Periodicities in Sequence Residue Hydropathy and the Implications on Protein Folds

Nancy Zhang

March, 2000

Biochemistry 118

I. Introduction

The deterministic folding of a polypeptide sequence into its convoluted 3-D structure is one of the most fascinating applications of nature's laws. With the current growth in the size of protein sequence databases and the distribution of sequence analysis tools on the internet, the classic problem of predicting a protein's structure from its amino acid sequence is becoming increasingly important. Currently, discounting homologies of over 35% identity, there are over 40,000 protein sequences identified, and yet only 4200 experimentally-determined protein structures. Being able to predict proteins structure from sequence is crucial to many fields of study, such as ligand-protein docking, as well as to the understanding of protein function at the molecular level.

The underlying hypothesis that motivates prediction efforts is that the complex packing arrangements of the main chain and side chains atoms of a folded protein is uniquely determined by two factors: its amino acid sequence and its folding environment. This has been supported by numerous experiments (1, 2), and is the foundation for current sequence analysis methods such as homology search, multiple sequence alignment, and motif identification. These methods evolve around the idea that if two proteins are similar in sequence, then the chances are high that the two proteins are similar in structure as well.

Despite decades of research, the accuracy of current methods is only around 60% (3). One of the main problems limiting the success of current prediction algorithms is that there are hidden variables effecting the protein folding mechanism that are not explicitedly accounted for in the algorithms. Non-local residue interactions is one of these hidden variables; to account for all such interactions would be impossible (more on this later). Solvent-chain interactions is another hidden variable, which many prediction algorithms often neglect. It has been shown that the propensity of amino acids for a certain secondary structure is environment dependent, and in particular, is dependent on its solvent accessibility (4, 5). Yet, since the solvent accessibility of a residue in the chain depends on the final folded structure, it is very hard to explicitly and fully acount for the solvent effect in structure prediction algorithms.

Although it is very hard to model the global characteristics of an amino acid sequence through pairwise interactions between residues, it is possible to represent them as frequencies- periodic patterns that span the entire sequence. The discrete Fourier transform has been used to find such patterns in the hydropathic content of sequences, and distinct frequencies in hydrophobicities have been identified to be strongly correlated with certain secondary structural elements (6, 7). The possibility that the two important "hidden variables" – solvent effect and non-local interactions – may be better represented in the frequency domain inspired the content of this paper. Can we use sequence alignments in the frequency domain to predict structural similarities between proteins? Do two proteins that are similar in structure necessarily have similar peak patterns in their hydrophobicity plots?

In section II of the paper, I will give a more detailed description of the solvent effect and explain why it is crucial to a protein's fold. In section III, I will explain how the Fourier transform simplifies the task of representing global sequence characteristics, and argue benefits of sequence analysis in the frequency domain. Finally, in sections IV and V I will describe the procedures and results of an experiment in which I tried to find a correlation between the structural distance of proteins and the distance in their frequency domain hydrophathy plots.

II. Solvent Effects on the Protein Folding Process

Although some protein structure prediction methods account for the hydropathic characteristics of the amino acids in their scoring functions. It is not yet sure how to explicitly model the solvent effects into fold prediction algorithms. However, studies have shown that the solvent plays a major role in the folding process. Just as a ball sliding along a rolling terrain, the folding chain continuously seeks for a local minimum in conformational free energy, given by the equation:

$$\Delta G = \Delta H - T\Delta S,$$

In vacuo, the nonconvalent binding energies between residues compete with chain entropy.

$$\Delta G_{chain} = \Delta H_{chain} - T\Delta S_{chain}$$

However, when the native, aqueous environment of the protein is taken into account, the equation becomes much more complicated:

$$\Delta G_{total} = \Delta H_{chain} - T\Delta S_{chain} + \Delta H_{solvent} - T\Delta S_{solvent}$$

The following table (8) shows the relative magnitudes of each for a folding chain in different environments:

	ΔG_{total}	$T\Delta S_{chain}$	$\Delta G_{transfer}$	ΔH_{chain}	$T\Delta S_{solvent}$	$\Delta H_{solvent}$
Polypeptide chain in vacuum	↓	↑		↓		
Nonpolar groups of chain in aqueous solvent	↓	↑	↓	↑	↓	↓

In the table, $\Delta G_{transfer}$ is the change in free energy in transfering a nonpolar side chain from water into the protein interior. It is clear that, in an aqueous environment, the energy gain from the interaction between side-chain and solvent $\Delta G_{transfer}$ accounts for a large contribution to protein stability. Moreover, the interaction between chain and solvent are of utmost importance in protein folding, elucidated by the fact that almost all proteins denature in ethanol or in aqueous urea (8).

