophobic residues in secondary structure elements of lengths L_j ensue in the real proportion $p_{i,j}$ (in percentage) measured by:

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

where $p_{i,j}$ ranges from 0 (whenever L_j does not possess large and hydrophobic residues, $t_{i,j}=0$) to 100% (every time that L_j entirely possesses these residues, $L_j=t_{i,j}$).

The estimated proportions $p_{i,j}$ of large and hydrophobic residues in helices and strands should be directly taken from prediction equations or expressions by means of linear fits in PDB data as below shown. The accurateness of our predictions is obtained by measuring $\Delta p_{i,j}$, the module of the dissimilarity between the real and estimated proportions $p_{i,j}$ through:

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

where $\Delta p_{i,j}$ can vary from zero (both $p_{i,j}$ are equals) to 100% (one $p_{i,j}$ is zero and another is 100%). More specifically, the prediction accuracy will be assumed excellent (whenever $\Delta p_{i,j}\leq 5\%$, that is with fluctuations $\Delta p_{i,j}\approx 0$), good ($5\%<\Delta p_{i,j}\leq 15\%$, $\Delta p_{i,j}\approx 10\%$), acceptable ($15\%<\Delta p_{i,j}\lesssim 25\%$ providing that $L_j\lesssim 15$), and bad (for further $\Delta p_{i,j}$).

3. Results and Discussion

3.1. Selection of Nonhomologous and Homologous Proteins

In the protein selection for each analyzed chain length N, proteins underwent post-translational modifications with non-natural amino acids (condition (i) above) were initially removed. Next, each pair of database proteins is aligned and compared by the residue sequences (via condition (i)) and, if necessary, also by the secondary structure elements (conditions (ii) and (iii)) and then excluded those redundant

chains; thus remaining the nonhomologous together with homologous macromolecules. After this stage they are partitioned into a nonhomologous or homologous (e.g. Figure 1) set, respectively. Figure 2a shows the residue sequence identity of 126 proteins with 70 residues that provide 7875 pair combinations $C_{126,2}$ (1). Figure 2b displays the helix and strand dissimilarity with at least four different residues (≥4 residues) for 61 homologous protein pairs with 100% sequence identity from Figure 2a.

The sequence identity (Figure 2a) for pairs of nonhomologous protein is frequently less than 25% and for homologous ones is equal to 100%. The homologous protein pairs (Figure 2b) have usually dissimilarities in helices or strands, but sometimes they occur in both secondary elements as shown for 8 pairs of numbers 11, 39, 42, 43, 47, 48, 57, 58. The results for the residue sequence identity (Figure 2a) and secondary structure dissimilarities (Figure 2b) are reasonably extensible for other chain lengths N, though here displayed only for N equal to 70 residues. From Figures 2a,b for 126 proteins, 94 nonhomologous and 32 homologous cases were selected. Also in other N, the nonhomologous proteins are in greater quantity and have more diversified residue sequences than those homologous ones; consequently, the nonhomologous macromolecules are first analyzed.

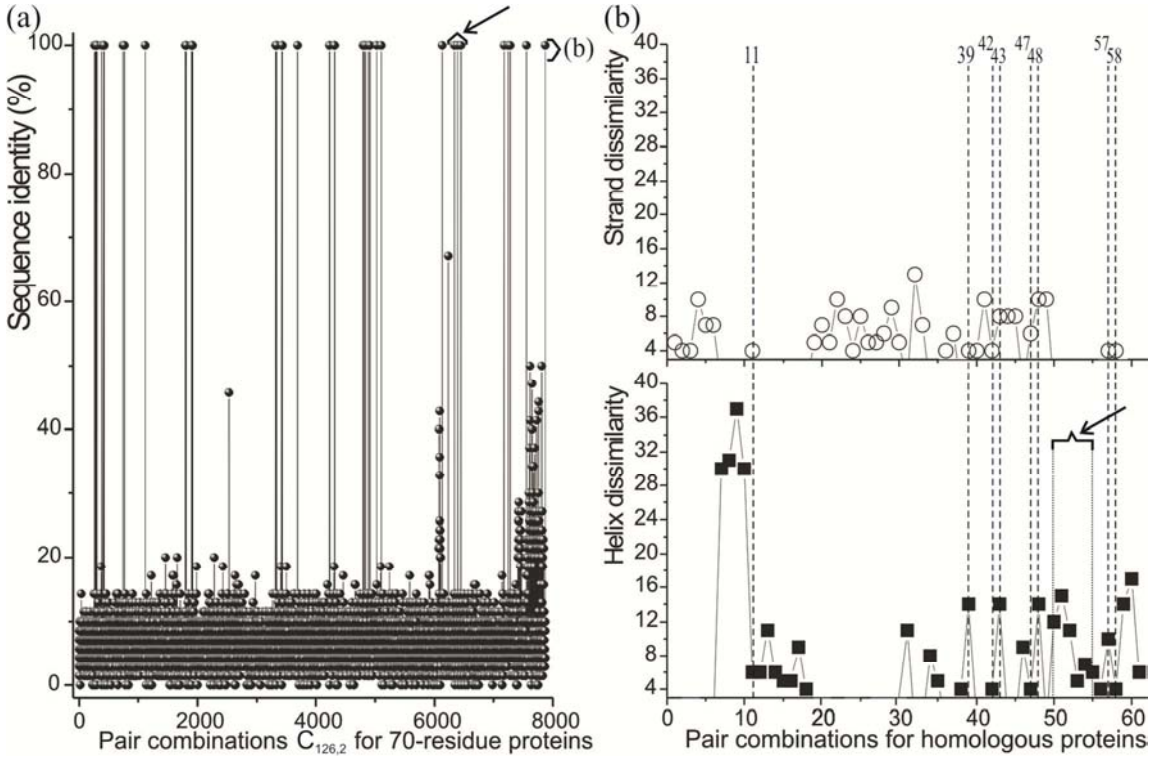

Figure 2. *(a) Percentage of sequence identity for 126 (non)homologous proteins that provide 7875 pair combinations $C_{126,2}$ and (b) helix and strand dissimilarity for 61 homologous protein pairs from (a). Four proteins in Figure 1 and their 6 pair combinations $C_{4,2}$ are precisely located into Figures a–b, as shown by inclined arrows. Here there are 32 homologous proteins into 8 conformational ensembles with 7, 5, 4, 6, 4 (Figure 1), 2, 2, 2 exemplars n that give rise to 21, 10, 6, 15, 6, 1, 1, 1 $C_{n,2}$ and thus a total of 61 $C_{n,2}$ for helices as well as strands (b).*

