Bobby Tesoriero

COMP 570 Final Paper

12/15/19

Understanding the Antiparallel β-Sheet Motif in β-Lactoglobulin and Ovalbumin

Introduction

The correct folding of proteins into functional conformations is perhaps one of the most complex and critical concepts in Biochemistry, and is arguably one of the most important, considering the critical importance of proteins in all modern living systems. As such, it is imperative to understand the delicate interplay of forces that drive protein folding, and the design principles behind how these folded states show functionality *in vivo*. Equally important is the analysis of what happens when protein folding goes wrong; what does it mean for a protein to 'misfold', and what factors may cause a protein to undertake this transition.

When looking at this question from a high-level, one of the first ideas that one may uncover is that of the 'protein folding funnel', also known as the 'folding landscape'. Coined by Onuchic *et al.* 1992, the folding funnel hypothesis dictates proteins fold to minimize their own free energy, with the native state of the protein being located at the global minima of this free energy well.[1] Numerous misfolded states also reside in minima, occurring in the many possible local minima present in the folding funnel. As such, introductions of energy to a protein can cause it to get 'stuck' in one of these misfolded states.

Table 1. Sequences from control proteins. short segments aligned Higher with eachother, whereas full sequences are the longest encapsulating sequences with the same sheet structure.

		Segment 1	Segment 2	Segment 3
β-Lactoglobulin	**Short**	YSLAMMA	LFCME	VCQC
	Full	YSLAMMA	YLLFCME	VLFFGRC
Ovalbumin	**Short**	SMLVLLP	LFCIK	FFGR
	Full	MSMLVLLP	FLFCIKH	VLFFGRC

As one might expect, the shape of a protein's folding funnel is primarily dictated by the amino acid sequence. Each amino acid possesses unique electronic and physical properties, and when formed into a polypeptide yield the seemingly limitless emergent properties found in proteins. This level of possibility comes at a cost; even in small proteins, there is an astronomical number of possible conformations in a single peptide chain, which makes the prediction of a peptide's structure based on sequence extremely computationally expensive, requiring massive server networks to obtain significant results.[2] Then how does nature fold proteins so quickly and effortlessly? The answer lies in protein nucleation, where small segments of residues promote folding of the entire chain, and by the same line of reasoning, may also be a critical factor in protein misfolding.[3] As such, being able to detect folding nucleation sites on a general scale would be a massive boon to our knowledge of protein folding.

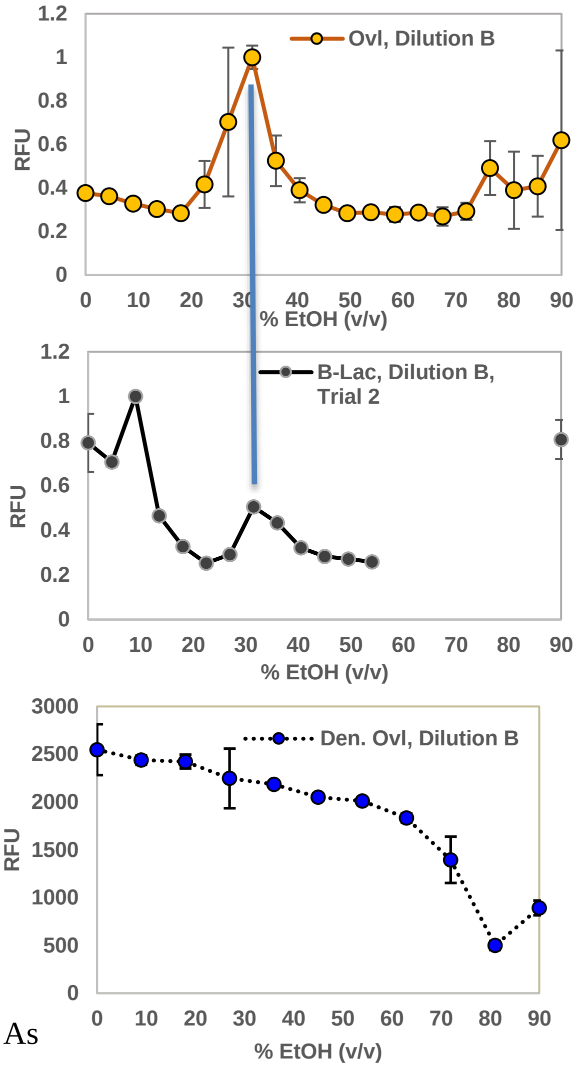

Figure 1. Thioflavin-T measurements of (A) Native ovalbumin, (B) Native β-lactoglobulin, and (C) denatured ovalbumin. Denatured β-lactoglobulin (not shown) shows the same behavior as (C).