The interaction between the peptide chain and the aqueous solvent depends on the hydropathic character of the residues in the chain. Amino acids with non-polar side chains, such as methionine and valine, energetically prefer to reduce their contact with water, while those with charged and polar side chains generally prefer to be immersed in the aqueous solvent. Thus, amino acids with hydrophobic side chains tend to be buried in the internal core of a globular protein, while those with hydrophilic side chains tend to reside on the surface. This tendency to minimize the accessible surface area of hydrophobic particles, and maximize that of the hydrophilic particles, is a major driving force in protein folding.

Various scales have been developed to measure the hydrophobicity/hydrophilicity of each of the twenty amino acids. Some scales, such as that of Janin (9) and Rose, et al. (10), are derived from examining proteins with known 3-D structure and defining the hydrophobic character of an amino acid as its tendency to be in the protein core as opposed to be on the surface, while others, such as that of Wolfenden, et. al. (11) and Kyte & Doolittle (12), are derived from the physio-chemical properties of the amino acids, such as the $\Delta G_{transfer}$ value of transfering the residue from a neutral, non-interacting solvent such as ethanol to water (in fact, it has been debated whether or not ethanol is a perfectly neutral solveng, see Kyte & Doolittle (12)). Due to the difference in their evaluation schemes, the scales vary significantly in their scoring of the amino acids.

Much work has been done to test for the importance of hydrophobicity/hydrophilicity in protein folding. There has also been much debate in this area. To begin with, a study by White and Jacobs in 1990 (13) contended that the distribution of the hydrophobic residues along the chain cannot be distinguished from that expected for a random distribution for a vast majority of soluble proteins, and thus, sequence hydropathic patterns are not a significant indicator of its structure. However, in the experiments of Cornett et al. (6) and Eisenberg et al.(7), it was shown using helical wheels and hydrophobic moments that patterns in amino acid hydrophobicity accurately detects amphipathic structures in proteins. Furthermore, the results of an experiment by Xiong et al. (5), showed that the hydropathic character of sequence residues has a larger effect on the sequence's choice for alpha-helix or beta-sheet, as compared to the intrinsic propensities of the amino acids for a particular secondary structure. In all contexts, the debate seems at present to favor the fact that a sequence's hydropathic pattern does effect its structure.

III. Representing global correlations among residues using Fourier analysis

The main drawback of current prediction algorithms is that they ignore the interactions between residues that are far apart in sequence. The Chou-Fasman algorithm assumes independence between any pair of amino acids, and most other algorithms, such as nearest neighbor and neural networks, use the "fixed-window-size" approach., assuming independence between residues inside and outside the window. The obvious explanation for these simplifying assumptions is that any algorithm that considers the interactions between all pairs (not even including triplets and multi-plets) of residues would be be NP hard in that its run-time would be exponential with respect to the length of the sequence. Furthermore, the problem of adjusting the parameters for such an algorithm would also be NP hard.

It is therefore necessary to steer away from the attempt to try to represent the global interactions in the seqpence as correlations between pairs of residues. Another approach is to seek for global patterns in the sequence, represented as periodicities in residue characteristics. A radio wave has a unique representation in both the time and frequency domains, with certain wave-characteristics that are obscured in the time domain elucidated in the frequency domain. If we can represent an amino acid sequence in its "frequency" domain, we may also discover some surprising results. Fourier Analysis has been applied by many scientists taking exactly this approach (7, 14).

