Supersecondary Structure Motifs and De Novo Protein Structure Prediction

Bryn Reinstadler
Williams College

Jennifer Van
George Mason University

Amarda Shehu
George Mason University

31 July 2012

Abstract

Defining physically-realistic structural fragments with which to assemble novel protein conformations is a central component of the ab initio protein structure prediction problem. In this work we pursue supersecondary structure motifs. The motifs are extracted from analysis of non-redundant experimental protein structures from the Protein Data Bank. Analysis of the motifs shows that they are present in all known protein folds and may therefore allow focusing de novo structure prediction to native folds. Our results show that these motifs can be promising for ab initio structure prediction.

1 Introduction

Because of the difficulty of empirically deriving protein structure through means such as x-ray crystallography or NMR, the methods of ab initio structure prediction and de novo protein design have come into popularity in recent years. Ab initio structure prediction is based on Anfinsen's premise that a protein's structure is based on the primary amino acid structure of the protein and the environment of the protein. [1] A protein's conformation, then, is the lowest free energy conformation given a set of global constraints that are provided by the environment. Thus, using the primary amino acid structure of a protein and a rough simulation of its environment, it is computationally possible to derive possible native conformations of a protein. [2] [3]

Unfortunately, though computationally possible it is also quite computationally difficult. Structure prediction engines such as those found in prediction competitions like CASP or CAPRI often produce tens of thousands of candidate structures before refining their candidates through the use of energy functions and geometric hashing.

However, the work of Fernandez-Fuendes et al provides interesting possibilities, not only in filtering candidate structures in ab initio protein structure prediction, but also in de novo protein design. [8] [12] Fernandez-Fuentes and his lab extract certain types of secondary structure motif elements from proteins and shows that secondary structure motifs made of two secondary structures connected by a loop, which were termed 'Smotifs' in the paper, can be found in all proteins, and therefore can be used for a variety prediction purposes. [8] [9]

Though they explored the use of Smotifs in the work of producing novel folds, in this paper we replicate Fernandez-Fuendes' work using a larger pool of proteins from which to draw data. We use the resulting library of Smotifs to explore the possibility of using Smotifs along with knowledge of their frequency in the general population of known protein structures (derived from the Protein Database, or PDB, as well as a server that predicts secondary structure known

as DSSP) to provide another way of filtering candidate structures in ab initio protein structure prediction. [4] [5] [6] [11]

2 Acknowledgements

The author would like to thank her mentor, Professor Amarda Shehu of George Mason University for her research support. She would also like to thank her undergraduate teacher Professor Morgan McGuire for his support and encouragement in seeking out this research opportunity.

The work of Bryn Reinstadler et al is supported in part by the Distributed Research Experiences for Undergraduates (DREU) program, a joint project of the CRA Committee on the Status of Women in Computing Research (CRA-W) and the Coalition to Diversify Computing (CDC), which is funded in part by the NSF Broadening Participation in Computing program (NSF CNS-0540631).

3 Results and Discussion

3.1 Smotif Frequencies

As discussed above, Smotifs are defined by two secondary structures connected by a loop. Each Smotif is also defined by some geometries: the length, the distance between the two secondary structure elements, the hoist, the packing, and the meridian. Using these geometries, which are further explained in the Methods section, we developed a histogram of the frequencies of the four major types of Smotifs. The four major types of Smotifs are alpha-alpha, alpha-beta, beta-alpha, and beta-beta, as named by the make-up of the consecutive secondary structures (See Figure 1). [8]

It is clear that the frequency of Smotifs across types is surprisingly consistent, with a high number of Smotifs in relatively the same place across all four types. Each local maxima, as well, is located in roughly the same place for each of the four bins. This is a novel finding, because originally it was assumed by our team that each of the four types, because they are so distinct in how they chemically bond to one another (ie, beta sheets use hydrogen bonds to link together in a certain way that alpha helices do not).