3.2. Measurement of Sequence-Structure Correlation for Nonhomologous Proteins

For nonhomologous proteins in each chain length N, the numbers of large and hydrophobic residues N_i from the primary structures are individually computed and then are observed the normalized quantities n_i (=N_i/N) with the real proportions $p_{i,j}$ (2) of these residues in the secondary structural elements, helices and strands. Though $p_{i,j}$ and n_i are apparently uncorrelated greatnesses, the plots of $p_{i,j}$ in function of n_i (Figure 3) are made for 317 helix and 223 strand data points, in a total amount of 1080 experimental data points, whose linear adjustments have general relations for estimated $p_{i,j}$ given by:

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

where m, b, and R are the slope, intercept, and linear correlation coefficient, and whose specific values (4(a)–(e)) are displayed in Figure 3.

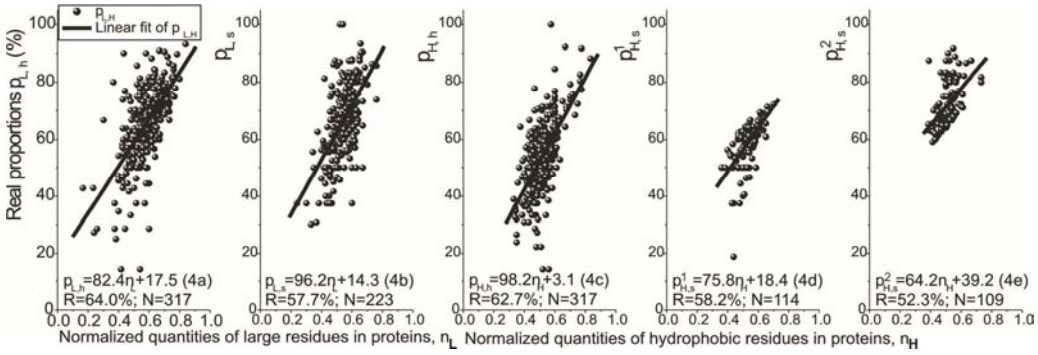

Figure 3. *Linear adjustments (4(a)–(e)) for 1080 real proportions $p_{i,j}$ in helices and strands in relation to normalized quantities n_i in primary structures of nonhomologous protein samples from PDB database. There are 465 proteins, of which 220, 100 and 145 cases own both, one and none helices/strands (with L_j>6) deriving 540 samples and accordingly 1080 $p_{i,j}$ on the whole.*

The 1080 points of $p_{i,j}$ versus n_i (linear relations (4(a)–(e)) in Figure 3) express how happen the information transference of large and hydrophobic residues in primary and secondary structures of folded conformations determined by X-ray crystallography, NMR spectroscopy or electron microscopy. These relations are dependent on the types of residues and secondary structure elements considering that the large residues in helices together with hydrophobic residues in strands have lesser sloped straight lines (with slopes m≲80.0 (4(a),(d),(e))) than the large residues in strands and hydrophobic ones in helices (m≈100.0 (4(b),(c))). 470 out of 540 nonhomologous samples possess both points (both excellent or good $\Delta p_{i,j}\leq 15\%$ (3)) around of their straight lines resulting from concurrent and efficacious use of the large and hydrophobic residue from primary to secondary structures via a doubly effective molecular mechanism.

Other 56 protein samples have a more efficacious and compensative employment of a residue type (just one $\Delta p_{i,j}\leq 15\%$) from primary to secondary structures by means of a singly effective mechanism. The 14 remaining samples possess a subtle employment of residues (both or one acceptable $\Delta p_{i,j}\lesssim 25\%$ in $L_j\lesssim 15$) so utilizing a partially effective mechanism. Consequently, no sample possesses native structures with a bad mechanism by both residue types (both together with bad dissimilarities, $\Delta p_{L,j}$ and $\Delta p_{H,j}$>15%). Far points from straight lines in the single and partial mechanisms of some protein samples have contributed for low linear coefficients R≈60.0% (4(a)–(e)).

The proportions $p_{H,e}$ of hydrophobic groups in strands had an unsatisfying straight line (with R<40% in $p_{H,e} = 61.9n_H+32.9$ (4f), N=223) that was only used to separate below/above $p_{H,e}$ points of it, and whose fits gave rise to $p^1_{H,e}/p^2_{H,e}$ (4(d)/(e)). This dual behavior of $p_{H,e}$ may simultaneously be due to long-range interactions into hydrophobic interplays [31], and non-local strands constituting β-pleated sheets [32–33].

The five linear relations $p_{i,j}$ ((4(a)–(e)) (Figure 3)) are dependent only of primary sequences (by n_L, n_H and N), and they will be validated by predictions in homologous protein samples similar to cross-validation assays in statistics [34]; though here is focused on a thorough case study by means of a rule-based approach ((3), (4(a)–(e))), so that it does not suffice to identify the occurrence and to determine the quantity of a type of mechanism, the protein names (PDB ID) should be precisely furnished whenever necessary. Furthermore, the four types of molecular mechanisms and their amounts should be confirmed, complemented or denied in the following more precise inspections for another detached sample set, the homologous proteins.

3.3. Molecular Interactions and Mechanisms in Homologous Proteins

In nonhomologous proteins (Figure 3), the empirical sequence-structure correlations between $p_{i,j}$ and n_i (4(a)–(e)) were determined, analyzed steric and hydrophobic interactions, and found out four types of molecular mechanisms. Such correlations as prediction rules are now employed to compute estimated $p_{i,j}$ that compared with real $p_{i,j}$ via the their dissimilarities $\Delta p_{i,j}$ (3) will permit us to survey molecular interactions and mechanisms in secondary structure elements of homologous proteins. Note that to reckon an estimated $p_{i,j}$, it suffices to know the primary sequence of the protein by the normalized quantities n_L or n_H.

The helical structures by means of $\Delta p_{i,h}$ (Figure 4), the module of dissimilarity between the real (2) and estimated (4(a),(c)) proportions of large and hydrophobic residues, are firstly inspected. In addition, a thorough analysis is proceeded in samples with troublesome dissimilarities ($\Delta p_{i,h}$>15%), so better known the individual occurrence of the steric and hydrophobic interplays and their acting mechanisms.

