5-28-2015

Why Do Protein Structures Recur?

Rebecca G. Leong
Dartmouth College

Why Do Protein Structures Recur?

Dartmouth Computer Science Technical Report TR2015-775

Rebecca Leong, Gevorg Grigoryan, PhD

May 28, 2015

Abstract

Protein tertiary structures exhibit an observable degeneracy in nature. This paper examines the connection between a protein motif's abundance in nature and its designability as measured by *in silico* methods. After generating a set of protein structures, we evaluated each structure's abundance in nature, ratio of possible contacts (contact degree) and *in silico* designability. Our results showed that any two or these metrics are moderately correlated. Together abundance and contact degree produced the strongest correlation with *in silico* designability. Our results suggest that abundance is indeed an indicator of designability. Furthermore, abundance and contact degree appear to correlate with some distinct components of *in silico* designability.

1 Introduction

Proteins are the focus of a host of cellular pathways and potential novel design applications. The function of any protein is dependent on its 3D structure. The ability to anticipate whether a particular structure can be realized with natural amino acids is of significant utility for novel protein design applications [11]. Sequence optimization and structure optimization are asymmetrical processes. A sequence that is optimized for a particular structure will often not fold to that same structure as its optimal conformation [11]. This is in part due to an infinite number of possible structural conformations. A structure is considered designable if it can be realized. An intuitive indicator of whether a structure can be designed is if that structure or a very similar structure has occurred in nature. This paper examines how a structure's abundance in nature can contribute additional insight into our understanding of what makes a structure designable.

On a high level, designability is a complex property that describes how easily a particular structure can be realized using the 20 naturally occurring amino acids. The designability of a structure is loosely defined as the number of sequences for which that structure is the lowest energy configuration [7], [11]. In order for a structure to be designable, it must be the optimal ground-state for some sequence [7]. The designability of a structure cannot be easily measured directly, however, there are other measurable characteristics that have been shown to be indicators of it. Designable structures are the ground state for many sequences and tend to be more thermodynamically stable and resistant to mutation than less designable ones [7]. Anfinsen *et al.*'s thermodynamic hypothesis states that the native state of a protein is the structure with the global minimum of free energy [1]. The free energy landscape of a structure will often have many metastable states that are less stable than the global minimum. This paper will measure a structure's *in silico* ability to be an energetic global minimum for some sequence as a proxy measurement for its designability.

Natural proteins adopt only a limited number of folds [7] and as a result, the protein universe exhibits degeneracy [11]. Previous studies have estimated that there are only about 1000 distinct natural protein folds (defined as the same major secondary structure elements, arrangement and

topological connections) [7]. Systematic categorization of loop connected secondary structure elements found that various classes occurred at highly variable frequencies in the Protein Data Bank (PDB) [5]. One explanation for this bias is that structures that occur frequently in nature are easier to realize than other less frequently occurring ones. This leads to the common hypothesis that abundant structures are more designable than less abundant structures [7], [6], [11] . A structure that has existed in nature is presumably designable to some extent since it has been realized at some point. However, low abundance structures are not necessarily non-designable. Structures may not exist in the PDB for a number of reasons. This study will examine the correlation between a structure's frequency in nature and Rosetta designability for a range of abundance values. In *de novo* protein design, the ability to quantify a structure's designability would allow researchers to filter unfoldable structures. It is often the case that a specific protein structure is desired, but there is no guarantee that a designed sequence would indeed fold to this conformation [3]. Efficient means of computing designability would improve novel protein design methods by allowing us to focus on achievable structures.

This study will examine to what extent various indicators of designability correlate with each other. Specifically, it will explore the hypothesis of whether a structure's abundance in nature is correlated with the ratio of potential contacts and Rosetta designability. To do this, we will examine a collection of randomly generated small structural motifs. Prior work from the Grigoryan lab has led to the search algorithm MASTER, which efficiently identifies matches to arbitrary disjoint backbone fragments [13]. This allows us to compute how frequently a given structural motif exists in the Protein Data Bank (PDB). Potential contacts describes the contacting amino acids that can exist in each position pair based on backbone orientation and environment of those positions [12]. A structure's designability will be estimated through a proxy of *in silico* methods to model folding and structural energy. *In silico* methods aren't ground truth, but they offer realistic representation of protein interactions through experimentally driven models of molecular physics. Furthermore, these modeling methods are widely used in other *in silico* predictions and applications.

2 Methods

In order to test our hypothesis, the high-level workflow described in figure 1 was used to generate a set of designed and corresponding folded structures.

2.1 Initial Structure Set

To generate the initial structures, a parallel beta strand and alpha helix were randomly combined. Both chains were initially aligned to the origin. The alpha helix was translated and rotated by random values within a set range (8Åfor translation and 360 degrees for rotation). The resulting conformation was then checked for backbone clashes (defined as any two alpha carbons within 4.0Åof each other). The clash-free structure was then run through MASTER to ensure that the motif had a specified minimum number of matches within 1.5Å. A range of minimum match values from 0 to 100 were used to generate the initial protein set. If the structure had insufficient number of hits, the chains were aligned to the fifth top match (ordered by RMSD distance) and then run through MASTER again. If no structure with sufficient matches was generated after three iterations of this method, the run was terminated. Certain numbers of matches were easier to satisfy than others, but we aimed to create a sufficient diversity in number of matches. This method resulted in 129 structures. Sequences were then designed for each of these structures.