In a culmination of the concepts introduced above, prior work (Tesoriero, unpublished) has uncovered a potential nucleation site in ethanol-induced aggregation of β-lactoglobulin and ovalbumin. By incubating samples of protein in different concentrations of ethanol and analyzing structure fluorometrically using the pleated-β-sheet-binding die thioflavin-T, it becomes possible to find specific concentrations that drive a particular spike in aggregation (Figure 1A, B). These aggregation peaks also appeared to be structure-dependent, as when the native structure was eliminated, as were the aggregation patterns (Figure 1C). By then cross referencing a library of these plots for different proteins, analyzing each in terms of sequence and structure, we then were able to deduce a specific region of shared structure and similar sequence in β-lactoglobulin and ovalbumin (Table 1).

In this paper, a computational follow-up to these experiments were carried out in MATLAB to uncover additional evidence of aforementioned aggregation nucleation site, utilizing large-scale BLAST and Protein Data Bank searches to analyze a library of segments similar to those found in Figure 1C. Proteins found to contain different combinations of segments in the same structural conformation as the test proteins are then proposed as possible candidates for aggregation testing. Because of the experimental aggregation behavior, one would expect that specific hydrophobicity values are critical in obtaining the aggregation peaks seen above, and so several calculations were performed to confirm this expectation.

Methods:

BLAST Searches: For each of the short sequence segments found for each protein, as well as sequences containing the immediate structurally-identical residues surrounding and including each segment (Figure 1C), BLAST searches were performed for the top 2000 hits using the ncbiblast() function, with parameter values optimized for short query sequences. Search results were automatically saved in .xml format, allowing for easy access using the blastread() function. Hits were limited to proteins found in Protein Data Bank (PDB), for homogeneity of results.

Creation of Structures Containing Relevant Protein Data: Using the gathered BLAST data, several pieces of information for each hit in a particular search was gathered into a large structure by first getting BLAST data from file for each hit using blastread() (up to a user-defined number of hits), and adding the PDB ID, chain identity, full chain sequence, alignment indices, alignment sequence, and alignment score to the struct directly. Then the PDB data for each hit was gathered using the PDB ID number with the getpdb() function. Using the obtained PDB structure, various metrics were obtained including the secondary structure, Full and average Kyte-Doolittle hydropathy scores (a metric for hydrophobicity[4]), central crystal structure coordinates, as well as a distance measurement between the aligned segment and the appropriate β-Lactoglobulin and Ovalbumin Sequences.

Secondary Structure Determination: Indices for each instance of β-sheets and Helices in a protein were first obtained from the PDB structure. Then, each index in the aligned sequence was compared against structural indices, and was assigned a matching letter representing the structural conformation for that residue. To simplify the algorithm, only three different conformations were considered: β-sheet, Helix, and other, represented by 'B', 'H', and 'X', respectively. The percentage contribution of each structure was calculated from these measurements. Note that there are conflicts between the MATLAB returned PDB structure and the online PDB information, with MATLAB results occasionally showing misaligned structure indices, which leads to errors in calculations.

Kyte-Doolittle Hydropathy Calculations: Each of the 20 amino acids were pre-assigned a hydropathy score, directly adapted from Kyte, J. and Doolittle, R.F 1982. The aligned sequence was then compared against this index, sequentially calculating the total hydropathy and average hydropathy scores for the segment.