Given that an amino acid sequence of length N can be represented by a sequence of numerical values R = $\{r_i\}$, i = 1…N, the Fourier transform of R would be:

$$\boldsymbol{F}(\mathrm{R})_i = \sum_{(j=1\ldots N)} r_j e^{(-2\pi ij/N)} \quad , \quad i = 1\ldots N$$

The resulting $\{\boldsymbol{F}(\mathrm{R})_i\}$ would be a complex vector in R^N. It would be convenient for analytical purposes to take the absolute value of this vector:

$$\boldsymbol{F}(\mathrm{R})_i = |\,\boldsymbol{F}(\mathrm{R})_i\,|$$

and result in the power spectrum of the original sequence in the frequency domain.

The function f(x) = 1 is the ideal "global function": everything that is true for f(x) at $x=x_0$ is also true for f(x) at every other point x. The Fourier transform of f(x) is the impulse function, $\delta(x)$, which can be thought of as only having local characteristics (at x=0). This exemplifies the fact that through the Fourier transform global features collapse into local features. This is exactly why Fourier analysis has the potential of great use in protein sequence analysis.

One major problem that we face in this approach is that there is yet no agreed upon best mapping from the amino acid sequence to R, the sequence of reals. What physical properties of the amino acids ought to be selected to encode in R? We can never be assured that enough properties has been represented, and in correct proportions, to completely describe the physio-chemical tendencies of the amino acids. One of the best mappings found so far is described in the results of Kidera et al (15), who carried out a factor analysis of essentially all of the data sets available for the amino acids, and were able to demonstrate that all of these data could be represented by a set of 10 factors. Using this representation, Rackovsky (14) identified a harmonic series of over-expressed frequencies that can be used to identify the TIM barrel. structure group. The complexity and size of the Tim Barrel makes it impossible to identify using sequence analysis that disregard nonlocal, periodic information.

IV. Procedure to Test for Correlation Between Periodic Hydropathic Features and Structure

In the remainder of this paper, I'll present the experiment in which I tested for correlation between the sequence periodicities of structurally similar proteins. In order to define the problem at hand, we need to first represent each amino acid as a numerical value. The problems that we face in this task has already been discussed in section IV. Here, we will simplify the procedure by considering only one property of the amino acids: their hydropathy. The reliance of the protein folding mechanism on the hydropathic patterns in its sequence has been widely studied in the last two decades and is delineated in section III. Considering that sequence hydropathic patterns is one of the major driving forces in protein folding, we expect that proteins with similar structures have similar hydropathic periodicity graphs. I used the Kyte & Doolittle hydropathy scale to quantify each of the amino acids, and experimented with the following two mappings from the amino acid sequence $\{A_i\}$ to their numerical representation $\{a_i\}$ (which we will here on after call the "hydropathy plot" of the sequence) :

1. M_1: $\{A_i\} \longrightarrow \{a_i\}$

$$a_i = \sum j=(i-u)\ldots(i+u)\ S(A_i)$$

2. M_2: $\{A_i\} \longrightarrow \{a_i\}$

$$a_i = \sum j=(i-u)\ldots(i+u)\ S(A_i)(1-\lambda)\lambda^{|j-i|-u}$$

$\lambda < 1$
$u = W/2$ (W = window size)
S(A) is a mapping from amino acid A to it hydropathic index in the Kyte & Doolittle scale.

M_1 simply maps each residue Ai to the average of the hydropathies amino acids within a certain window of size W centered around Ai. This "smoothes" out the hydropathy curve and eliminates noise in the data, and is better suited for evaluating the hydropathic characteristics of regions in the segment at a birds-eye view. M_2 places more emphasis on the individual hydropathies by polynomially decreasing the weight, by a factor of λ, on each amino acid going further away from the center of the window. Thus, M_2 may capture site-specific characteristics of the sequence, which is important in predicting features such as the accessible surface area (ASA) of the residues in the sequence's folded form. The ASA of residues along a folded chain changes rapidly. Thus, the periodic changes of buried and exposed residues with short periods of 3.6 and 2.0 residues play a significant role in the formation of alpha-helix and beta-strand, respectively (16).

