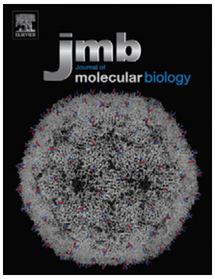
Article



Compact Structure Patterns in Proteins

**Bhadrachalam Chitturi[1,2,4,†], Shuoyong Shi[2,†],
Lisa N. Kinch[3] and Nick V. Grishin[2,3]**

*1 - **Department of Computer Science and Engineering,** Amrita School of Engineering, Amritapuri, Amrita Vishwa Vidyapeetham, Amrita University, India*
*2 - **Departments of Biophysics and Biochemistry,** University of Texas Southwestern Medical Center, 5323 Harry Hines Boulevard, Dallas, TX 75390-9050, USA*
*3 - **Howard Hughes Medical Institute,** University of Texas Southwestern Medical Center, 5323 Harry Hines Boulevard, Dallas, TX 75390-9050, USA*
*4 - **Departments of Computer Science,** University of Texas at Dallas, Richardson, TX 75083, USA*

***Correspondence to Nick V. Grishin:** Departments of Biophysics and Biochemistry, University of Texas Southwestern Medical Center, 5323 Harry Hines Boulevard, Dallas, TX 75390-9050, USA. grishin@chop.swmed.edu*

http://dx.doi.org/10.1016/j.jmb.2016.07.022
Edited by Sarah A. Teichmann

Abstract

Globular proteins typically fold into tightly packed arrays of regular secondary structures. We developed a model to approximate the compact parallel and antiparallel arrangement of α-helices and β-strands, enumerated all possible topologies formed by up to five secondary structural elements (SSEs), searched for their occurrence in spatial structures of proteins, and documented their frequencies of occurrence in the PDB. The enumeration model grows larger super-secondary structure patterns (SSPs) by combining pairs of smaller patterns, a process that approximates a potential path of protein fold evolution. The most prevalent SSPs are typically present in superfolds such as the Rossmann-like fold, the ferredoxin-like fold, and the Greek key motif, whereas the less frequent SSPs often possess uncommon structure features such as split β-sheets, left-handed connections, and crossing loops. This complete SSP enumeration model, for the first time, allows us to investigate which theoretically possible SSPs are not observed in available protein structures. All SSPs with up to four SSEs occurred in proteins. However, among the SSPs with five SSEs, approximately 20% (218) are absent from existing folds. Of these unobserved SSPs, 80% contain two or more uncommon structure features. To facilitate future efforts in protein structure classification, engineering, and design, we provide the resulting patterns and their frequency of occurrence in proteins at: http://prodata.swmed.edu/ssps/.

Introduction

The topological connectivity and arrangement of secondary structure elements (SSEs) in three-dimensional (3D) space define protein folds. Identifying and enumerating common substructure motifs within folds, such as the helix-turn-helix [1], the βαβ [2], and the Greek key [3,4], have aided in predicting protein structure [5–9] and function [1,10,11] and in understanding fold evolution [12,13]. These named substructures represent different super-secondary structure patterns (SSPs) that encompass two or more closely packed SSEs. SSPs can be defined by the order, connection topology, orientation, and packing of SSEs.

The knowledge of SSP composition within folds helps us understand the structural and evolutionary relationships among proteins. Previous studies have revealed a number of frequently recurring SSPs within diverse folds, such as the Rossmann fold [14], the β-grasp [15], and the Greek key [3,16], and have established the value of using these common SSPs to outline structural relationships among large families. Databases such as CATH [17] and SCOP

0022-2836/© 2016 Published by Elsevier Ltd.

[18] have provided large-scale classification of protein structures according to these relationships. For instance, SCOP describes folds by conserved combinations of SSEs in the common structure core.

The SSPs that occur with high frequency confirmed some basic rules of protein folding. For example, an investigation of crossover connections in β-sheets highlighted a strong preference for right-handedness in βαβ units [19], while enumeration of β-sheet structures detected the absence of sheets with order 3142 or 2413, known as the “pretzels” [20]. Distributions of open β-sheets have suggested a preference for a lower number of β-strand pairs adjacent in sequence but separated in the β-sheet, that is, “jumps” [6]. Recently, the ability of these rules to dictate the probability of SSP occurrence in two-layer architectures (i.e., consisting of two planes) was evaluated, helping explain the limited number of SSE arrangements seen in protein structures [21]. These rules can also aid in protein engineering and design. For example, fundamental rules such as chirality and orientation preference of SSPs, along with additional rules such as the angle between SSEs and loop length, were used to successfully guide the design of ideal proteins [22].