2.2 Structure Optimization Algorithm

All structure optimizations were performed using a variable-temperature MonteCarlo algorithm implemented in PyRosetta [4]. Each iteration involved a small random change in structural con-

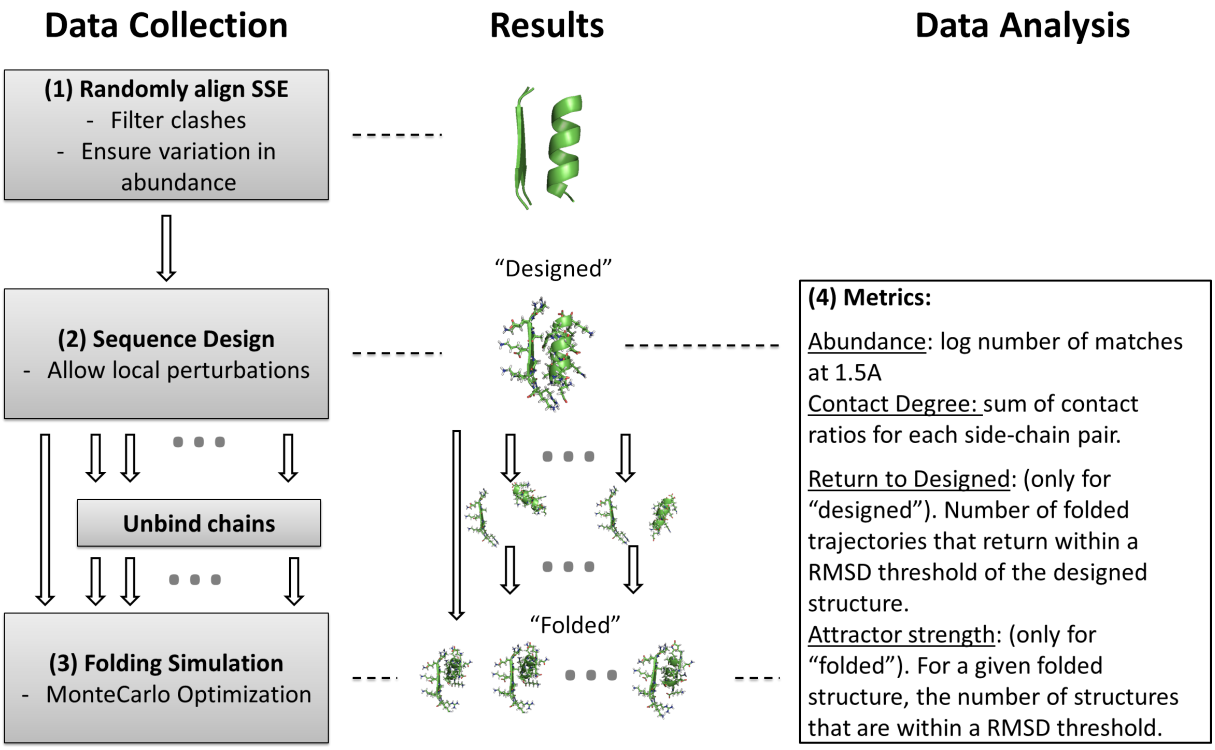

Figure 1: The above diagram outlines the high-level work flow of both the data collection and analysis as well as introduces key terms to be used throughout this paper. (1) Structures are generated by randomly aligning a parallel beta sheet and alpha helix. Structures are filtered for clashes and evaluated for abundance to ensure that sufficient diversity exists in the initial structure set. (2) These initial structures each undergo an iterative sequence design phase in PyRosetta. The output structure from this step is referred to as the "designed" structure. (3) Using Rosetta, ten simulated folding trajectories are performed on each of the designed structures. With the exception of one trajectory, the chains of all the other trajectories are first unbound. The lowest energy structure is saved from each run to form the set referred to as the "folded" structures. (4) To analyze the designed and folded structures, a set of metrics were measured for each of them. The abundance in nature (abundance) and ratio of potential contacts (contact degree) were calculated for all structures. Return to designed (RTD) is calculated for the designed structures and attractor strength is calculated for the folded structures. Both the RTD and attractor strength measures refer to the number of folded trajectories whose optimal structures are within a specified RMSD cutoff of the structure. See the methodology section for additional details about each stage of this process.

formation (translational and rotational) followed by side chain repacking. Structures were then scored using PyRosetta's scoring function, which uses an empirically based energy function that accounts for features such as molecular bonds, angles and environment to estimate its free energy. Scores are accepted or rejected using the Metropolis criterion [2]. The Metropolis criterion determines the probability P that a structure is accepted:

$$P = \begin{cases} e^{-\Delta E/kT} : \Delta E \geq 0 \\ 1 : \Delta E < 0 \end{cases}$$

$$\Delta E = energy_i - energy_{i-1}$$

for the structure of the current iteration i [2].

