Journal of Molecular Structure 1063 (2014) 242–250

Contents lists available at ScienceDirect

Journal of Molecular Structure

journal homepage: www.elsevier.com/locate/molstruc

Toward a better understanding of structural divergences in proteins using different secondary structure assignment methods

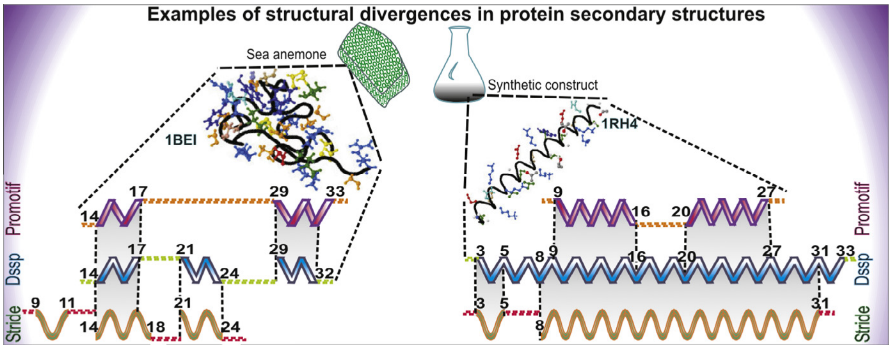
L.F.O. Rocha [*]

Department of Physics and Chemistry, Faculty of Pharmaceutical Sciences at Ribeirão Preto, University of São Paulo, Av. do café, s/n, 14040-903 Ribeirão Preto, São Paulo, Brazil

HIGHLIGHTS

- Our investigation makes 116 structural evaluations within a template group.
- 88 stereochemical predictions for target subgroups display high success amounts.
- 42 comparisons between three methods show high compatibility scores between them.
- We identify a triple molecular mechanism by steric and hydrophobic interactions.
- The structural divergences in proteins are better understood and appreciated.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 22 October 2013
Received in revised form 24 January 2014
Accepted 24 January 2014
Available online 29 January 2014

Keywords:
Nonconsensual assignment
Methods DSSP, PROMOTIF, STRIDE
Secondary structure definition
Protein stereochemistry
Structure comparison

ABSTRACT

Structural disagreements on the location and quantity of secondary structure segments comprise a current challenging problem leading to several limitations for theoretical and applied research. This paper presents 116 structural evaluations by steric and hydrophobic interactions in secondary structures within a specific template group; determines simple prediction rules that calculate 88 occurrence frequencies of large and hydrophobic residues into target intra- and inter-subgroups with structure disagreements; and utilizes 42 comparisons between the methods PROMOTIF, DSSP and STRIDE. In the stereochemical predictions inside the subgroups there are predominantly excellent and/or good success amounts with their expected values, and the disclosure of a triple molecular mechanism by residue volumetric and hydrophobic ingredients. The method comparisons show high compatibility scores between them, therefore validating their seemingly incompatible assignments. Thus, the nonconsensual ascriptions are better understood and appreciated. Furthermore, such results suggest a broad utility of our assignment method for other benchmark datasets and known methods.

1. Introduction

Indispensable to existence of living organisms, peptides and proteins are the most complex known molecules having multidimensional structures and often undergoing co- or post-translational modifications [1]. They exist in isolated forms, complex mixtures, or as a part of multi-component systems; are the building blocks of cells until organs; and can have two or more different biological roles [2]. Their specialized functions are dependent on their particular native three-dimensional (3D) shapes constituted by strategical secondary structure elements that in turn are dictated by the physicochemical properties and compositions of the amino acid residue sequences [3], in proper physiological environments. Analyses from secondary structures have been successfully utilized in several studies, such as fold-recognition procedure [4],

[*] Tel.: +55 16 3602 4169; fax: +55 16 3602 4880.
E-mail address: luiz@fcfrp.usp.br

http://dx.doi.org/10.1016/j.molstruc.2014.01.060
0022-2860/© 2014 Elsevier B.V. All rights reserved.

conformation control [5], and tertiary structure prediction methods [6].

The assignments of native (non-)repetitive structures are usually made employing various automatic secondary structure assignment methods (SSAMs), based on different definitions and distinct characteristics of these structures [7]. SSAMs handle atomic coordinate data and residue sequences of experimentally solved macromolecules from X-ray crystallography and nuclear magnetic resonance spectroscopy, for instance. Among the SSAMs, DSSP, PROMOTIF and STRIDE have been highlighted and widely utilized. DSSP [8] assigns secondary structural segments based solely on the hydrogen bonding pattern from the geometrical features and an electrostatic model, thus attributing eight states: alpha-, 3_{10}-, pi-helix, isolated and extended strand, hydrogen-bonded turn, curved bend, tight loop, and the remaining residues are outside of secondary elements. A number of software programs make use of DSSP definitions; e.g., RasMol [9] and Gromacs [10].

PROMOTIF [11] is also based on the hydrogen binding design similar to the DSSP approach, but with a slightly modified algorithm and small alterations at the extremities of helices and strands, where possible. This method is a key tool in the PDBsum database [12], and provides information regarding structural features, including interactions and geometry of α-, 3_{10}-, π-helices, β-strands, β-, γ-turns, β-bulges, β-hairpins, α-β-α units and ψ-loops. STRIDE [13] has been used by molecular visualization programs, such as VMD [14], and attributes secondary structure types using backbone torsional angles and hydrogen bonds that are detected from an empirical energy function, thus distinguishing seven states: α, 3_{10}, π helix, extended, isolated strand, turn, and coil. The several secondary structural states from the SSAMs are usually reduced to three elementary classes: helix, strand and coil.

