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 $R = \{r_i\}$, $i = 1 \ldots N$, the Fourier transform of R would be:

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

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

$$F(R)_i = | F(R)_i |$$

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.