The Metropolis criterion varied in strictness throughout the simulation to balance the trade off between depth and breadth while sampling the structure's free energy landscape. The kT value was initialized to one Rosetta Energy unit. If a structure is accepted, the kT value is reduced by 1%. If a structure is rejected, the kT value is increased by 1%. This allows the algorithm to both search local minima and escape such minima without relaying on full run resets. After a preset number of iterations, the lowest energy structure was retrieved and saved. Variations of this method were used to create both the designed and folded structures.

2.3 Designed Structure Set

The designed structures were created by taking the initial structures from the random generation method and performing sequence design on them. Using PyRosetta and the algorithm described above, each structure underwent 10^3 iterations of local perturbations and side chain repacking with sequence design. This allowed the designed structure to find an optimized model without drifting too far from the initial structure. Local perturbations restricted the designed structure to remain within 0.5Åof the initial structure. Structures retrieved from this method were labeled as the designed structures. Their abundance and ratio of potential contacts were then recalculated.

2.4 Folded Structure Set

The folded structures used the same MonteCarlo optimization algorithm on the designed structures but only allowing for side chain repacking (no sequence design). To understand the energy landscape of the designed structures, 10 independent trajetories of folded structures were collected. Prior to running the optimization, the chains were first randomly dissociated (chains were separated by at least 5Å) for 9 of the runs. The final run started at the designed conformation. Each run ran for 10^6 iterations with only side chain repacking and required trajectories to remain within 10Åof the designed structure. These 10 trajectories created the corresponding folded structures for each of the designed structures. Again, abundance and ratio of potential contacts were calculated for each of these structures.

2.5 Measurements for analysis

As described briefly in figure 1, four key measurements were collected and analyzed in this study: contact degree, abundance, return to designed rates and attractor strength. Both return to designed (RTD) rates and attractor strength are proxy measurements of relative folding ease. For the designed structure, its RTD score is calculated based on the number of folded trajetories whose lowest energy structure is within a specified RMSD cutoff of the designed structure. The attractor strength score captures a similar property: for a particular folded structure, it is the number of folded runs (from the same designed structure) whose lowest energy structure are within a specified RMSD threshold. Both measurements should be examined at a variety of RMSD thresholds. The attractor strength can be calculated for any of the folded structures while RTD can only be calculated for a designed structure. Both the RTD and attractor strength values measure the strength of the folding funnel for the particular structure, but RTD value focuses on the relative folding ease of the structure for which the sequence is optimized.

The abundance of all structures was calcu-

lated using the MASTER software previously developed by the Grigoryan Lab [13]. MASTER takes a given motif and queries a given database (the PDB) for structurally similar (ordered by RMSD) proteins. MASTER returns the number of matches in the PDB within a specified RMSD cutoff of the given structure. The term abundance refers to the *log* of the number of unique protein matches (plus a small value to avoid taking the *log* of 0) found by MASTER within 1.5Å. The contact degree of a structure captures the fraction of potential contacting amino acid pairs that could exist in each pair of positions. Two positions were considered interacting if they could structurally influence the amino acid identity and conformation of each other. This accounts for backbone orientation and structural environment of the two positions being considered. The equation in figure 2 was used to determine the fraction of rotamer pairs $f(i,j)$ forming close contacts for positions i and j. Finally the contact degree is the sum of $f(i,j)$ for all pairs of positions.

3 Results

Structures are binned by either abundance and/or contact degree to enable analysis.

3.1 Abundance vs. Rosetta Metrics

The first question examined is whether abundance is correlated with our *in silico* energetic proxy for designability. The data suggests that abundance and relative folding ease, as measured by Rosetta, were moderately correlated. Figure 3 shows that highly abundant structures are generally more likely to return to designed. The RTD rates are shown for 4 different RMSD thresholds to determine whether a folded trajectory has returned to the designed structure. At a threshold of 1.5Å, an average of 0.37 trajectories return to designed for the highest abundance bin, while only 0.27 trajectories return in the lowest abundance bin. The correlation is most apparent for a threshold of 2.0Å. The increase in correlation strength as the RMSD threshold increases suggests that more abundant structures perhaps have a wide-folding funnel with additional local minima features within it. For all definitions of native, the second to lowest abundance bin has the lowest RTD fraction indicating that structures that do not occur at all in nature are not necessarily energetically unfavorable, but may represent a special case for further examination.

Attractor strength measures the relative folding ease for any structure for a particular sequence. Figure 4 shows that there's a stronger correlation between attractor strength and abundance. Looking at the strictest threshold for attractor strength (>8 neighbors), 0.51 of the highest abundance structures meet this criteria compared to only 0.10 of the lowest abundance structures. The trend appears exponential suggesting that the correlation between abundance and designability is even more pronounced at abundance levels higher than those examined. The RTD metric measures the relative folding ease specific to the structure for which the sequence was designed. The stronger correlation in attractor strength than RTD suggests that there is a correlation between abundance and specific encodability, but that a designable starting point does not ensure a global minimum. This shows an unbiased tendency for highly abundant structures to be an optimal conformation.