In general, the assignments of secondary structures simultaneously assessed by various SSAMs display an extensive consensus, mainly in spiral helices and extended strands of β-sheets, which are the common and most prominent structural building blocks by different methods. However, sometimes 3D polypeptide conformations present interestingly considerable discrepancies concerning the location, quantity, and length of secondary structure fragments, especially in the boundaries between secondary states, mobile regions, and chain ends [15], which provoke fascinating and challenging matter subject to questioning. Nonetheless, in almost all cases no deeper evaluations and implications are frequently provided, other than arguments of intrinsic incompatibilities between methods, difficulty by non-ideal secondary arrangements, or due to different definitions and criteria employed in each SSAM. Consensus approaches generally propose either to ascribe to each residue the state assigned by the majority of the SSAMs [16] or to use standard-of-truth diagnoses provided from experimentalists' attributions [13,17].

In the secondary state ascriptions, structural divergences create additional theoretical difficulties and applied limitations for biological macromolecule analyses and employments, such as in surveys and predictions by *ab initio* methods, and in specific biophysicochemical characterizations and applications of functional native conformations. These divergences raise uncertainty about the quality of the SSAMs used and which outcomes to employ; hence, the necessity and importance of a better and deeper understanding of such divergences from in-depth inspections and case studies, including computational, empirical and mathematical approaches, as made here.

2. Materials and methods

2.1. Benchmark dataset, secondary structures, and residue ingredients

Currently in the post-genomic era, there have been a reasonable and growing number of experimentally determined polypeptides in each given extension, and the first step of this work is the careful selection of these polypeptides for our benchmark dataset. Initially, one imposes some selection conditions for obtaining a well-defined and unbiased dataset: (i) to examine only unrelated peptide/protein polymers of equal residue or bead numbers [18], thereby avoiding direct comparisons among chains of diversified extensions; (ii) the non-redundant dataset must contain samples with less than 25% of homology with the others, or otherwise reasonably different secondary structure fragments; and (iii) the selected native conformations should explicitly have many non-consensual assignments and structure divergence of at least three residues between SSAMs.

After breaking down several chain extensions and taking as basis the three pre-stated conditions i–iii above, this study opted for 35-residue natively folded macromolecules recruited from the Protein Data Bank, PDB [19], resulting in a benchmark dataset containing 55 nonstatistical samplings. Utilizing these particular samplings, we studied the cylindrical helices and β-pleated sheets [20] by means of the two primary and indispensable amino acid ingredients [21], volume and hydrophobicity. Furthermore, such samplings will make detachedly up the template group to numerically quantify the sequence–structure relationships, and the target subgroups with structural divergences. Although we chose only 35-residue macromolecules, our results, shown in the next section, are suitably applicable to other peptide and small protein datasets and known SSAMs, but this fact is out of scope of the current report and will be pertinently revealed in elsewhere.

The macromolecular samples of the dataset have 20 genetically encoded and nine non-proteinogenic amino acids that are independently classified into a reduced alphabet: large or small, L or S [22], and hydrophobic or polar, H or P [23], representing their volumes and hydrophobicities, respectively. The standard amino acids (single letters) and by similarity to them the non-proteinogenic unities (three letters) are: large-hydrophobic (F, H, I, L, M, V, W, Y; iil (allo-isoleucine), nle (norleucine)), large-polar (E, K, Q, R), small-hydrophobic (A, C, P, T; aba (alpha-aminobutyric acid, a parent of the alanine), dpr (D-proline), pca (pyroglutamic acid)) and small-polar (D, G, N, S; ace (acetyl group), dnp (3-amino-alanine), nh2 (amino group), sin (succinyl)). The nine non-standard unities are too few, totaling 26 residues into 16 modified polypeptides, for a total of 1925 residues into 55 polypeptides. In the 20 standard amino acids there is predominance of large and hydrophobic constituents (both separately having 12 units); therefore, these two major constituents are taken as reference and used hereafter.

2.2. Mathematical formulations, structural disagreements, and anticipated accuracies

In the template group of our benchmark dataset, any polypeptide chain has its primary sequence with total numbers n_i of large (L) and hydrophobic (H) residues, where $i = \{L \text{ or } H\}$ in the primary and secondary structures. Strategic portions of these n_i residues somehow belong to helices (h) and/or strands (e) in specific and nontrivial percentage proportions $p_{i,j}$, where $j = \{h \text{ or } e\}$. Henceforth, the subscript indices i and j refer to the amino acid ingredients and secondary structure types, respectively. Taking as an example the PDB code 1e4r (having $n_L = 17$ and $n_H = 18$) with strands from PROMOTIF (Fig. 1a), a total length L_e of 11 residues with expected occurrence frequencies $t_{L,e}$ of seven large and $t_{H,e}$ equal to six hydrophobic residues are observed, consequently the proportions of these residues are $p_{L,e} = 63.6\%$ and $p_{H,e} = 54.5\%$, derived from the following generic algebraic equation:

$$p_{i,j} = (t_{i,j}/L_j)100 \quad (1)$$

where for 1e4r, the indexes i and j represent L, H and e, respectively.