We denote the set of all SSPs composed of *n* SSEs as S*n*; for instance, S3 consists of all SSPs with three SSEs. Numerous small domains or protein fragments, corresponding to SSPs consisting of two (S2) or three (S3) SSEs, are present in the PDB, and the knowledge of their local interactions guides structure prediction. One popular hypothesis suggests that stable SSPs serve as folding nuclei [23–26]. Accordingly, correctly recognizing such core SSPs may help ab initio structure prediction. Hidden Markov models for predicting SSP 3D context have been used to aid local structure prediction when a template is not available [27]. In CASP5, the FRAGFOLD server obtained the most accurate models for two new fold targets by assembling SSPs using a simulated annealing algorithm [9]. SSP classification has also helped loop modeling. A library of small SSPs consisting of two SSEs linked by a loop, called SMotif [28], has been used to reduce the loop search space by selecting both the candidate loop fragments that match loop length and also the “bracing SSEs” that bound the loop and meet geometrical requirements [29]. Recently, SMotif [28] and chemical shift information were combined to model larger structures [8].

Several theoretical models have been proposed to enumerate SSPs in protein folds. Owing to the difficulty of complete enumeration due to the very large number of possible SSPs, many of these models restrict their scope to various protein fold subsets. Early models describe α-helices packing onto β-sheets in a small subset of α/β folds [30], β-strand orientations in packed β-sheets [31,32], and α-helical arrangements in globular proteins [33]. Enumeration of β-strand arrangements in open β-sheets is widely studied [3,6,20,34]. For instance, a systematic analysis of topology preference for four-stranded β-sheet patterns found that 42 out of 96 possible topologies were identified in protein structures, and 50% of these structures were covered by only four topologies [34]. SSPs in β-sandwich structures have also been thoroughly investigated [35–40]. A comprehensive survey of Greek key motifs among β-barrels and β-sandwiches suggested basic rules that reflect their topological constraints and preferences [35]. Recently, models describing β-strand arrangements in β-sandwich structures have identified a characteristic feature among existing structures, termed “interlock”, and used it as a rule to distinguish and predict β-sandwiches [38–40]. Using such rules drawn from the analysis of recurring SSPs in proteins, Efimov proposed a method that models fold growth through stepwise addition of one SSE to a root structure pattern [41]. With this method, Efimov outlined possible folding pathways for five protein superfamilies of diverse folds.

We propose a more general theoretical model of fold growth by generating all possible up-and-down, compact SSPs built on a hexagonal lattice. Here, up-and-down refers to the antiparallel orientation of the successive (as dictated by the sequence) SSEs in an SSP. Compactness broadly requires that we form tight clusters without holes in the middle of the SSP or concavities along the contour (periphery); particularly, when an SSE is added to an SSP (with at least two SSEs) to obtain a larger SSP, the added SSE must be adjacent to at least two of the SSEs of the SSP in the lattice. Within the definition of compactness, additional rules for combining two SSPs that both consist of at least two SSEs are detailed in the Appendix, along with some additional exceptions. Instead of growing structures by the addition of a single SSE, we extended Efimov's idea by treating larger SSPs as the combination of two smaller ones, that is, structural tree construction [41], where a new SSP was built by adding an additional SSE to the root SSP. However, Efimov's root SSP was predefined with certain common patterns. For example, all-β structure enumeration was limited to SSPs containing a specific root composed of only β-strands, and the α/β structure enumeration was confined to SSPs containing a βαβ unit. Moreover, Efimov used certain strict rules to guide the SSE addition so that the resulting SSP was much more likely to occur in the protein database. Compared with Efimov's work, our enumeration initiates from elementary SSEs (i.e., β-strand and α-helix) and grows without preference for handedness or connection type. Thus, our SSPs are more comprehensive and enable the identification of rare and unobserved SSPs in proteins. This idea of growing a larger SSP by combining smaller ones is a likely path of protein origin in nature [12,50,51].

Our model builds upon the basic root SSPs (helix–helix, strand–strand, helix-strand, and strand-helix),