In 194 homologous samples with helix structures and their 388 values $\Delta p_{i,h}$ (Figure 4), 162 of them have both residue types (324 $\Delta p_{L,h}$ and $\Delta p_{H,h}\leq 15\%$) inside gray rectangles, and therefore making use of a doubly effective mechanism. In contrast, 31 once underlined samples possess only one (those of numbers 2, 12, 18...182, 183), already the twice underlined sample, number 45, with none of residue type $\Delta p_{i,h}$ (both acceptable $15\%<\Delta p_{i,j}\lesssim 25\%$ in L_h=7) inside gray rectangles work with singly and partially effective mechanisms, respectively. Among the once underlined samples, the steric interactions are better than the hydrophobic ones with 21 samples inside gray rectangles.

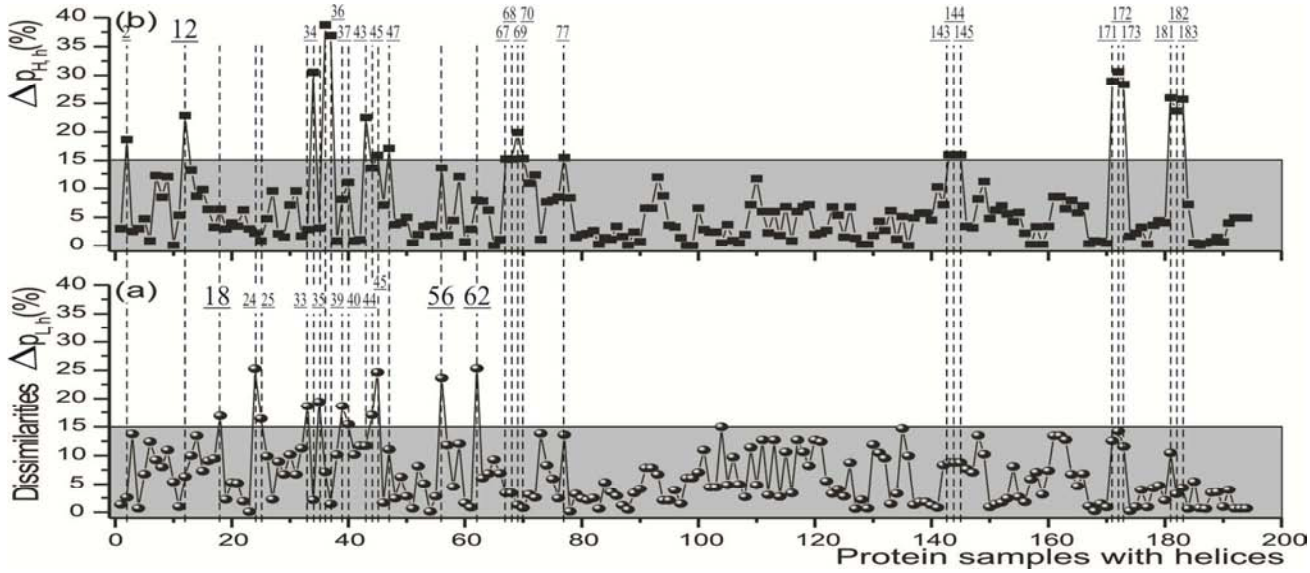

Figure 4. *388 module of dissimilarities for helical structures $\Delta p_{L,h}$ (a) and $\Delta p_{H,h}$ (b) in 194 homologous protein samples. The samples with both points inside gray rectangles, once and twice (number 45) underlined samples employ doubly, singly and partially effective mechanisms, respectively. Consecutive points with very near or equal $\Delta p_{L,h}$ and $\Delta p_{H,h}$ (e.g., samples of numbers 67, 68, 69, 70) often represent the same helix ensemble.*

After analyzing helices, a similar proceeding is assumed for 148 samples with strands and their 296 $\Delta p_{i,e}$ (Figure 5) between the real proportions (2) and the estimated proportions of large (4b) and hydrophobic (4(d),(e)) residues. The choice (4(d)/(e)) for estimated $p_{H,e}$ was based on low/high $p_{H,e}$ values, as used previously for nonhomologous proteins (Figure 3).

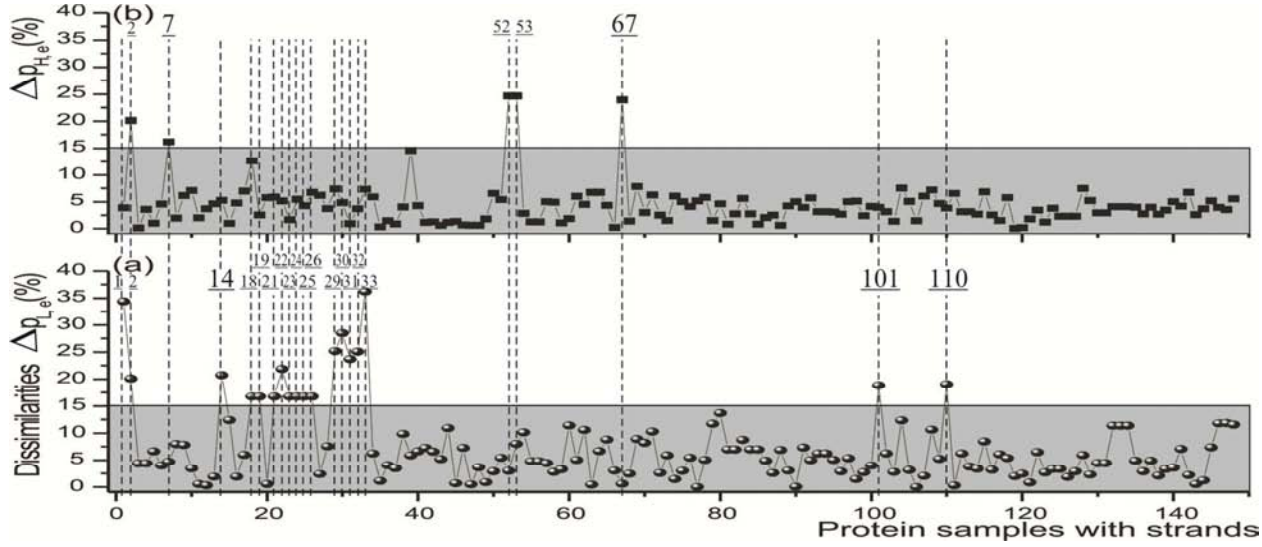

Figure 5. *296 module of dissimilarities for strand structures $\Delta p_{L,e}$ (a) and $\Delta p_{H,e}$ (b) in 148 homologous samples. The double, single and partial mechanisms are utilized by samples with both points inside gray rectangles, once and twice (number 2) underlined ones.*