From now on, we distinguish between hydropathy plots H in the sequence domain as simply the output of one of the mappings M_i above, and hydropathy plots H^ in the frequency domain as the absolute value of the Fourier transform of H. We now face the issue of finding an appropriate distance measure in the space of all hydropathy plots in the frequency domain. Given an alignment of two hydropathy plots, their distance can be taken as simply the mean of the absolute value difference of all of the aligned residues, minus a penalty for the unaligned residues. Ideally, we would allow gaps in the alignments. However, to simplify the task, we disallow gaps and only consider those alignments that have an outlie (fraction of residues not aligned) within a fixed parameter μ..

Correspondingly, we need to find the distance between any two sequences in structure space. That is, how would we measure the degree to which two structures differ? There is a vast array of structural alignment tools available on the web, each with its own alignment algorithms and scoring functions. I used the LOCK superposition program developed by Brutlag Informatics Group (17). This program allowed me to compare the structure of one sequence against a PDB subset that is at most 25% similar in homology to the query sequence. The program outputs, for each target protein, the number of aligned alpha-carbons and the rmsd deviation of the alpha-carbon positions of aligned atoms. To combine these two figures into one score, which we'll call the "structural distance" between the proteins, we will use the equation:

$$\text{Structural distance} = (\text{rmsd of aligned atoms}) + \frac{(\text{average num. of aligned atoms})}{10^{*}\ (\text{num. of aligned atoms for target protein})}$$

Having defined the hydropathy plots and distance measures, the hypothesis for the experiment is that there is a positive correlation between the distance in hydropathy plots of two sequences and the distance in their structures. I used the following procedure to test this hypothesis:

1. Conduct LOCK alignment on a query protein against a PDB subset of maximum 25% homology.
2. Obtain the hydropathy plot of the query protein and each protein returned by the alignment.
3. Obtain the frequency power spectrum of the hydropathy plots.
4. Obtain the frequency domain hydropathy distance (hdp_distance) between the query protein and each of the proteins in the PDB subset of step 1.
5. Plot the hdp_distance versus the structural dstance (obtained in step 1) to find any correlations that may exist.

Given the shortness of the time for this project, I only experimented with two proteins as the query protein: human hemoglobin (PDB code 1BAB strand B), which is a mostly alpha protein belonging to the structural class Globins, and amicyanin (PDB code 1AAJ), which is a mostly beta protein belonging to the Greek-Key-I structural class.

V. Results

The hydropathy distance versus structural distance plots, with sequence hydropathy plot calculated using M_1 and M_2, are given in the appendix. Although the plots are very scattered, they seem to indicate a weak positive correlation. Comparing the plots for M_1 and those for M_2, we can observe that those for M_2 are more scattered. This is indeed true when we compare the correlation coefficients:

	M_1	M_2
Hemoglobin	0.25	0.20
Amicyanin	0.44	0.37

The higher correlation of the plots obtained using the M_1 hydropathy mapping suggests that, in the structural alignment algorithm used by the LOCK program, general periodic trends in the sequence is more important than site-specific features. Due to the low correlation values, we can not be certain of the statistical significance of the suggestive positive slope of the linear regression line. Thus, the results obtained so far from the two tests are not strongly in favor of the hypothesis.