3.2 A Scoring Function

We also used our library of Smotifs to help us characterize a preliminary scoring function based on the frequency of Smotifs in a protein. The scoring function finds the frequency of each Smotif based on the number of Smotifs in its sub-bin divided by the number of Smotifs in its primary bin, then multiplies the various frequencies of all of the Smotifs in a given protein. We used three CASP models, ranked easy, medium, and difficult. The models were ranked by difficulty of accurate prediction as defined by the number of correct sequence identifications versus the number of incorrect. [10] If the number of correct structure predictions was more than 3/4ths of the total predictions, then that protein was counted as an 'easy' protein to guess correctly. If the number of correct structure predictions was less than 1/4th of the total predictions, then that protein was counted as a 'difficult' protein to guess correctly. Proteins with numbers of correct structure predictions in between those two numbers were counted as 'medium' proteins to guess correctly.

After statistically analyzing the Smotif frequencies and probability score (as discussed above) of one of each type of protein, we came up with the following data.

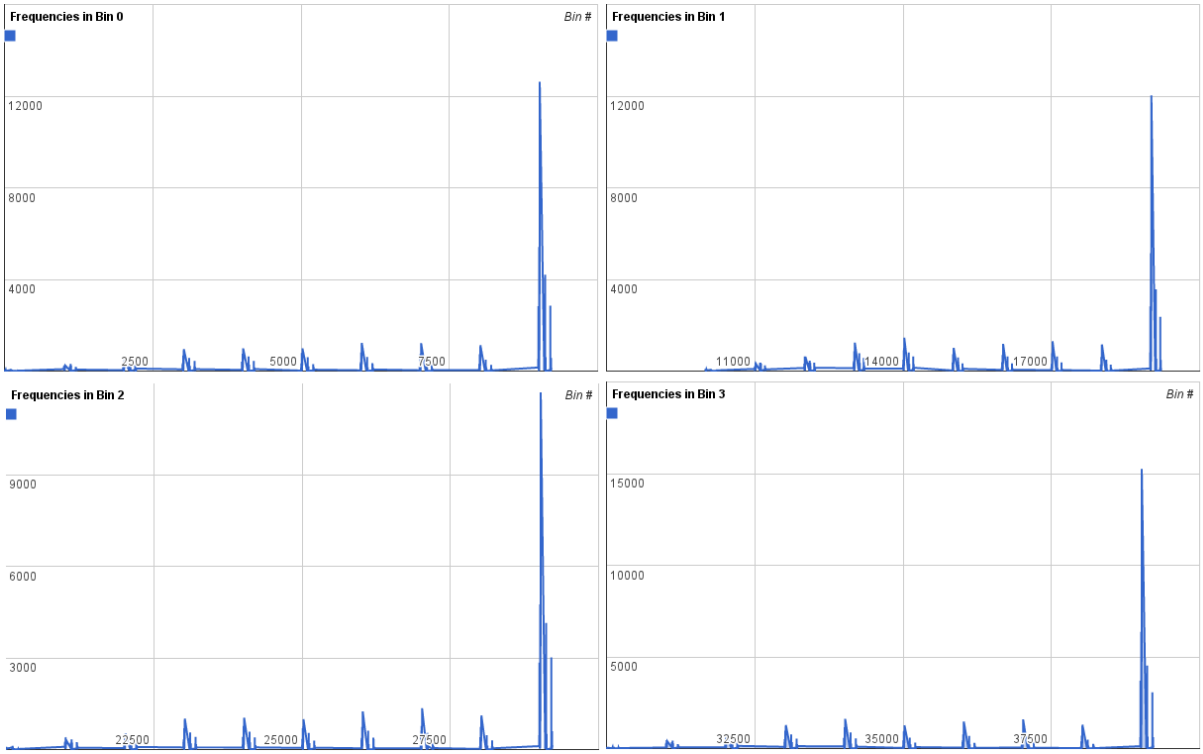

Figure 1:
From left to right, top to bottom:

(i) A graph showing the frequencies of Smotifs from Bin 0 - Alpha-Alpha.
(ii) A graph showing the frequencies of Smotifs from Bin 1 - Alpha-Beta.
(iii) A graph showing the frequencies of Smotifs from Bin 2 - Beta-Alpha.
(iv) A graph showing the frequencies of Smotifs from Bin 3 - Beta-Beta.