126 out of a total of 148 samples with strand structures (Figure 5) possess both points inside gray rectangles with excellent or good predictions $\Delta p_{i,e}$ (252 points with $\Delta p_{i,e} \leq 15\%$), and consequently using a doubly effective mechanism. On the other hand, 21 samples with one (once underlined those of numbers 1, 7, 14…101, 110) and one sample with none (the twice underlined number 2 with both acceptable $\Delta p_{i,j}$ in L_e=7) point inside gray rectangles work with a singly and partially effective mechanism, respectively. In 17 out of 21 once underlined samples (except for those of numbers 7, 52, 53, 67 in Figure 5b), the hydrophobic interplays are more effective than the steric ones.

All the 342 homologous protein samples with helix and strand structures (Figures 4 and 5) use double, single or partial mechanisms by the steric and hydrophobic interactions, as disclosed by their 684 dissimilarities $\Delta p_{i,j}$. Now we pass to visually analyze homologous samples into conformational ensembles (like Figure 1), and perceive that different arrangements of amino acid residues from primary to secondary structures can have or not more than one type of mechanism in these ensembles (Figure 6), and therefore hypothesizing the stability levels of such ensembles.

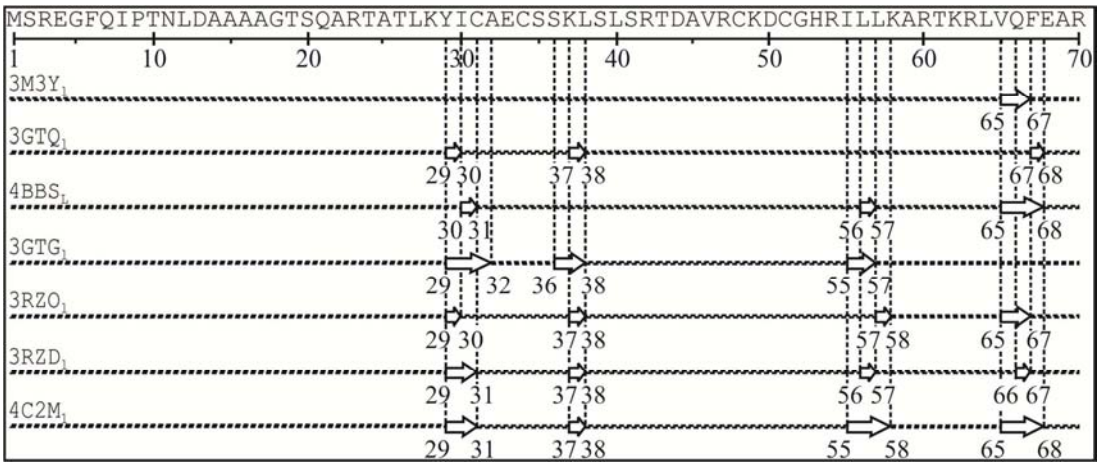

Figure 6. *Strand ensemble with seven 70-residue homologous transcription/transferase proteins that utilize two types of effective mechanisms, except for 3M3Yl and 3GTQl. In the 7 strand samples above, the first two cases (3M3Yl, 3GTQl) with $L_e \leq 6$ have not computed their mechanisms by $\Delta p_{i,e}$, and the 5 others correspond to samples of numbers 30 to 34 in Figure 5.*

Figure 6 shows that the particular disposition of amino acid residues in each strand segment leads to specific fulfillments of the steric and hydrophobic interactions ($\Delta p_{i,e}$) and by consequence the less or more stable forms given by the simply (in 4BBSL, 3RZOl, 3RZDl, 4C2M1) or doubly (3GTGl) effective mechanisms, respectively. In addition, these results evidences the intrinsic interaction instability in short strand lengths L_e that vary from 3 residues (3M3Y1) with one isolated strand not forming β-sheet to 13 residues (4C2M1) with four extended strands constituting two antiparallel β-sheets. In consequence of such instability, the need of the cutoff length $L_j>6$ for more precise measures in $p_{i,j}$ and $\Delta p_{i,j}$ previously adopted in this paper.

It is substantive to point out that for each conformational ensemble our rule-based approach allows to individually compute the strategic performance of steric and hydrophobic interactions (observing $\Delta p_{L,j}$ and $\Delta p_{H,j}$) in each native conformation as well as identifying the existence of one or more sorts of molecular mechanisms. Such interaction performance and detection of mechanisms are visualized in Figure 7 for a conformational ensemble with 18 homologous ribosomal proteins constituted by diversified helical segments.