The data supports that sequence optimization for a particular motif, even if such structure is fairly abundant, does not ensure that the sequence will fold to that motif in isolation. One could test this hypothesis by examining if folded trajectories end at a more abundant conformation than the designed structure. Examining this metric revealed that of structures that both did not return to designed and were strong attractors, just over 1/3 ended at a more abundant structure. Increased abundance alone does not account for structures not returning to designed. This result suggests that abundance is a fairly complex property that is not completely captured by the folding simulations of that motif alone in Rosetta. In order to further understand the relationship between abundance and *in silico* designabiltiy, contact degree was also considered.

$$f(i,j) = \frac{\sum_{a=1}^{20}\sum_{b=1}^{20}\sum_{r_i \epsilon R_i(a)}\sum_{r_j \epsilon R_j(a)} C_{ij}(r_i,r_j)Pr(a)Pr(b)p(r_i)p(r_j)}{\sum_{a=1}^{20}\sum_{b=1}^{20}\sum_{r_i \epsilon R_i(a)}\sum_{r_j \epsilon R_j(a)} Pr(a)Pr(b)p(r_i)p(r_j)}$$

Figure 2: The equation to calculate the fraction of possible contacts between two positions taken from previous work by Zheng [12]. $R_i(a)$ is the set of non-clashing rotamers of amino acid a and position i, $C_{ij}(r_i, r_j)$ is a binary variable indicating whether rotamers r_i and r_j have heavy atom pairs within 3A of each other. $Pr(a)$ is the frequency of amino acid a in the structural database. $p(r_i)$ is the probability of rotatmer r_i from the rotamer library by Richardson *et al.* [8]

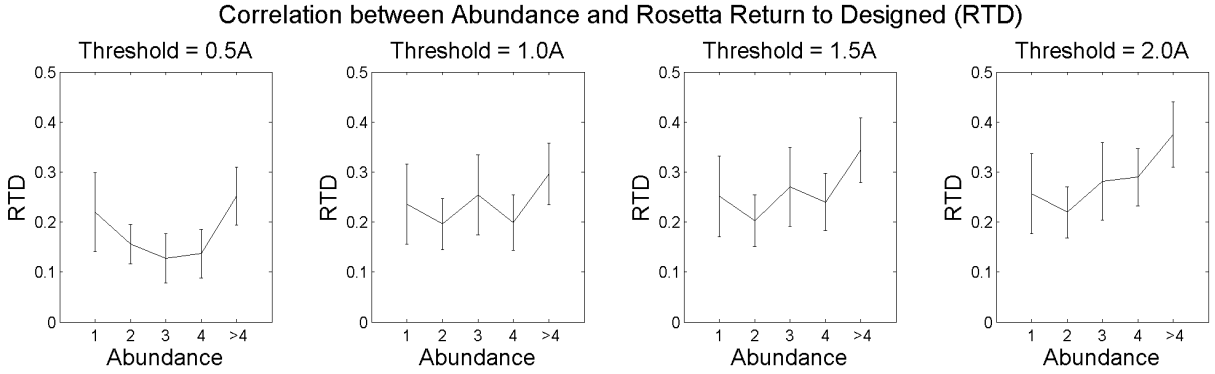

Figure 3: Each plot shows the fraction of structures that return to designed for each bin of structures. Returning to designed was defined as finding the folded structure within a certain RMSD threshold (0.5Å, 1.0Å, 1.5Åor 2.0Å) of the designed structure. Each structure belongs to exactly one bin in which the X value is representative of a range of abundance values. A structure in bin 2 has an abundance between 1 and 2. For each designed structure, the fraction of structures was determined by the number of search iterations that found their lowest energy structure within the specified RMSD cutoff.The error bars show the standard error for each bin.

3.2 Contact Degree vs. Rosetta Metrics

Contact degree measures the number of potential amino acid pair interactions between the two chains. Previous work with simple lattice models has found correlations between increased contact degree and designability [10]. The data suggests a strong correlation between contact degree and Rosetta designability. Similar to abundance, figure 5 shows that contact degree has a moderate correlation with RTD rates. At a threshold of 1.5Å, 0.36 of structures in the highest contact degree bin returned to designed, compared to the 0.19 of the lowest contact degree bin. The increase in correlation is unsurprising as increased contact degree would give Rosetta more potential amino acid pairs to design.

Contact degree and attractor strength show a prominent positive correlation. Figure 6 shows increases in contact degree are almost always accompanied by an increase in fraction of structures that meet the attractor strength criterion. Most notably, for the strictest cutoff (>8 neighbors) 0.40 of the highest contact degree structures were strong attractors compared to only 0.04 of the lowest contact degree structures. Unlike abundance, the correlation is fairly consistent across the four contact degree bins. This consistency even in the lowest bin suggests that contact degree is informative across all ranges examined. Furthermore, the data suggests that