<u>*Central Crystal Structure Coordinates*</u>*:* For each hit, the central X, Y, and Z components of the segment from the PDB crystal structure data was obtained by directly grabbing the coordinates of the alpha carbon in the central residue (s). For odd-numbered segments, these coordinates directly match the middle residue, whereas in even-numbered segments, the average X, Y, and Z components of the two centermost residues was used. Note that for some PDB entries, it is impossible to find said coordinates, such as when the entry represents the superposition of protein conformations in response to stimuli, and as such the coordinates were recorded as NaN. For Proteins that contained more than one of the aligned segments, the root mean square distance was calculated using these coordinates using the following equation:

<u>*Superposition Distance Measurements*</u>*:* Two Distance measurements as defined by the pdbsuperpose() function to represent the sum of squared errors and the root mean square distance between the superimposed segment and either the associated β-lactoglobulin and ovalbumin segment were calculated using the PDB info for each hit. Note that the short length of the segments can lead to issues in the superposition process, either leading to empty results or false

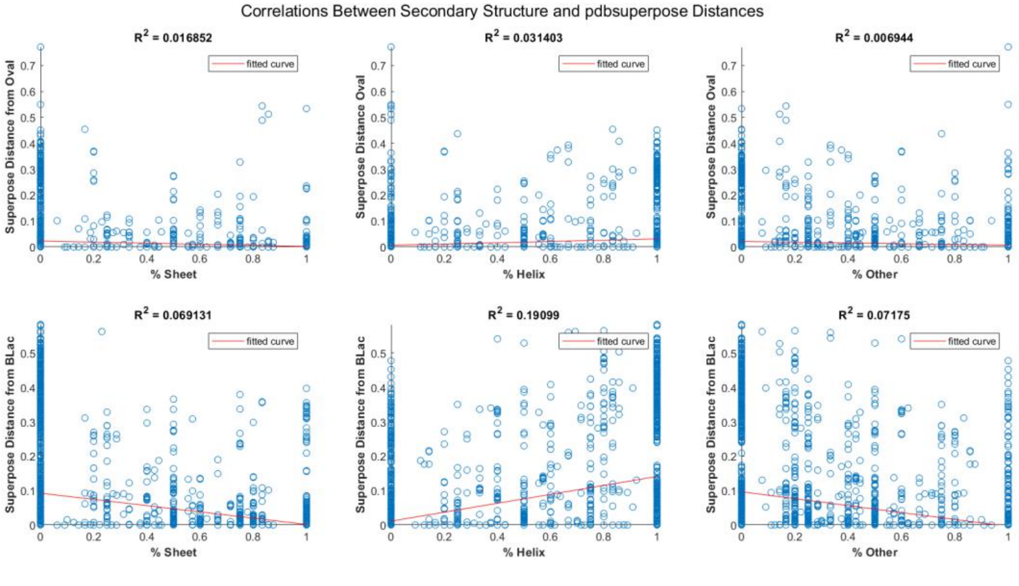

Figure 2. Percent Secondary Structure plotted against pdbsuperpose Distance for each control protein.

positive results of 0. Since this is an issue with the function itself, this intrinsic source of error was accounted for by way of the β-sheet content of each segment.

Results:

Table 2. Average Hydropathies for short and full segments.

	Short Seg. Avg. Hydropathy	Full Seg. Avg. Hydropathy
β-Lactoglobulin	1.4036	1.6659
Ovalbumin	1.4240	1.6845

Analysis between secondary structure content versus the pdbsuperpose distance and RMSD outputs overall, regardless of segment origins can be found in Figure 2. As the Distance measurements are compared against ovalbumin or β-lactoglobulin, with the segments for both of the aforementioned proteins are completely β-Sheet *in vivo*, one would assume that there would be a linear correlation between distance and secondary structure, however in all cases no significant correlation was found.