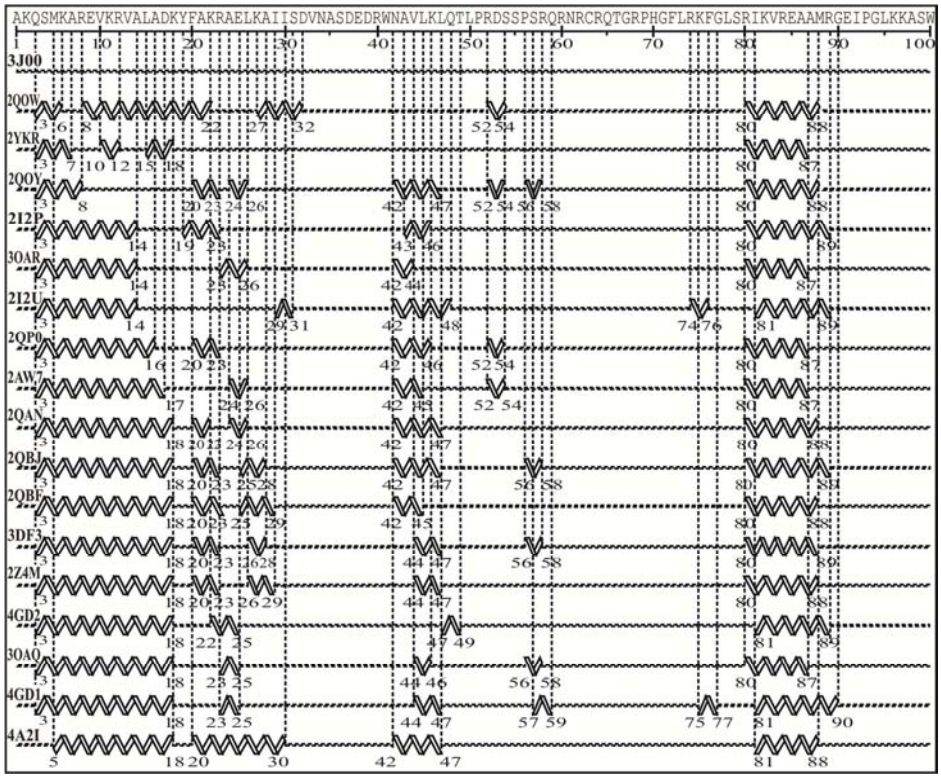

Figure 7. *Helix ensemble with eighteen 100-residue homologous samples that make use of doubly effective mechanisms excepting 3J00. The first sample 3J00 has $L_h=0$ and consequently without $\Delta p_{i,h}$. The sample numbers 15 (4GD2) and 16 (3OAQ) possess strand segments ($L_e>6$ not shown here) and utilize a simple and double mechanism, respectively.*

In Figure 7, four samples from numbers two to five (2QOW with $\Delta p_{L,h}$ and $\Delta p_{H,h}$ equal to 0.6% and 1.3%; 2YKR, 3.1%, 2.3%; 2QOY, 2.2%, 0.2%; 2I2P, 10.5%, 4.3%) use the steric and hydrophobic interactions with a subtle predominance of one on another considering that both have excellent or good performances ($\Delta p_{L,h}$ and $\Delta p_{H,h} \leq 15\%$). The interaction strengths for our 4 samples with helices should be extended for the 13 other homologous partners that also employ $\Delta p_{L,h}$ and $\Delta p_{H,h} \leq 15\%$, and therefore the 17 samples work of stable form with only the doubly effective mechanism. When used this individual numerical characterization in proteins of Figure 1, the helix ensemble of 4 homologous samples shows to possess one (1BQT with $\Delta p_{L,h}$ and $\Delta p_{H,h}$ equal to 23.7% and 13.6%) and three cases (1PMX, 11.8%, 1.9%; 2GF1, 4.5%, 4.5%; 3GF1, 12.1%, 12.0%) with simple and double mechanisms, respectively.

With regard to coupled acting of $\Delta p_{L,j}$ and $\Delta p_{H,j}$, the homologous (and nonhomologous) samples are comparable having in percentage 84 (87), 15 (10), 1 (3) and 0 (0) of the samples successfully working via doubly, simply, partially and badly effective molecular mechanisms, respectively. The quantitative agreement between both types of samples through three types of mechanisms indicate that proteins make use of an interactional plasticity, since depending of the sample the secondary structure elements of 3D native structures utilize either more stably both interactions by a large majority (>80%) of the cases, or less stably one or partially both interactions in a smaller amount of cases, <20%. Therefore, the singularity of sequence and plurality of structures in homologous proteins keep uniformly the interaction performances and four types of molecular mechanisms, in addition to validating the five rules $p_{i,j}$ ((4(a)–(e)), Figure 3) that were originated from the singularity of sequences and structures in nonhomologous proteins.

None of the 1764 inspections in nonhomologous and homologous samples (Figures 3–5) possess native state structures with a malfunctioning or bad mechanism by both residue types ($\Delta p_{L,j}$ and $\Delta p_{H,j} > 15\%$). The occurrence of a bad mechanism by $\Delta p_{L,j}$ and $\Delta p_{H,j}$ in a protein native conformation could indicate atypical interactional behaviors that would call for more inquiry, such as occurrences of specific interaction among other parts of the protein or with another macromolecule, or still a direct influence of other molecular interactions biasing the steric and hydrophobic driving forces measured by $\Delta p_{i,j}$. Although, this unsatisfactory mechanism is plausible to happen of relevant and measurable form by $\Delta p_{i,j}$ in conformation ensembles of denatured states and folding intermediates during events of protein dynamics. In the folding and dynamics processes and others such as protein design and evolution, our rule-based approach by both $\Delta p_{L,j}$ and $\Delta p_{H,j}$ in four mechanism types can be a useful tool for investigating the strategic power and nature of steric and hydrophobic forces.

The current approach based on two coarse-grained models is insufficient for sharper measures of the secondary structure composition, as traditionally occur in these types of models [35–36], and in consequence other approaches, such as semi-empirical ones, or via other higher resolution levels with more letter codes or atomic models, should be evaluated. However, the detailed protein approaches are also limited in many features, since they frequently demand too many computational resources and details of molecular interactions and cellular environments that they use or try to catch are still not fully understood [37–38]. In summary, our current low-resolution approach is a suitable instrument to succeed at capturing pivotal insights and principles of homologous proteins when quantitatively accurate estimations and systematic investigations are needed and furthermore particular details can be suppressed.

4. Conclusion

Numerous studies have examined homology-derived proteins by template-based methods via search optimization for sequence-sequence comparisons, multiple sequence and sequence-structure alignments that incorporate information about protein families or folds [39–44] and have been utilized in several investigations, including homology inference, structure modeling, functional prediction and phylogenetic analysis. In the sequence-structure context despite expressive research efforts, some relevant questions as the key acting of fundamental interactions (e.g., steric and hydrophobic/hydrophilic ones), the driving mechanisms resulting from these interactions as well as their implications for analyzing conformation ensemble in homologous proteins are not fully understood, and therefore such questions have been analyzed here by means of a rule-based approach.

Firstly, nonhomologous proteins showed a direct synchronism between the employment ($p_{i,j}$) in folded structures with the availableness (n_i) from primary structures by the steric and hydrophobic interactions through five empirical linear relations $p_{i,j}$ (Figure 3) that modulate different strategies employed by the residue volume and hydrophobicity. Then, when used as prediction rules in homologous proteins, such linear relations inside modules of dissimilarities $\Delta p_{i,j}$ (in 684 $\Delta p_{i,j}$, Figures 4–5) measure the strengths of the individually steric and hydrophobic interactions, check the stability level by both coupled interactions (looking for $\Delta p_{L,j}$ and $\Delta p_{H,j} \leq 15\%$), identify four types (obtaining 84% double, 15% single, 1% partial and 0 bad) of molecular mechanisms for homologous protein, as well as we can visualize the occurrence of one or more type of these mechanisms in helix and strand ensembles (Figures 1, 6–7) of native conformations.