Protein Probabilities based on Component Smotif Frequencies

Easy	$5.96 * 10^{-4}$
Medium	3.0826
Difficult	$1.043 * 10^{-8}$

Because we only used one protein from each category, the data may be skewed. However, our goal was only to use these figures as a proof of concept, that perhaps something could be done with this scoring method in the future. For example, both the easy and medium proteins showed higher probabilities of being found in nature than the difficult protein, which was to be expected.

However, our results were only partially consistent with the expected results. It was expected that 'easier' proteins would be those with a high percentage of high-frequency Smotifs, and therefore a high scoring using our simple scoring function. This was the expectation because current ab initio protein building methods are partially built on the assumption that the secondary structure motifs of one protein can be used to help inform the building of another; therefore if a particular Smotif has a high frequency in the total population of discovered proteins, then it would make a sensical template for ab initio protein building and protein structure

guessing. However, the fact that the easy protein showed a lower level of predictability than the medium protein does not fit with the previous hypothesis. There could be many reasons for this; chief among them is the fact that there were significantly fewer Smotifs in the easy protein as compared to the medium and difficult proteins, and so any abnormality in the number of low-frequency proteins could skew the data.

4 Methods

During this study, we utilized a self-made protein parsing system that took PDB and DSSP files of select proteins and gave as output the Smotifs found in each protein. [4] [6] [10] [11]

We selected the proteins that we would parse using the PISCES server's culled protein lists [7], selecting for a list of proteins that had a less than 80

Our parser took the aforementioned PDB and DSSP files from the culled list of proteins from the Dunbrack Lab's PISCES server and extracted the following information: crystallization factor, number of atoms in each residue, the xyz coordinates of each atom in the protein, the number of residues in each secondary structure (as predicted by DSSP, the secondary structure prediction algorithm), and the number of secondary structures in the protein itself.

Several factors could render a Smotif incomplete and therefore inadequate for future studies. A Smotif is defined by two consecutive secondary structures connected by a loop. If any residues in a Smotif had a crystallization factor greater than 1.0, or if it were more than 0.25 greater than the mean of the protein's total average crystallization factor, that residue (and the secondary structure it was contained in) were both considered 'missing' and as such those 'missing' secondary structures were not used in the extracting of a protein's Smotifs. If a loop had more than 12 residues, then the secondary structures it was adjacent to were counted as not being attached to one another.

After the Smotifs were correctly identified from the structure and arrangement of the protein, the c-termini and n-termini of each secondary structure in each Smotif were used to determine the geography of the Smotif. Consider the following:

The N and C-terminal secondary structures of a Smotif are henceforth referred to as SS1 and SS2. They are each represented by their principal moments of inertia. P0 is the startpoint, in cartesian coordinates, of SS1; P1 is the ending point of SS1, P2 is the starting point of SS2, and P3 is the ending point of SS2. Plane is defined by M1 (the moment of inertia of SS1) and L, which is defined by the space between P1 and P2. Plane is defined by M1 and the normal to plane . [13] The geometry of a Smotif, then, is defined by five measures. [8]

1. The measure of the length of the Smotif as determined by how many residues are in the Smotif from P0 to P3. This helps distinguish proteins that are otherwise similar but have a difference of more than 3 residues in length.

2. The distance between the C-terminal of SS1 and the N-terminal of SS2 (which are, respectively, P1 and P2). This is henceforth referred to as (D).

3. The 'hoist' of the Smotif, defined by the angle between L and M1, known as (δ).

4. The 'packing' of the Smotif, defined by the nagle between M1 and M2 (which is the moment of inertia of SS2), known as (θ).