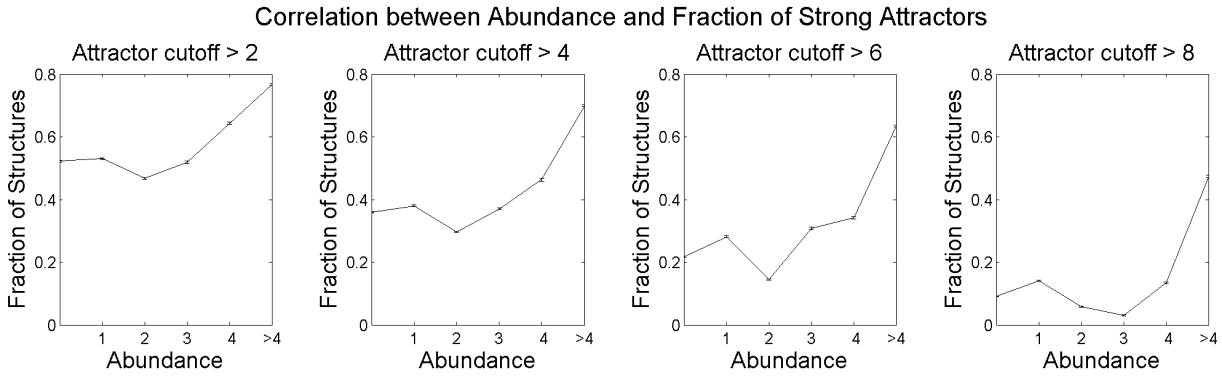

Figure 4: Each plot shows the fraction of structures that have at least N (20, 40, 60,80) neighbors within 1.5A for each bin of structures. A neighbor is defined as another search structure that is within 1.5Åof RMSD of the current structure. Each structure belongs to exactly one bin in which the X value is representative of a range of abundance values. A structure in bin 2 has an abundance between 1 and 2.The error bars show the standard error for each bin.

Rosetta's methods more accurately capture the properties of contact degree than abundance.

3.3 Abundance vs. Contact Degree

Both abundance and contact degree are indicators of designability to some extent. Naturally one might question whether each measure captures the same features of designability. The data suggests that contact degree and abundance are also positively correlated especially in the higher abundance structures. Figure 7 shows that low abundance structures are not very informative of the contact degree of a structure, but higher abundance structures are fairly informative. More abundant structures tend to have a higher contact degree. The reverse is not necessarily true as high contact degree structures appear even in the lowest abundance bin. This is not surprising given that structures may not be in the Protein Data Bank for a number of reasons. Contact degree and abundance to some degree appear to capture similar features, but do not completely determine one another. Abundance is a complex feature that is not explained by contact degree alone.

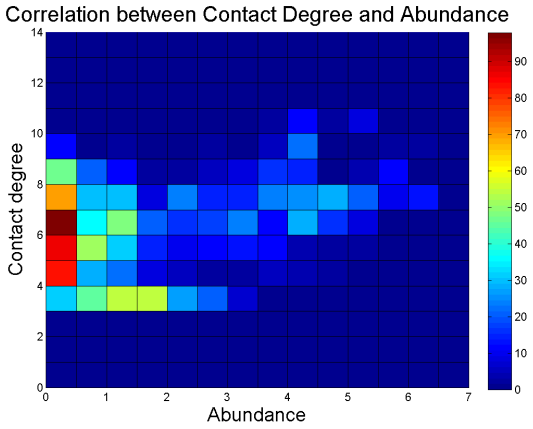

Figure 7: This plot visualizes the frequency at which each contact degree and abundance pairing occurs.

3.4 Combined Analysis

Each pair of metrics are positively correlated to some degree. A natural question to examine next is whether there are correlations in the combined analysis of the three metrics. Particularly, we were interested in whether abundance along with contact degree could produce a better indicator of the *in silico* designability of a structure. Figure 8 shows no obvious trend in RTD rates within the abundance-contact degree space. The small data set results in few to no structures in many of

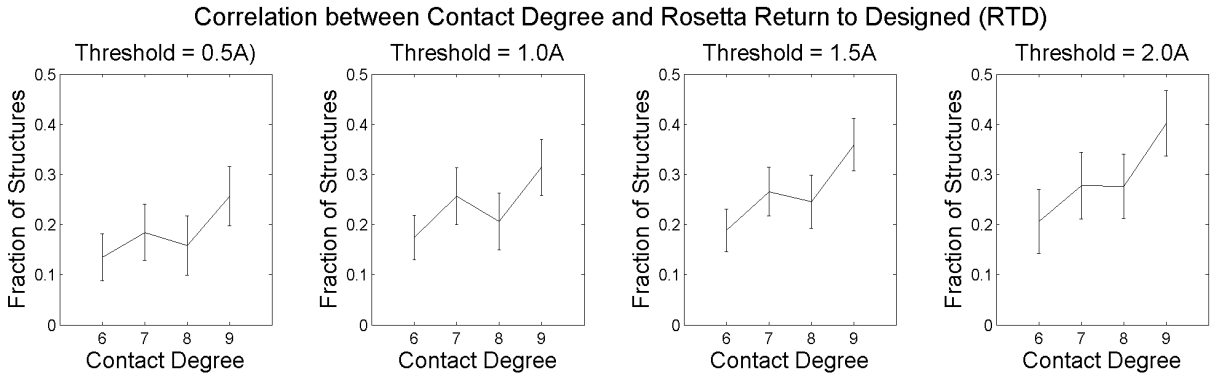

Figure 5: Each plot shows the fraction of structures that return to designed for each bin of structures. Returning to designed was defined as finding the optimal structure within a certain cutoff (0.5Å, 1.0Å, 1.5Åor 2.0Å) of the designed structure. For each designed structure, the fraction of structures was determined by the number of search iterations that found their lowest energy structure within the specified RMSD cutoff. Each structure belongs to exactly one bin in which the X value is representative of a range of contact degree values. A structure in bin 6 has a contact degree between 6 and 7.The error bars show the standard error for each bin.

the bins making any analysis difficult and prone to uncertainty.