In summary, taken together our 1764 inspections intend to contribute with better criteria in the conformational ensemble selection, capture protein fundamental aspects and furnish structural-interactional insights for native conformations of homologous proteins, as well as support the inference that our rule-based approach can potentially to be applied to study other proteins and to better understand conformational

variability in biomolecular processes such as protein evolution, design, dynamics and folding [29, 45–49]. Such inspections obtained via two coarse-grained models work complementally with other results from simplified approaches, including misfolding and unfolding events [50], comparative modeling to explore protein-like features [51], lattice models for protein folding [52], energy landscape mapping methods for structure predictions [53], and evaluation of knots in proteins [54]. Furthermore our approach intend to join with other tools and resources [55–58] to help researches in the protein sequence-structure correlations and to pave the way for improving the general understanding of conformational ensembles in further proteins.

Supplementary Materials

In the present paper, the homologous proteins have been inspected according to their secondary structure compositions that form conformational ensembles with different sizes and component quantities (Table 1). Such helix and strand ensembles were displayed and analyzed in Figures 1, 4, 7 and Figures 5, 6, respectively.

Table 1. *List with conformational ensemble sizes their total quantities of homologous proteins and respective PDB identifications scanned in this paper. In third column, it is only shown protein chains with L_j>6, different ensembles into each size are separated by semicolons, and chains of Figures 1, 6 and 7 are pointed out.*

Ensemble size	Protein total quantity	Homologous proteins (PDB IDs)
2	102	1FC3$_a$,1FC3$_b$; 1HLO$_a$,1HLO$_b$; 1ICA,1L4V; 1LQ1$_a$,1LQ1$_c$; 1O3Q,1O3T; 1UTG,2UTG; 1VOQ,1PNS$_s$; 1WU9$_a$,1WU9$_b$; 1XF5$_l$,1XF5$_m$; 1YUG,3TGF; 1Y04,1Y03; 2COB,2COO; 2H8C$_a$,2H8C$_c$; 2IWO$_a$,2IWO$_b$; 2JPK,2JPM; 2J0N$_a$,2J0N$_b$; 2K7Y,1VPU; 2LTT,2LTD; 2M4Z,1MB6; 2OIH,1VC5; 2O52$_a$,2O52$_b$; 2O8M$_a$,2O8M$_b$; 2O97,1MUL; 2WW9$_b$,2WWA$_b$; 2XTC$_a$,2XTC$_b$; 2XZE$_q$,2XZE$_r$; 2ZA4$_b$,2ZA4$_d$; 3B5N$_k$,3B5N$_c$; 3FB9$_a$,3FB9$_b$; 3FCG$_a$,3FCG$_b$; 3FYL,3G6P$_b$; 3H8M$_a$,3H8M$_b$; 3IM3,2EZW; 3I00$_a$,3I00$_b$; 3J44,3J21$_P$; 3J5W,1RQU; 3OAR$_p$,3OFY$_p$; 3PYR$_u$,3PYV$_u$; 3P5T$_l$,3P5T$_m$; 3V6W$_5$,4G5N$_5$; 3ZWH$_q$,4CFQ; 4B6X$_a$,4B6X$_b$; 4EMP$_v$,4EMP$_g$; 4FZ0,1LMM; 4HBM$_a$,4HBM$_d$; 4HWD$_d$,4HWD$_b$; 4H13,2D2C; 4H6U$_a$,4H6U$_b$; 4JKR,3IYD; 4KJ7$_h$,3J5O; 4UJH$_H$,4UJW$_H$
3	27	1AML, 1BA4, 1BA6;1FDM, 2C0W, 2CPB;1QN4, 1QN7, 1QNC;1TCP, 1KIG, 1TAP;2LUQ, 2LUP, 2LBS;2VTX$_a$, 2VTX$_b$, 2VTX$_j$;2X78$_a$, 2X78$_b$, 2X78$_c$;3IGK, 3KZ8, 3D0A;4BWG$_b$, 4BWG$_c$, 4BWG$_j$
4	32	1BQT, 1PMX, 2GF1, 3GF1 (Figure 1); 3U5E$_h$, 2WW9$_N$, 2WWA$_N$, 2WWB$_N$; 4BBS$_j$, 3M3Y$_j$, 3GTG$_j$, 3GTQ$_j$; 4BBS$_K$, 3M4O, 3H3V, 2NVT$_K$; 4IBU, 4IBV, 4IBW, 3KZ8; 4KFK$_r$, 4KHP$_r$, 3PYN$_r$, 3PYS$_r$; 4L6J$_1$, 4K0Q$_4$, 4KFL$_4$, 3PYV$_1$; 4UJF$_V$, 4UJH$_y$, 4UJM$_y$,4UJP$_V$
5	15	3J19$_l$, 3OFQ$_l$, 3FIK$_l$, 4TOV, 4TP7; 3PIO$_o$, 3PIP$_o$, 3DLL$_o$, 2ZJR, 1YL3; 4C2M$_l$, 4BBS$_L$, 3RZD$_l$, 3RZO$_l$, 3GTG$_l$ (Figure 6)
6	6	4WAP, 4WAR, 4TP7$_N$, 3R8T$_N$, 3OAS$_N$, 3OAT$_N$
7	7	3V6W$_3$, 4EJB$_3$, 3V23$_3$, 3UYE$_x$, 3UZK$_x$, 2XTG$_3$, 3D5B$_3$
9	9	4GAR, 4GAU, 3SGF, 3UOS, 3J01$_t$, 3ORB$_t$, 3I1R$_t$, 2AW4$_t$, 2AWB$_t$
10	10	4GD1$_f$, 4GD2$_f$, 4ADV, 2YKR$_f$, 3OAQ$_f$, 3OAR$_f$, 3OFA$_f$, 3OFB$_f$, 3OFO$_f$, 3OFP$_f$
11	11	4GD1$_q$, 4GD2$_q$, 4A2I$_q$, 2YKR$_q$, 3OAQ$_q$, 3OAR$_q$, 3OFA$_q$, 3OFB$_q$, 3OFX$_q$, 3OFY$_q$, 3OFP$_q$
12	12	4B3S$_n$, 4B3T$_n$, 4L6K$_n$, 4L6M$_n$, 4KDJ$_n$, 4KFH$_n$, 4K0K$_n$, 4KHP$_n$, 4AQY$_n$, 3PYN$_n$, 1PNS$_n$, 1VOS$_n$
17	17	4GD1$_n$, 4GD2$_n$, 4A2I$_n$, 2YKR$_n$, 3OAQ$_n$, 3OAR$_n$, 3DF3$_n$, 2QAN$_n$, 2QBF$_n$, 2QBJ$_n$, 2Z4M$_n$, 2QOW$_n$, 2QOY$_n$, 2QP0$_n$, 2I2P$_n$, 2I2U$_n$, 2AW7$_n$ (Figure 7)