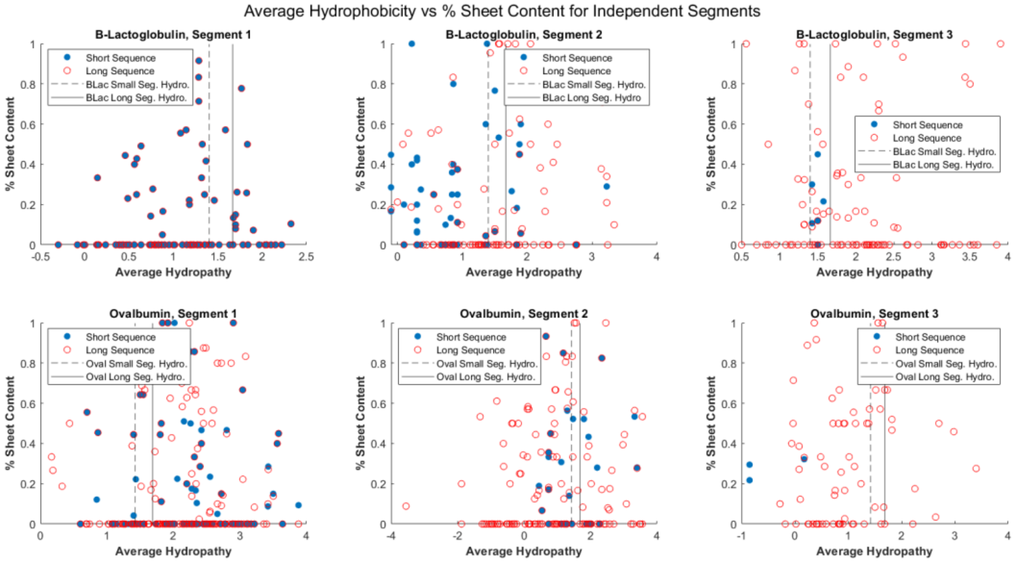

Figure 3. Comparisons between the Average Hydropathy and percent sheet content for each segment. Blue filled circles represent the short sequence hits, and empty red circles represent the longer sequence hits. Dashed and Solid vertical lines represent the hydropathy values for the relevant short and long segments, respectively.

Average hydropathy compared against percent sheet content for each segment can be found in Figure 3. Average hydropathy values for the short and long segments of the appropriate control protein are shown as the vertical dotted and solid lines, respectively. There is no clear correlation between the Average Hydropathy and percent sheet content of the segments at this stage in analysis.

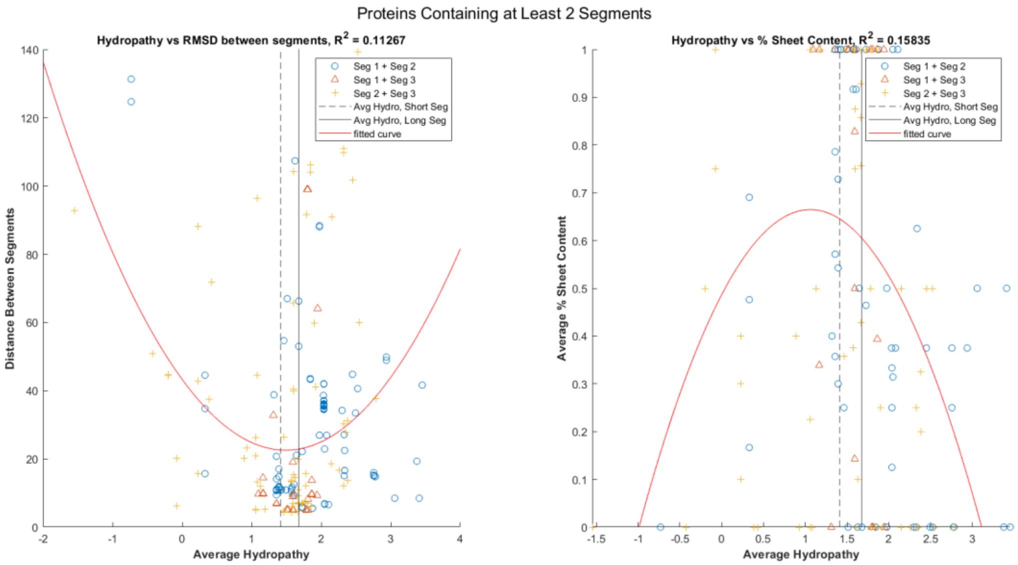

Figure 4. Comparisons between average hydropathy, root mean squared distance between segments, and percent sheet content for proteins containing two or more aligned hits. Dashed and solid vertical lines represent means between the average hydropathies for the short and long sequences, respectively.

Narrowing down the results from Figure 3, Figure 4 shows the correlations between the average hydropathy, root mean squared distance and percent sheet content for all protein hits that aligned with two or more segments. There was no cross-pollination between lists of BLAST hits; that is, only matches from within the same original protein were considered in this and future analyses. Note that the average hydropathy plotted in Figure 4 represents the average between the two aligned segments present in the protein. Additionally, the vertical lines represent the average between the short or long hydropathies for β-lactoglobulin and ovalbumin. Given the very similar values between the two, averaging them simplifies the plot without changing the interpretation (Table 2). While the correlation is still not significant, there is a slight quadratic pattern emerging in both groups.