On the other hand, Figure 9 shows a prominent distinction between high abundance-contact degree bins and low abundance-contact degree bins. Notably, the higher abundance-contact degree bins show an increased likelihood of being a strong attractor (more than 7 neighbors). Plots are normalized to gain a sense of significance even though some bins only contain a few structures. A t-test found the difference in rates of strong attractors between high CD (11-12) - high abundance (11-12) structures to be significantly different from low CD (5-6) - low abundance $(1-2)$ structures ($p < 0.05$).

Based on the pair-wise analysis we know that each pair of metrics is loosely correlated and measure different characteristics about a structure. It appears that together contact degree and abundance capture characteristics similar to those captured by Rosetta's *in silico* methods. The combined analysis suggests that a combination of abundance and contact degree correlates more strongly with Rosetta designability than either alone. Of structures with both high abundance (5-6) and contact degree (10-11), 0.78 were strong attractors. Only 0.56 of the structures with high contact degree (10 - 11), but low abundance (0 -2) were strong attractors. A t-test ($p = 0.38$) indicates that the difference was not significant, but additional samples would be useful in examining this hypothesis. Similarly, compared to abundance alone, only 0.27 of high abundance (4.5-5.5), but low contact degree (4-6) structures were strong attractors. Similar to contact degree, a t-test indicates that this trend is not significant (p=0.14). Thus we see that high abundance and contact degree has a stronger correlation than either metric alone. Although the differences are not significant, with additional data, this hypothesis deserves additional examination.

Figure 9 shows a bin of structures that contradict our hypothesis: Structures with very low abundance and high contact degree have an unexpectedly high fraction of strong attractors. Figure 10 shows one such structure for which the designed structure had a reasonable abundance and contact degree, yet all the folded trajectories ended up at an alternate conformation. By visual inspection, as a motif component, the final folded structure does not appear to be stable or more stable than the designed structure. The perpendicular orientation of the folded structure's chains does not appear to support enough contacts to maintain that conformation. There may be two factors contributing to

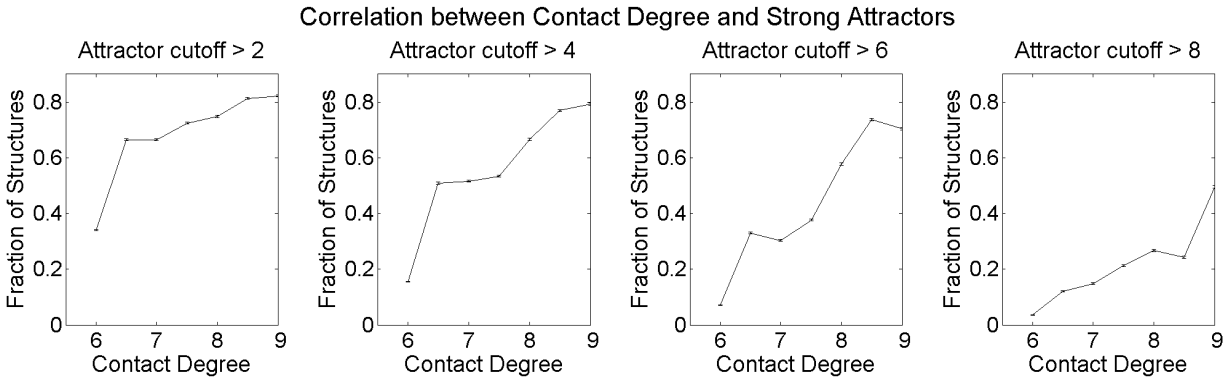

Figure 6: Each plot shows the fraction of strong attractors based on increasingly threshold definitions (2, 4, 6, or 8 neighbors) of what is a strong attractor. A neighbor is defined as another folded structure that is within 1.5Åof RMSD of the current structure. Each structure belongs to exactly one bin in which the X value is representative of a range of contact degree values. A structure in bin 6 has a contact degree between 6 and 7. The error bars show the standard error for each bin.

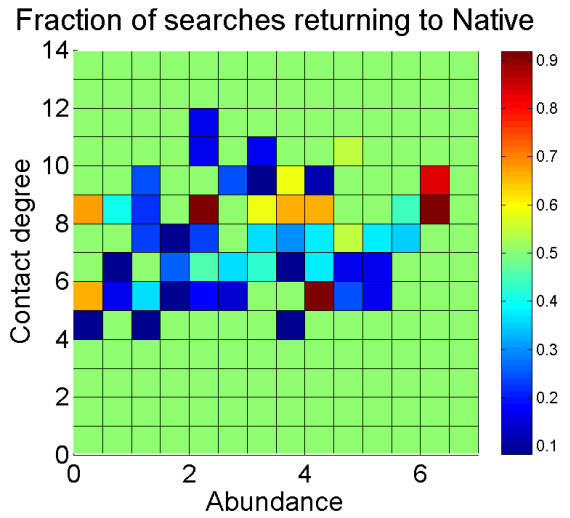

Figure 8: The graph shows the average fraction of structures that return to designed(0.5Å) correlated by contact degree and abundance. The fraction has been smoothed assuming 0.5 naive return rates.

the low correlations between the examined metrics. First, high abundance of a motif in nature in the context of an entire structure does not ensure that such motif will fold independently. Second, Rosetta scoring methods may not accurately find the optimal structure for small, two-chain motifs.