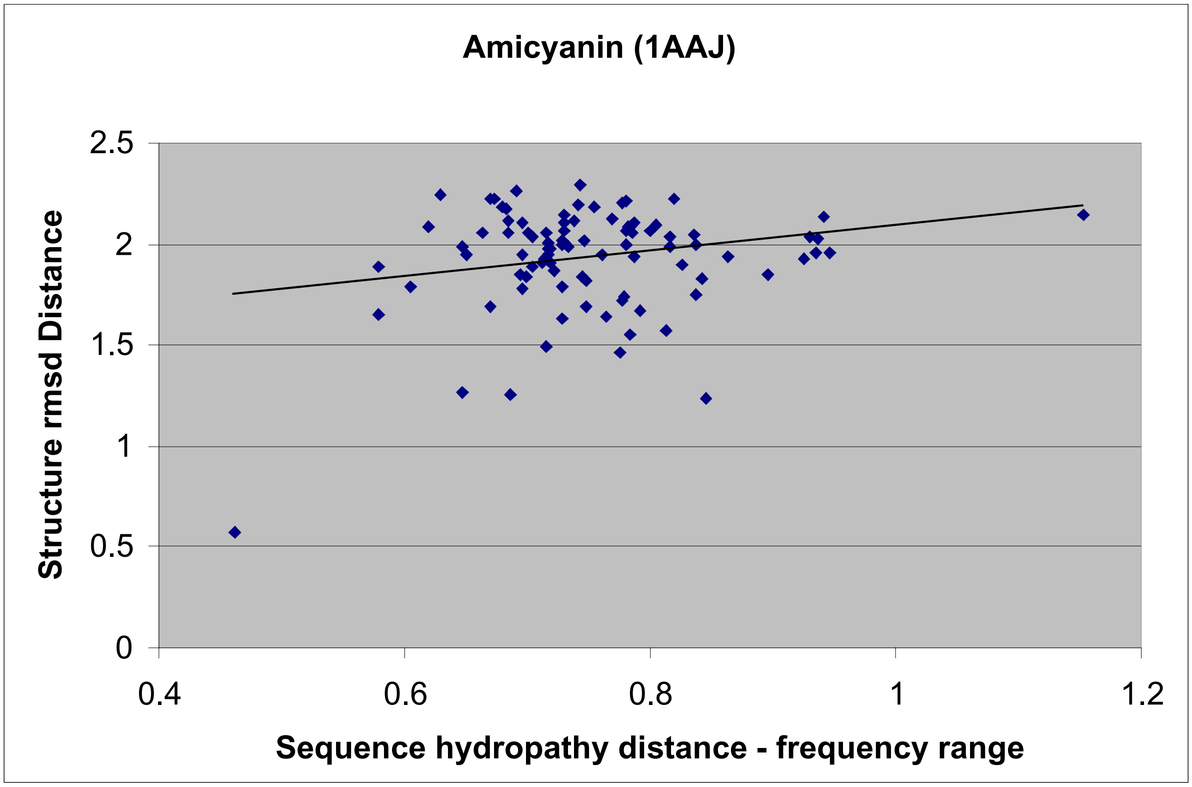
The existence of a large amount of noise in the data, as evident in the scattered nature of the plots, can be attributed to many factors. Both of the distance functions are very crude (one is derived from a simple, ungapped alignment while the other is an arbitrary linear combination of the rmsd and the inverse of the number of residues aligned). Better distance functions may give rise to more significant results. Furthermore, the fact that the LOCK structural alignments allow gaps while my hydropathy plot alignment program disallow gaps partially accounts for the discrepancy between the two sets of data.