Finally, the correlations between average hydropathy (calculated in the same way as for Figure 4), root mean square distance, and percent sheet content were examined for all proteins showing hits for all three segments from the assigned protein list (Figure 5).

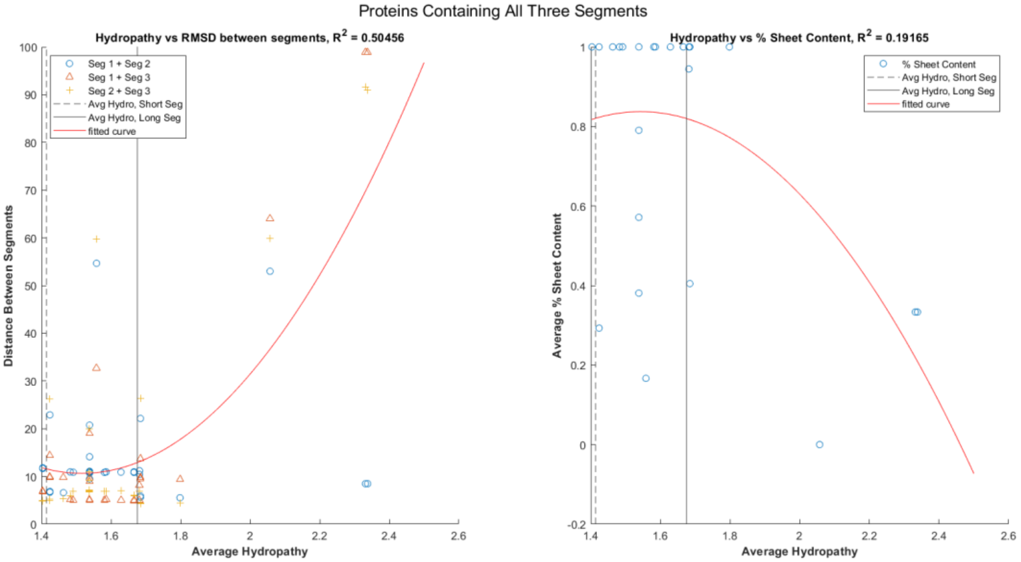

Figure 5. Comparisons between average hydropathy, root mean squared distance between segments, and percent sheet content for proteins containing three aligned hits. Dashed and solid vertical lines represent means between the average hydropathies for the short and long sequences, respectively.

Compared to the plots above, there is a much more significant correlation between average hydropathy and root mean squared distance between segments, with a local minimum located right between the average hydropathies for the short and long control segments. All segment point RMSD values represent the distance between only two out of the three aligned segments. Overall however, regression analysis shows that there is some statistical significance supporting that the aligned segments with average hydropathies close to that found in the control proteins are more likely to be located adjacently *in vivo*. While the correlation between average hydropathy and percent sheet content does not show nearly as much significance, it is still more than the figures above, possibly suggesting that if more points were added, they may follow a similar trend to the current fit. Note that there appears to be some discrepancy between the secondary structures shown in the PDB database online and the returned MATLAB structure, generally underestimating the percent sheet content. If this were to be corrected for, it would be expected that some of the low-sheet points between the two vertical lines would approach 100% β-Sheet.

To confirm that the aligned segments all share similar hydropathy to each other, and thus would make more sense to self-assemble *in vivo* the average and standard deviation of the hydropathy values were plotted against each other and fit to both linear and quadratic functions, each further supporting a minimum of standard deviation within the range of the control segments (Figure 6). There seems to be moderate statistical evidence for a quadratic model to fit this data, however without the upper left two points, a linear model would fit much better, so more data would have to be obtained to confirm the behavior of the data.