5. The 'meridian' of the Smotif, defined by the angle between M2 and , known as (ρ).

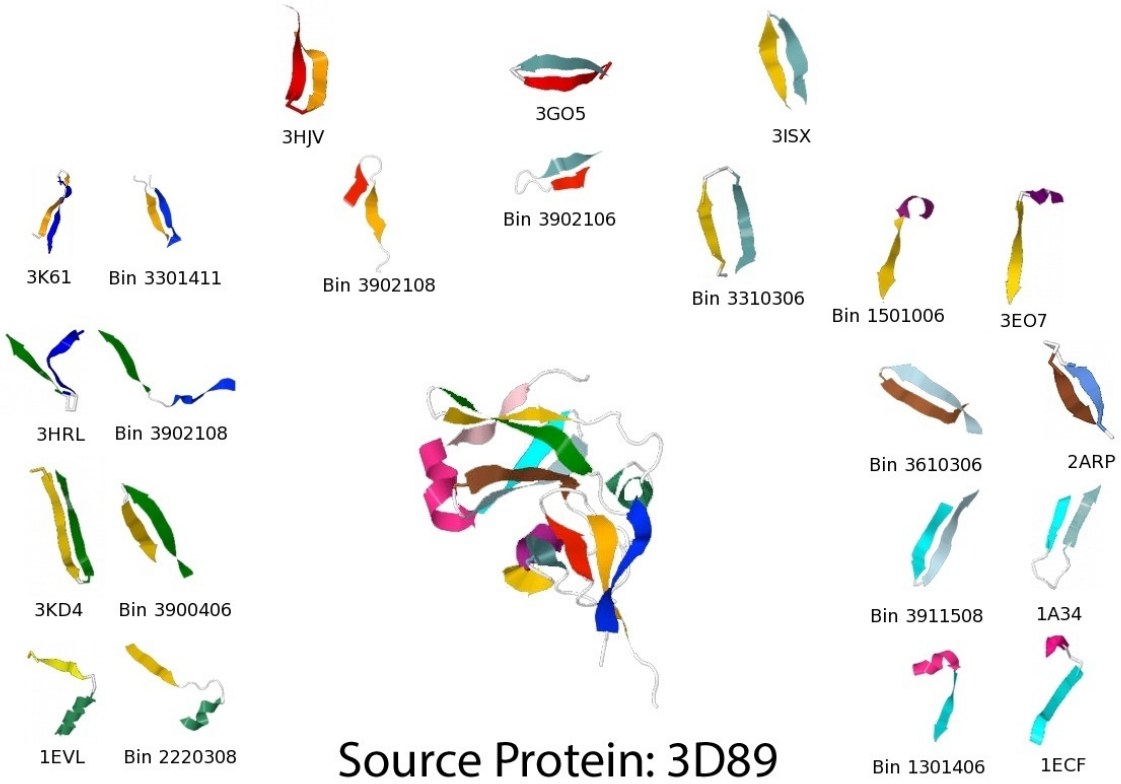

Figure 2: This figure shows Smotifs mined from a single protein and subsequently binned. The bin numbers can be found under the medial Smotifs, which are taken from the protein 3D89. The outer, or distal, Smotifs are mined from other proteins (as named beneath them) but have the same bin as the Smotif they are next to.

After parsing the Smotifs, we identified them by the geometries by 'binning' them. Each bin had a 7-digit characterization. All sub-bins, which are represented by one digit of the total bin, are zero-indexed. The first digit, with a range of 0 to 3, represented the types of secondary structure that the Smotif contained. The four types categorized were those with two alpha-helix secondary structures, an alpha and a beta, a beta and an alpha, or two beta-sheet structures, respectively. The second digit represented the distance D, which ranges from 0 to 40 Angstroms. Values larger than 40 were assigned to 40. There would 10 4-Angstrom bins for distance. Delta and theta angles are represented by the next two digits. Both types of angles span from 0 to 180 degrees. There are 60 degree sub-bins for delta and theta. The sub-bins for rho, which spans between 0 and 360 degrees, are also 60 degree bins, but they start from 30 degrees. Again, all of the bins are zero-indexed, so if there are 3 bins the digits used to represent them are 0, 1, and 2. [8] We categorized length as between 0 and 200, with a Smotif of any other length between set to 200 residues in length. There were 100 bins for length, from 00 to 99.

Using the binning information, we used simple statistical analysis to determine the relative frequency of Smotifs within the population of Smotifs we had discovered.

We also utilized a basic scoring mechanism to score the probability of a structure being found in nature based on the frequency of its component Smotifs. We did this by multiplying

the frequencies of each Smotif in the protein.