The results considered 248 homologous proteins (total sum in second column of Table 1) comprising conformational ensembles in 11 sizes (line numbers in Table 1) and a total of 78 ensembles (amount of second divided by first column in lines of Table 1) segregated by semicolons. Some pair combinations ($C_{n,2}$ (1)) of n homologous proteins possess dissimilarities in both helices and strands, others in helices or strands, so that the proteins are segregated in 194 samples with helix ensembles (Figure 4) and 148 ones with strand ensembles (Figure 5) totalizing 684 dissimilarities $\Delta p_{i,j}$.

Inside each ensemble (third column in Table 1), the protein pair combinations have 100% primary sequence identity or none different residue. However, five (toxin proteins 2M4Z with 1MB6, antibacterial 1ICA with antibiotic 1L4V, viral 1FDM-2C0W, DNA binding 4IBU-4IBW, DNA binding 4IBV-4IBW) and seven (viral 2K7Y-aids 1VPU, 2C0W-virus 2CPB, 4IBU-transcription 3KZ8, 4IBV-3KZ8, 4IBW-3KZ8, transcription 3IGK-3KZ8, 3KZ8-transcription 3D0A) of these combinations possess one and two different residues, respectively.

References

[1] Z. Dosztányi, C. Magyar, G. E. Tusnády, M. Cserzõ, A. Fiser, I. Simon. Servers for sequence–structure relationship analysis and prediction. Nucleic Acids Res. 31(13), 2003, 3359–3363.

[2] D. J. Selkoe. Folding proteins in fatal ways. Nature 426, 2003, 900–904.

[3] K. A. Dill, S. B. Ozkan, M. S. Shell, T. R. Weikl. The Protein Folding Problem. Annu. Rev. Biophys. 37, 2008, 289–316.

[4] Z. Zhu-Qing. Folding Mechanism of De novo Designed Proteins. Acta Phys.-Chim. Sinic. 28(10), 2012, 2381–2389.

[5] Z.-R. Xie, J. Chen, Y. Wu. A coarse-grained model for the simulations of biomolecular interactions in cellular environments. J. Chem. Phys. 140, 2014, 054112–1–12.

[6] L. F. O. Rocha. Quantifying Steric and Hydrophobic Influence of Non-Standard Amino Acids in Proteins That Undergo Post-Translational Modifications. Biochem. Mol. Biol. 2(2), 2017, 12–24.

[7] A. G. Murzin, S. E. Brenner, T. Hubbard, C. Chothia. SCOP: A Structural Classification of the Protein Database for the Investigation of Sequences and Structures. J. Mol. Biol. 247, 1995, 536–540.

[8] N.-K. Kim, J. Xie. Protein Multiple Alignment Incorporating Primary and Secondary Structure Information. J. Comp. Biol. 13(9), 2006, 1615–1629.

[9] E. Krissinel. On the relationship between sequence and structure similarities in proteomics. Bioinformatics 23(6), 2007, 717–723.

[10] Y. Wang, R. I. Sadreyev, N. V. Grishin. PROCAIN server for remote protein sequence similarity search. Bioinformatics 25(16), 2009, 2076–2077.

[11] M. Kosloff, R. Kolodny. Sequence-similar, structure-dissimilar protein pairs in the PDB. Proteins 71, 2008, 891–902.

[12] K. Arnold, L. Bordoli, J. Kopp, T. Schwede. The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling. Bioinformatics 22(2), 2006, 195–201.

[13] J. Pei, B.-H. Kim, M. Tang, N. V. Grishin. PROMALS web server for accurate multiple protein sequence alignments. Nucleic Acids Res. 35, 2007, W649–W652.

[14] Y.-C. Hsu, C.-M. Chen, T.-W. Pai, J.-F. Jeng, C.-H. Hu, W.-S. Tzou. Length Encoded Secondary Structure Profile for Remote Homologous Protein Detection. A. Hua, S.-L. Chang (Eds.): ICA3PP 2009, LNCS 5574, Springer-Verlag Berlin Heidelberg 2009, 1–11.

[15] N. Echols, D. Milburn, M. Gerstein: MolMovDB: analysis and visualization of conformational change and structural flexibility. Nucleic Acids Res. 31, 2003, 478–482.

[16] M. Gerstein, N. Echols. Exploring the range of protein flexibility, from a structural proteomics perspective. Curr. Opin. Chem. Biol. 8, 2004, 14–19.

[17] S. K. Mani, H. Balasubramanian, S. Nallusamy, S. Samuel. Sequence and structural analysis of two designed proteins with 88% identity adopting different folds. PEDS 23(12), 2010, 911–918.

[18] C. O. Nwamba, K. Ibrahim. The Role of Protein Conformational Switches in Pharmacology: Its Implications in Metabolic Reprogramming and Protein Evolution. Cell Biochem. Biophys. 68, 2014, 455–462.

[19] H. Berman, K. Henrick, H. Nakamura. Announcing the worldwide Protein Data Bank. Nat. Struct. Biol. 10(12), 2003, 980.

[20] C. Chothia, A. M. Lesk. The relation between the divergence of sequence and structure in Proteins. EMBO J. 5(4), 1986, 823–826.

[21] A. Baruah, P. Biswas. The role of site-directed point mutations in protein misfolding. Phys. Chem. Chem. Phys. 16, 2014, 13964–13973.

[22] L. J. Fox. An Introduction to Mathematical Probability. Chapter 2. Human Nature, Biology, and Social Structure: A Critical Look at What Science Can Tell Us About Society. Vol. V, 1987.

[23] E. G. Hutchinson, J. M. Thornton. PROMOTIF--A program to identify and analyze structural motifs in proteins. Protein Sci. 5, 1996, 212–220.