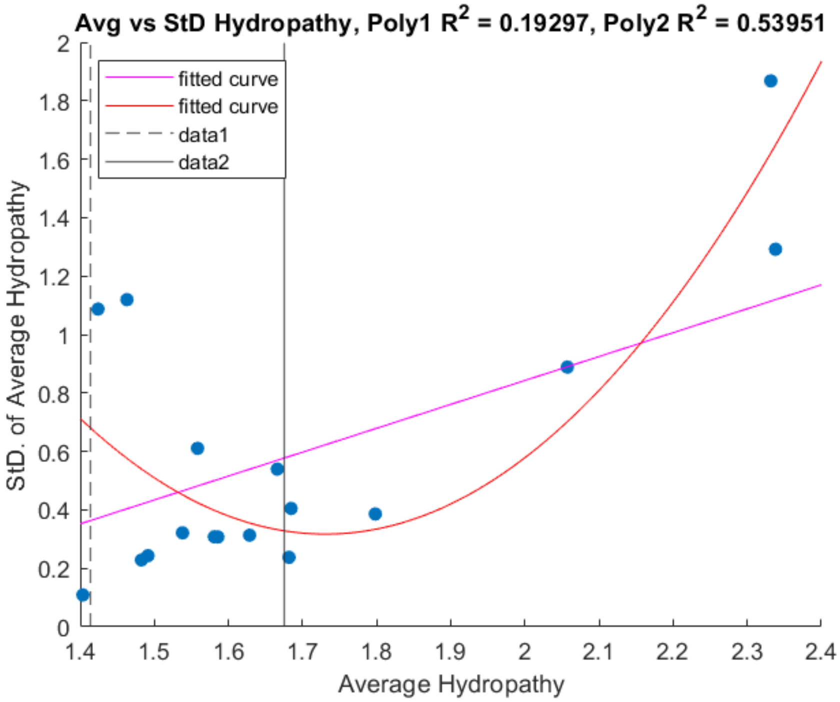

Figure 6. Comparison between the average and standard deviation of hydropathy values for proteins aligning to all three segments.

Discussion:

By sequentially analyzing the proteins gathered from BLAST searches with increasing alignment between segments, it becomes clear that this motif shared between β-lactoglobulin and ovalbumin is at least partially formed as a function of the average hydropathy of the region. While this is a simplified approach to understanding and explaining the complex dynamics behind protein folding, it logically follows that segments of a protein that show similar hydrophobicity would react similarly when placed in a more hydrophobic solution, such as 31.5% ethanol.

Furthermore, the degree to which these segments are buried likely result in the aggregation pattern seen in Figure 1; many cases of aggregation occur when a protein ‘flips inside out’ to expose its hydrophobic residues to the environment, and thus a region of approximately uniform hydrophobicity would result in a simultaneous inversion and aggregation of the proteins structure. This uniform hydrophobicity between segments is exhibited well in Figure 6, where the segment groups within the bounds of the control appear to show low standard deviation relative to segment groups outside of the control.

That being said, there are several ‘trouble spots’ within the program that should be fixed before this work is finalized. Importantly is the intrinsic error within the secondary structure indices needs to be corrected for somehow, with a method more efficient than manually changing errors (disclaimer: manually ‘fixing’ data points was not performed in this work). Also, the discrepancy between pdbsuperpose measurements, and secondary structure measurements should be understood, to know how to most effectively use all possible tools for analysis. It would also be beneficial to sort out duplicates of the same protein; PDB contains several entries for each protein (generally with small mutations), and it may be important to confirm that the results do not change when duplicate or semi-duplicate points are removed.

To lend further support to the argument that localized uniform hydrophobicity within the center of a protein leads to some form of aggregation peak, a two pronged approach must be taken. First, more data points must be obtained computationally, with less intrinsic error than what is presented in this work. This can be done simply by expanding the BLAST result list, or artificially engineering segments with the desired hydrophobicity to query into a BLAST search. Secondly, said data points must be tested experimentally, in order to produce a full, accurate representation of this effect.

Overall, while these experiments may seem mundane and unimportant to some, it is critical to analyze every aspect of how proteins fold and misfold to achieve a complete understanding of Biology, and this work hopefully leads us one inch closer to that goal.

References

1) Leopold, P.E., Montal, M., and Onuchic, J.N. (1992) Protein folding funnels: A kinetic approach to the sequence-structure relationship. *Proceeds to the National Academy of Sciences* **15,** 89, 8721-8725.
2) Rohl, C.A., Strauss, C.E.M., Misura, M.S., Baker, D. (2004) Protein Structure Prediction Using Rosetta. *Methods in Enzymology* **383**, 66-93.
3) Shakhnovich, E., Abkevich, V. & Ptitsyn, O. (1996) Conserved residues and the mechanism of protein folding. *Nature* **379**, 96–98.