4 Discussion

Throughout this study, it has been observed that abundance and contact degree both show a positive correlation with designability as measured by Rosetta. It is often the case that high abundance structures are highly informative of their designability. Lower abundance structures are less informative about the designability of a structure. Although the low abundance structures tend to be less designable, there is still a large potential range of designability values. This aligns with prior hypotheses which state that abundance may be useful to confirm a structure's designability, but cannot necessarily confirm that it will be non-designable. Furthermore, these data suggests that some aspects of designability is not captured by abundance. Identifying specific cases where discrepancies occur may provide insights into the abundance-designability relationship and means in which *in silico* evaluation methods could be improved.

Contact degree appears to be slightly more informative along all possible values. As a single metric, current *in silico* methods better model the properties captured in contact degree than those involved with abundance. But results also show that, as expected, abundance is not simply a measure of contact degree and is indeed a more complex property. It further suggests that con-

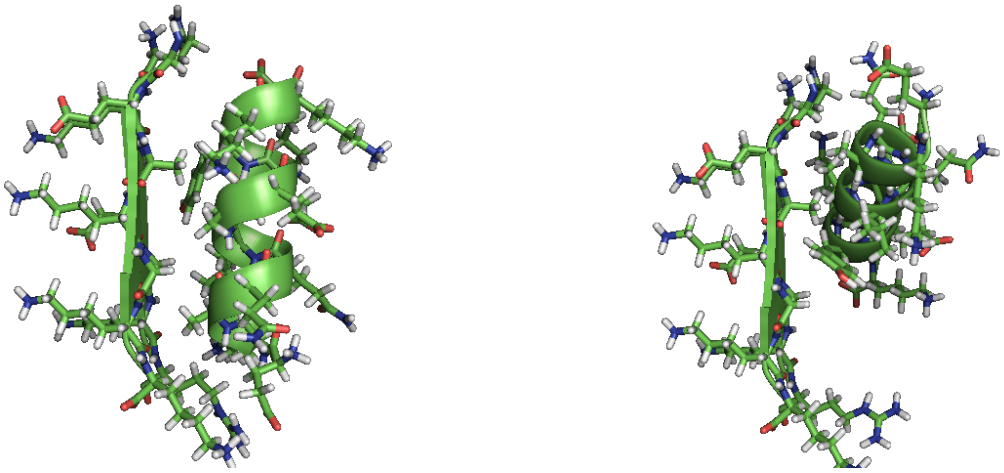

Figure 10: An example of designed (left) and folded (right) structure that demonstrate the apparent contradiction in the data. The designed structure on the left has a high contact degree and appears in nature. The example folded structure on the left has a high contact degree and never appears in nature. But the all folded structures are optimized by Rosetta as conformations within 0.5Åof the structure shown on the right. Visual inspection would not lead one to believe that such a structure is stable.

tact degree and abundance are each correlating with some unique features of Rosetta designability. Using both contact degree and abundance shows a strong correlation with Rosetta metrics. Abundance evaluates how often the given motif occurs in nature. These motifs exist within other protein structures which affect its folding. Thus abundance is not a perfect measure of the isolated folding of a motif.

Our results thus far hint towards the distinction between a motif existing as part of a structure and being folded independent of that structural context. The concept of "independent designability" refers to the ability for a motif to fold independently of other structural context. Incorporating contact degree as an additional indicator along with abundance helps account for the independent designability of a particular motif. Even if a motif is abundant in nature, it may not exhibit high designability in isolation. High contact degree ensures that there are sufficient residue contacts to allow interaction and to be designed upon. Thus high abundance of a motif perhaps does not correlate directly with designability of the particular motif in isolation. But rather the appropriate environment of the motif must be considered. Independent modeling of these structures may not represent the designability of the motifs as part of a larger structure. Motif abundance may be a better indicator of designability when motifs are examined as building blocks to a larger structure.

Our data reveals two cases that contradict our hypothesis: (1) The motif is abundant but not designable or (2) The motif is designable but not abundant. Each of these cases can potentially be explained by the unaccounted context of the motif. In order for a two chain motif to be abundant in nature, it must be stable and capable of being completed within a structure. The first case can be explained by motifs needing other structural context for stability as discussed earlier. In nature, each of these motifs could have external contacts that might help stabilize a particular conformation and thus make it more designable in that context. In case two, structures that are highly designable, but not abundant might suffer from the opposite problem. Although the motif is stable *in silico* on its own, it is difficult to integrate into a larger structure. In order to be abundant, the motif itself must be common and there must be a designable way in which