[24] E. S. Huang, S. Subbiah, M. Levitt. Recognizing Native Folds by the Arrangement of Hydrophobic and Polar Residues. J. Mol. Biol. 252, 1995, 709–720.

[25] M. Gerstein, C. Chothia. Packing at the protein-water interface. Proc. Natl. Acad. Sci. USA 93, 1996, 10167–10172.

[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] A. A. Zamyatnin. Protein volume in solution. Progr. Biophys. Mol. Biol. 24, 1972, 107–123.

[28] S. Moelbert, E. Emberly, C. Tang. Correlation between sequence hydrophobicity and surface-exposure pattern of database proteins. Protein Sci. 13, 2004, 752–762.

[29] 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.

[30] K. H. DuBay, G. R. Bowman, P. L. Geissler. Fluctuations within Folded Proteins: Implications for Thermodynamic and Allosteric Regulation. Acc. Chem. Res. 48, 2015, 1098−1105.

[31] M. U. Hammer, T. H. Anderson, A. Chaimovich, M. S. Shell, J. Israelachvili. The search for the hydrophobic force law. Faraday Discuss. 146, 2010, 299–308.

[32] M. M. Gromiha, S. Selvaraj. Protein secondary structure prediction in different structural classes. Protein Engng. 11(4), 1998, 249–251.

[33] B. Wathen, J. Zongchao. A hierarchical order within protein structures underlies large separations between strands in β-sheets. Proteins 81, 2013, 163–175.

[34] H.-C. Chung, C.-P. Han. Conditional confidence intervals for classification error rate. Comput. Statist. Dat. Analys. 53, 2009, 4358–4369.

[35] S. Moreno-Hernández, M. Levitt. Comparative modeling and protein-like features of hydrophobic–polar models on a two-dimensional lattice. Proteins 80, 2012, 1683–1693.

[36] W. G. Noid. Perspective: Coarse-grained models for biomolecular systems. J. Chem. Phys. 139, 2013, 090901–1–25.

[37] J. Hu, T. Chen, M. Wang, H. S. Chan, Z. Q. Zhang. A critical comparison of coarse-grained structure-based approaches and atomic models of protein folding. Phys. Chem. Chem. Phys. 19(21), 2017, 13629–13639.

[38] H. C. Li, Y. Y. Chang, J. Y. Lee, I. Bahar, L. W. Yang. DynOmics: dynamics of structural proteome and beyond. Nucleic Acid Res. 45(W1), 2017, W374–W380.

[39] R. Schneider, A. de Daruvar, C. Sander. The HSSP database of protein structure–sequence alignments. Nucleic Acids Res. 25(1), 1997, 226–230.

[40] C. Guda, L. R. Pal, I. N. Shindyalov. DMAPS: a database of multiple alignments for protein structures. Nucleic Acids Res. 34, 2006, D273–D276.

[41] A. Zemla, B. Geisbrecht, J. Smith, M. Lam, B. Kirkpatrick, M. Wagner, T. Slezak, C. E. Zhou. STRALCP—structure alignment-based clustering of proteins. Nucleic Acids Res. 35(22), 2007, e150–1–8.

[42] J. Pei, M. Tang, N. V. Grishin. PROMALS3D web server for accurate multiple protein sequence and structure alignments. Nucleic Acids Res. 36, 2008, W30–W34.

[43] R. C. Edgar. Quality measures for protein alignment benchmarks. Nucleic Acids Res. 38(7), 2010, 2145–2153.

[44] D. Gront, M. Blaszczyk, P. Wojciechowski, A. Kolinski. BioShell Threader: protein homology detection based on sequence profiles and secondary structure profiles. Nucleic Acids Res. 40, 2012, W257–W262.

[45] J. Wang, W. Wang. Simplification of complexity in protein molecular systems by grouping amino acids: a view from physics. Adv. Phys.-X. 1 (3), 2016, 444–466.

[46] R. L. Baldwin, G. D. Rose. Is protein folding hierarchic? II. Folding intermediates and transition states. TIBS 24, 1999, 77–83.

[47] L. C. James, D. S. Tawfik. Conformational diversity and protein evolution – a 60-year-old hypothesis revisited. Trends Biochem. Sci. 28(7), 2003, 361–368.

[48] V. N. Uversky, A. K. Dunker. Understanding protein non-folding. Biochim. Biophys. Acta 1804, 2010, 1231–1264.

[49] A. Perez, A. Roy, K. Kasavajhala, A. Wagaman, K. A. Dill, J. L. MacCallum. Extracting representative structures from protein conformational ensembles. Proteins 82, 2014, 2671–2680.

[50] A. Kumar, A. Baruah, P. Biswas. Role of local and nonlocal interactions in folding and misfolding of globular proteins. J. Chem. Phys. 146 (6), 2017, 065102.

[51] E. Ferrada. The Amino Acid Alphabet and the Architecture of the Protein Sequence-Structure Map. I. Binary Alphabets. Plos Comp. Biol. 10(12), 2014, e1003946–1–20.

[52] D. L. Shaw, A. S. M. S. Islam, M. S. Rahman, M. Hasan. Protein folding in HP model on hexagonal lattices with diagonals. BMC Bioinformatics 15 (Suppl 2), 2014, 1–13.

[53] A. G. Citrolo, G. Mauri. A local landscape mapping method for protein structure prediction in the HP model. Nat. Comput. 13, 2014, 309–319.

[54] T. Wüst, D. Reith, P. Virnau. Sequence Determines Degree of Knottedness in a Coarse-Grained Protein Model. Phys. Rev. Lett. 114, 2015, 028102–1–5.

[55] C. R. Terrell, L. L. Listenberger. Using molecular visualization to explore protein structure and function and enhance student facility with computational tools. Biochem. Mol. Biol. Educ. 45(4), 2017, 318–328.

[56] D. Vlachakis, A. Armaos, S. Kossida. Advanced Protein Alignments Based on Sequence, Structure and Hydropathy Profiles; The Paradigm of the Viral Polymerase Enzyme. Math. Comput. Sci. 11(2), 2017, 197–208.

[57] L. F. O. Rocha. Analysis of molecular structures and mechanisms for toxins derived from venomous animals. Comput. Biol. Chem. 61, 2016, 8–14.

[58] M. L. Hekkelman, G. Vriend. MRS: a fast and compact retrieval system for biological data. Nucleic Acids Res. 33, 2005, W766–W769.