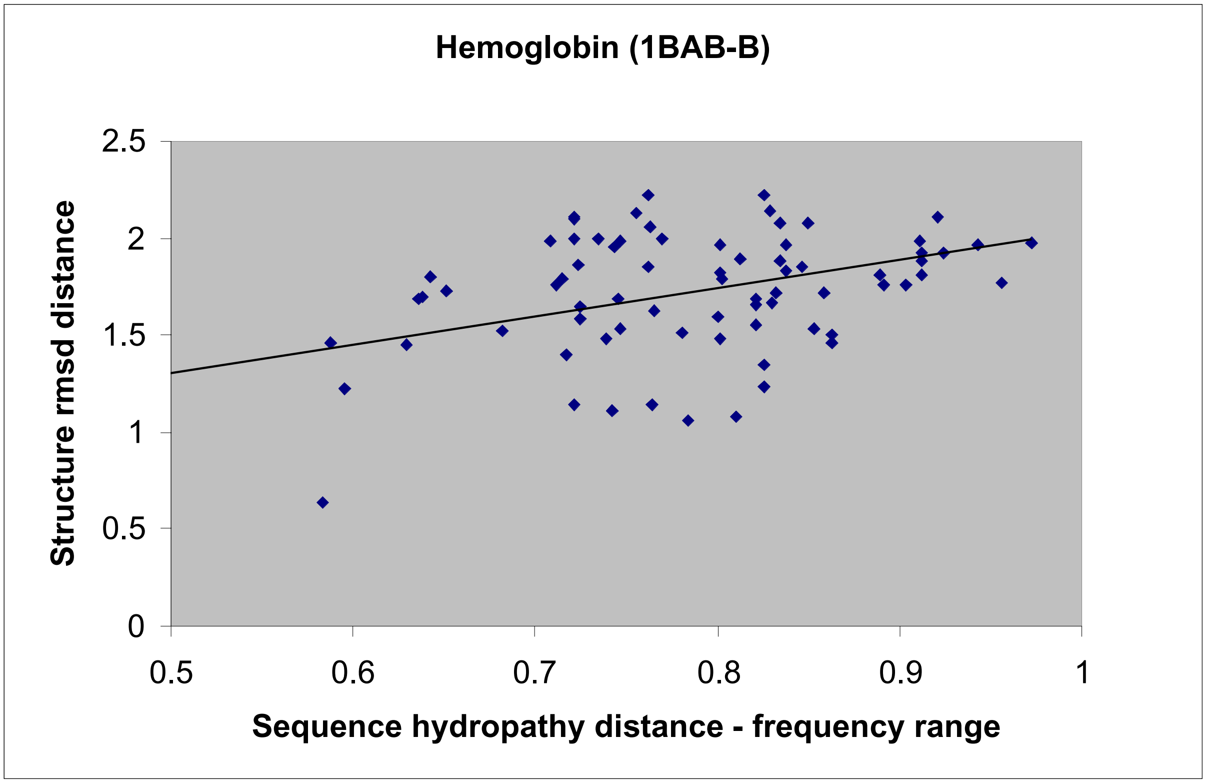
Another contributor to noise in the data is the fact that residue hydropathy is only one of the many determinants of sequence structure. As long as we can not encode all of the physio-chemical properties of each of the amino acids into their numerical representation, noise will always be a problem. However, the experiment can be improved if a more sophisticated mapping (such as that invented by Kidera et al.) was used instead of a simple hydropathy plot. Using a more sophisticated mapping, we would be able to observe a more complete correlation between sequence periodicity and structure. However, the mapping currently used allows us to observe the isolated effect of the solvent on protein folding, which is also very valuable.

Finally, given more time, I would like to test on more proteins. Even if we observe a strong positive correlation in the plots, they would not be significant in arguing the existence of a trend applicable to all soluble proteins, because only two query sequences were used in this experiment. To prove any general trends, an experiment of much broader scope, testing on hundreds of proteins, preferrably taken from all structural classes, would be necessary.