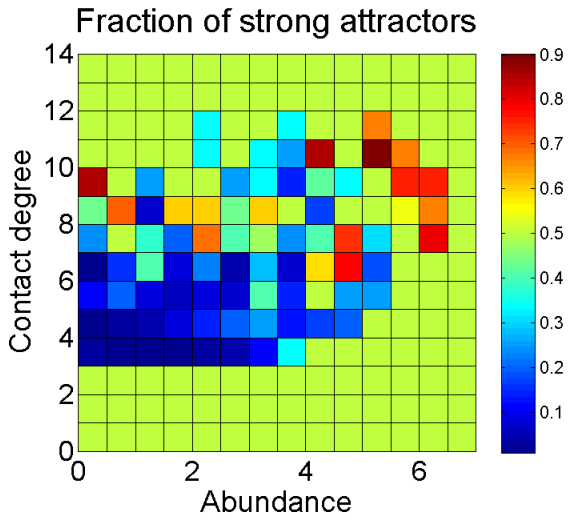

Figure 9: The graph shows the average fraction of structures that are strong attractor (> 7 neighbors) correlated by contact degree and abundance. The fraction has been smoothed assuming 0.5 naive return rates.

to integrate the motif. To further examine this characteristic, we attempted removing all motifs with zero abundance to account for the possibility that there was no way to "complete" the motif in a structure. Even with this correction, the graphs changed only minimally, suggesting that the external designability is a similarly complex feature. Simply having at least one way to complete the structure may be insufficient. Thus in order to understand the designability of a motif, we must analyze both the internal and external designability. Designability as a product of abundance cannot be completely understood as a single independent motif, but rather must be examined as an entire structure. Future work, may explore the manners in which abundant motifs compose a structure and how that correlates to a structure's overall designability.

5 Conclusion

In summary, the data support a positive correlation between abundance and *in silico* designability, while drawing attention to some edge cases. This correlation is enhanced by the additional information provided by contact degree. Although higher abundance structures tend to exist within folding funnels, the funnel is often not around the structure for which the sequence was designed.

6 Future Work

A major limiting factor in this study was the sample size of the structures (particularly on the upper spectrum of both abundance and contact degree). For many of the plots examined, the most interesting regions (high abundance and high contact degree) contained less than ten designed structures. Future work should examine a larger structure set, which would help address a number of issues when examining these designed structures. More structures would also allow a more data driven form of clustering to determine bins. A more disciplined structure generation method may give more control over the abundance of structures produced, but also risks introducing biases into the structure set. The maximum abundance of designed and folded structures was noticeably higher than the initially randomly generated maximum, indicating that random sampling alone is insufficient for generating structures that represent the whole range of motif abundance.

This study only examined motifs composed of one alpha helix and one parallel beta sheet. The other five possible pair combination of helix, parallel sheets and anti-parallel sheets should also be examined in a similar manner. This may reveal characteristics specific to certain secondary structure interactions. Furthermore, it would be interesting to evaluate motif abundance as an indicator of designability on an entire structure rather than just motifs in isolation.

7 Acknowledgements

I would like to thank Professor Gevorg Grigoryan and all the lab members for their expertise and guidance throughout this project.

References

[1] Anfinsen, Christian B. "Principles that govern the folding of protein chains." Science 181.4096 (1973): 223-230.

[2] Beichl, Isabel and Francis Sullivan. "The Metropolis Algorithm."Computing in Science and Engineering, January/February 2000: 66-69.

[3] Butterfoss, Glenn L., and Brian Kuhlman. "Computer-based design of novel protein structures." Annu. Rev. Biophys. Biomol. Struct. 35 (2006): 49-65.

[4] Chaudhury S., Lyskov S. & Gray J., "PyRosetta: a script-based interface for implementing molecular modeling algorithms using Rosetta," Bioinformatics, 26.5 (2010): 689-691.

[5] Fernandez-Fuentes, Narcis, Joseph M. Dybas, and Andras Fiser. "Structural characteristics of novel protein folds." PLoS computational biology 6.4 (2010): e1000750.

[6] Govindarajan S, Goldstein RA. "Why are some proteins structures so common?" Proc Natl Acad Sci USA. 1996;93:3341–3345.

[7] Helling, Robert, et al. "The designability of protein structures." Journal of Molecular Graphics and Modelling 19.1 (2001): 157-167.

[8] Lovell, Simon C., et al. "The penultimate rotamer library." Proteins: Structure, Function, and Bioinformatics 40.3 (2000): 389-408.

[9] Wingreen, Ned S., Hao Li, and Chao Tang. "Designability and thermal stability of protein structures." Polymer 45.2 (2004): 699-705.

[10] Yang, Jian-Yi, Zu-Guo Yu, and Vo Anh. "Correlations between designability and various structural characteristics of protein lattice models." The Journal of chemical physics 126.19 (2007): 195101.

[11] Zhang, Jian, and Gevorg Grigoryan. "Mining tertiary structural motifs for assessment of designability." Methods in enzymology 523 (2013): 21.

[12] Zheng, Fan, Jian Zhang, and Gevorg Grigoryan. "Tertiary Structural Propensities Reveal Fundamental Sequence/Structure Relationships." Structure (2015).

[13] Zhou, Jianfu, and Gevorg Grigoryan. "Rapid search for tertiary fragments reveals protein sequence–structure relationships." Protein Science (2014).