5 Future Work

Though our work helped the field of bioinformatics in some ways, there are opportunities for future work. For example, the scoring mechanism can and should be expanded to larger protein samples that have been similarly rated by CASP as to the difficulty of prediction by human or server teams. Using a larger spread of information, it would be possible to ascertain whether the findings as to the disparity between the medium and hard proteins were statistically significant and whether or not the findings with the easy protein were accurate.

In addition to the previous, it was found in the paper by Fernandez-Fuentes that proteins are made up of a smattering of low-frequency proteins and a majority of high-frequency proteins. Future work could consist of trying to find an algorithm for ab initio protein formulation based on the frequency of the component Smotifs.

References

[1] Anfinsen CB. Principles that govern the folding of protein chains. Science 1973;181:223230.

[2] Aung, Zeyar, and Jinyan Li. "Mining Super-secondary Structure Motifs from 3D Protein Structures: A Sequence Order Independent Approach." Genome Informatics 19 (2007): 16-26. Japanese Society for Bioinformatics. Secretariat of Japanese Society for Bioinformatics, 2007. Web. 31 July 2012.

[3] Ausiello, Gabriele, Pier F. Gherardini, Elena Gatti, Ottaviano Incani, and Manuela Helmer-Citterich. "Structural Motifs Recurring in Different Folds Recognize the Same Ligand Fragments." BMC Bioinformatics 10.182 (2009): n. pag. BioMed Central. Springer Science+Business Media, 15 June 2009. Web. 31 July 2012.

[4] A series of PDB related databases for everyday needs. Joosten RP, Te Beek TAH, Krieger E, Hekkelman ML, Hooft RWW, Schneider R, Sander C, Vriend G, NAR 2010; doi: 10.1093/nar/gkq1105

[5] Colloc'h, N., C. Etchebest, E. Thoreau, B. Henrissat, and JP Mornon. "Comparison of Three Algorithms for the Assignment of Secondary Structure in Proteins: The Advantages of a Consensus Assignment." Protein Engineering 6.4 (1993): 377-82. National Center for Biotechnology Information. U.S. National Library of Medicine. Web. 31 July 2012.

[6] Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Kabsch W, Sander C, Biopolymers. 1983 22 2577-2637. PMID: 6667333; UI: 84128824.

[7] G. Wang and R. L. Dunbrack, Jr. PISCES: a protein sequence culling server. Bioinformatics, 19:1589-1591, 2003.

[8] Fernandez-Fuentes, Narcis, Joseph M. Dybas, and Andras Fiser. "Structural Characteristics of Novel Protein Folds." Ed. Ruth Nussinov. PLoS Computational Biology 6.4 (2010): 1-11. Print.

[9] Fernandez-Fuentes, Narcis, Baldomero Oliva, and Andrs Fiser. "A Supersecondary Structure Library and Search Algorithm for Modeling Loops in Protein Structures."Nucleic Acids Research 34.7 (2006): 2085-097. National Center for Biotechnology Information. U.S. National Library of Medicine, 14 Apr. 2006. Web. 31 July 2012.

[10] Fidelis, Krzysztof, Andriy Kryshtafovych, and Bohdan Monastyrskyy. Critical Assessment of Protein Structure Prediction (CASP). Protein Structure Prediction Center. US National Institute of General Medical Sciences (NIH/NIGMS), n.d. Web. 31 July 2012. ¡http://predictioncenter.org/index.cgi¿.

[11] H.M. Berman, K. Henrick, H. Nakamura (2003): Announcing the worldwide Protein Data Bank. Nature Structural Biology 10 (12), p. 980.

[12] Verschueren, Erik, Peter Vanhee, Almer M. Van Der Sloot, Luis Serrano, Frederic Rousseau, and Joost Schymkowitz. "Protein Design with Fragment Databases." Current Opinion in Structural Biology 21.4 (2011): 452-59. Print.

[13] Zemla, Adam. "LGA: A Method for Finding 3D Similarities in Protein Structures." Nucleic Acids Research 31.13 (2003): 3370-374. Oxford Journals. Oxford University Press, June 2003. Web. 31 July 2012.