Appendix: Hydropathy Distance Versus Structure Distance Plots

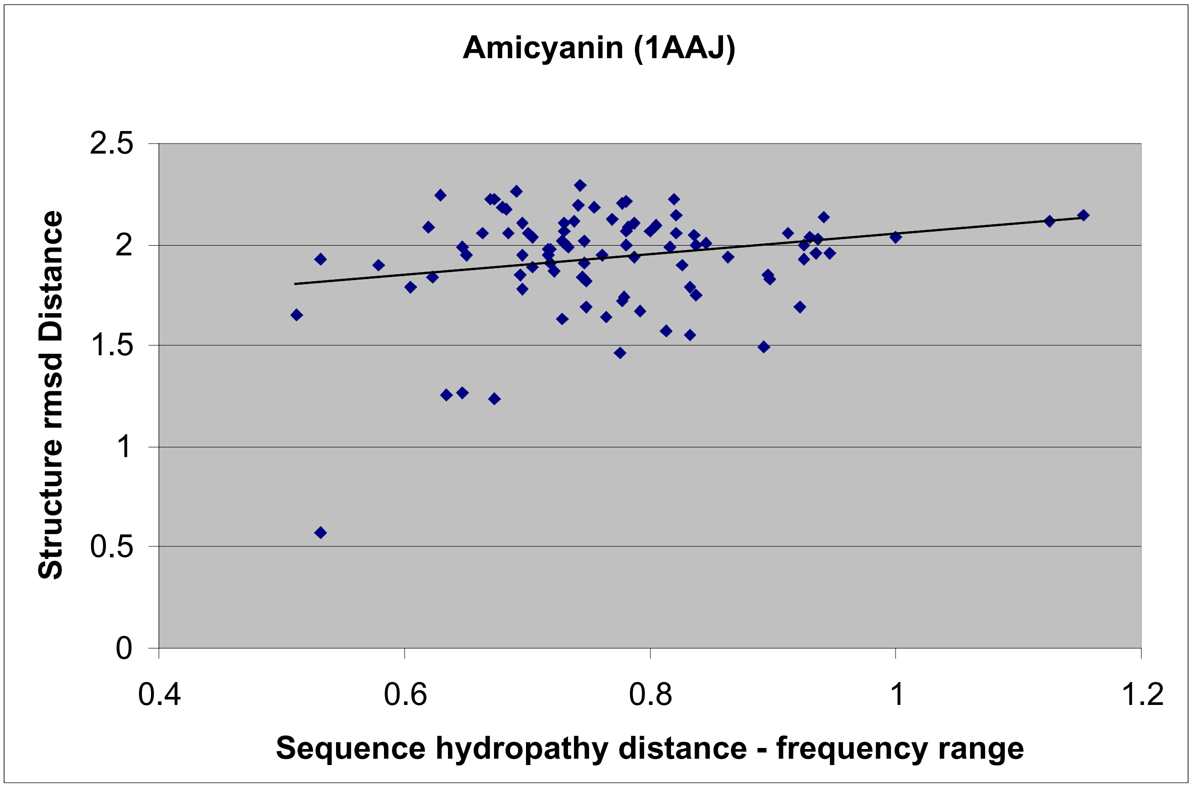
Sequence Hydropathy Calculated Using the M_1 Mapping

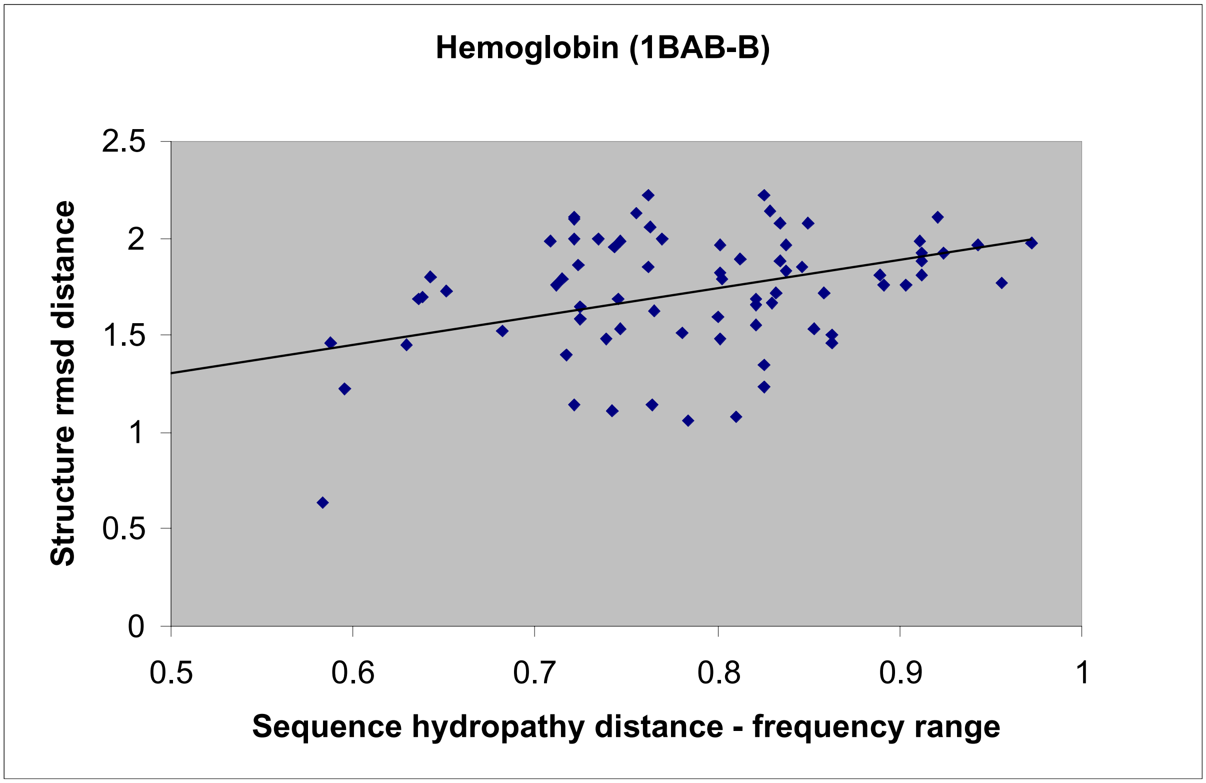




Appendix: Hydropathy Distance Versus Structure Distance Plots

Sequence Hydropathy Calculated Using the M_1 Mapping





References

1. Anfinsen, C. B. (1974). *Science*, 181, 223.
2. Seckler, R. and Jaenicke, R. (1992). *FASEB J*., 6, 2545.
3. Kabsch, W. & Sander, C. (1983). How good are predictions of protein secondary structure? *FEBS Letters*, 155, 179-182.
4. Hiroshi, W. & Blundell, T. (1994). Use of Amino Acid Environment-dependent Substitution Tables and Conformational Propensities in Structure Prediction from Aligned Sequences of Homologous Proteins. *Journal of Molecular Biology*, 238, 682-708.
5. Xiong, H., Buckwalter, B., Shieh, H., & Hecht, M. (1995). Periodicity of polar and nonpolar amino acids is the major determinant of secondary structure in self-assembling oligomeric peptides. *Proc. Natl. Acad. Sci. USA*, 92, 6349-6353.
6. Cornette, J. L., Cease, K. B., Margalit, H., Spouge, J. L., Berzofsky, J. A. & Delisi, C. (1987). Hydrophobicity scales and computational techniques for detecting amphipathic structures in proteins. *Journal of Molecular Biology*, 195, 659-685.
7. Eisenberg, D., Weiss, R. M. & Terwilliger, T. C. (1984). The hydrophobic moment detects periodicity in protein hydrophobicity. *Proc. Natl. Acad. Sci. USA*, 81, 140-144.
8. Schulz, G. E., & Schirmer, R. H. (1979). In *Principles of Protein Structure.* (ed. C. Cantor). p. 37. Springer Verlag New-York Inc.
9. Janin, J. (1979). Surface and Inside Volumes in Globular Proteins. **Nature**. 277, 491-492.
10. Rose, G., Geselowitz, A., Lesser, G., Lee, R., & Zehfus, M., (1985). Hydrophobicity of Amino Acid Residues in Globular Proteins, *Science* 229, 834-838
11. Wolfenden, R., Anderson, L., Cullis, P. & Southgate, C.. (1981). Affinities of Amino Acid Side Chains for Solvent Water. *Biochemistr,*. 20, 849-855.
12. Kyte, J. & Doolite, R. (1982). A Simple Method for Displaying the Hydropathic Character of a Protein, *Journal of Molecular Biology*, 157, 105-132.
13. White, S. H. & Jacobs, R. E. (1990). Statistical distributin of hydrophobic residues along the length of protein chains. *Biophysics*, 57, 911-921.
14. Rackovsky, S. (1998). “Hidden” sequence periodicities and protein architecture. *Proc. Natl. Acad. Sci. USA*, 95, 8580-8594.
15. Kidera, A., Konishi, Y., Oka, M., Ooi, T. & Scheraga, H. A. (1985). *Journal of Protein Chemistry*. 4, 23-54.
16. Schiffer, M. & Edmundson, A. B. (1967). Use of helical wheels to represent the structures of protein and to identify segments with helical potential. *Biophysics. J.* 7, 121-135.
17. Singh, A. P. & Brutlag, D. L. LOCK Hierarchichal Protein Structure Superposition. www.gene.stanford.edu/lock/.