Ab Initio Prediction of Helical Segments in Polypeptides

J. L. Klepeis and C.A. Floudas*
Department of Chemical Engineering
Princeton University
Princeton, N.J. 08544-5263

*Author to whom all correspondence should be addressed; Tel.: (609) 258-4595; Fax: (609) 258 -0211; email: floudas@titan.princeton.edu

Abstract

An ab initio method has been developed to predict helix formation for polypeptides. The approach relies on the systematic analysis of overlapping oligopeptides to determine the helical propensity for individual residues. Detailed atomistic level modeling, including entropic contributions and solvation/ionization energies calculated through the solution of the nonlinear Poisson-Boltzmann equation, is utilized. The calculation of probabilities for helix formation is based on the generation of ensembles of low energy conformers. The approach, which is easily amenable to parallelization, is shown to perform very well for several benchmark polypeptide systems, including bovine pancreatic trypsin inhibitor, the immunoglobulin binding domain of protein G, and chymotrypsin inhibitor 2.

Keywords : Protein folding; secondary structure prediction; free energy; alpha helix; global optimization

1 Introduction

Proteins are essential molecules that exhibit complex structural and functional relationships. Biological functionality is defined by the native three-dimensional structure of the protein, which in turn depends on the intricate balance of molecular interactions of the system. It is well known that many proteins fold spontaneously from random disordered states into compact (native) states of unique shape. However, the ability to explain the mechanisms that govern this transformation has not yet been realized. The protein folding problem is to understand this folding process and to predict the three dimensional structure of proteins from their one dimensional amino acid sequence.

An important question regarding the prediction of the native folded state of a protein is how the formation of secondary and tertiary structure proceeds. Two common viewpoints provide competing explanations to this question. The classical opinion regards folding as hierarchic, implying that the process is initiated by fast formation of secondary structural elements, followed by the slower arrangement of the tertiary fold. The opposing perspective is based on the idea of a hydrophobic collapse, which suggests that tertiary and secondary features form concurrently.

Inherent to the hierarchical view of protein folding is the dominant role of local forces in determining the formation of secondary structure. These local forces denote those interactions between neighboring residues, rather than nonlocal forces that may arise during tertiary structure formation. In other words, local sequence information should be sufficient to predict native secondary structure if folding is hierarchic. In considering the local prediction of secondary structure elements, such as α-helices, β-strands and turns, most methods rely on statistical treatments[1]. More recent work has led to the proposal of a physical theory for secondary structure formation based on local interactions and sterics[2,3,4]. The basis for this theory hinges on the role of intrinsic propensities for backbone conformations and backbone hydrogen bonding.

The alternative perspective stresses the importance of the hydrophobic collapse rather than local propensities in determining a protein's fold. In this view, hydrophobic forces drive the collapse through the desolvation of side chains. It is believed that these non-local side chain interactions influence the formation of tertiary as well as secondary structural elements[5]. In addition, these ideas suggest that simple side chain models of protein folding may be sufficient to predict folding behavior.

For both cases experimental evidence has been produced to support the underlying claims. For example, kinetic studies have shown that elements of secondary structure common to the native fold are able to form before substantial tertiary structure arrangement. The boundaries of helical structure can also be identified through local sequence information, implying that local interactions dominate helix formation. Finally, fragments of longer protein sequences can form native-like folds in absence of long range interactions[6]. On the other hand, support for non hierarchical folding through a hydrophobic collapse includes experiments showing that protein folds are less affected by mutations on their surfaces than in their hydrophobic cores[7]. In addition, hydrophobic collapse, like secondary structure formation, occurs rapidly in certain cases[8]. Other results, such as the formation of β-sheet folds through α-helical intermediates[9], imply that secondary units are not preassembled and can be driven by tertiary structure formation.

It is interesting to note that simulations of a hydrophobic collapse through side chain models fail to predict the formation of α-helices[10]. This indicates that simplified models for protein folding may not be sufficient because they lack a full structural and energetic description of secondary structure formation. Other methods, such as those based on a statistical mechanical treatment for helix determination, have been effective, but lack a true physical basis[1].

In this work, the principles of hierarchical folding are used to develop a method for the prediction of α helices in protein systems. The support for this procedure for α-helix determination is based on observations that native like segments of helical secondary structure form rapidly. The ability for helices to overcome Levinthal's paradox suggests that α helix formation can occur during the earliest stages of protein folding. Such a mechanism for the helix-coil transition is based on local interactions which induce nucleation and propagation of the helix[11].

2 Secondary Structure Prediction

Secondary structure prediction is often an important precursor in tackling the overall protein folding problem, and many methods have been developed in an attempt to accurately predict the location of α helices and β strands. The most successful methods rely on homology modeling or multiple sequence alignments to predict secondary structure using only the amino acid sequence. If the databases of experimental structures contain significantly similar (homologous) sequences to the predicted sequence, then local conformation patterns, such as α helices and β strands, can be predicted with accuracy that in certain cases can exceed 70 percent. However, most protein sequences do not possess known structural homologues, which causes a significant decrease in prediction accuracy. For these cases the natural extension of the comparative modeling approach to fold recognition and threading techniques has shown some success.

For target sequences possessing known folds, the technique of comparative modeling begins with the process of sequence alignment; in other words, the search for homologous proteins. This procedure is practical when sequence identities are greater than 30 percent[12,13]. Since the goal of sequence alignment is to identify and accurately align segments of related sequences, the use of multiple sequence alignment has been an important development that has led to the ability to better identify sequence variability, insertions and deletions[14]. The most successful sequence alignment techniques use profiles derived from databases of sequence families[15,16,17]. More recently, advanced sequence alignment methods have been based on hidden Markov models [18,19] and genetic algorithms[20].

The success of sequence alignment, as measured by the sequence identity score, directly determines the complexity of the homology modeling process. For sequence identities greater than 70 - 90 percent, the backbone template of the homologous protein provides a very accurate model for the target sequence[21,22]. The only remaining step is to correctly place the side chains of the target sequence onto the backbone of the template sequence. The task becomes more complex as sequence identities decrease to the vicinity of 30 percent. Aligned sequences in this range generally adopt the same fold, however the sequence is dominated by the modeling of loops, which introduces additional challenges[23,24].

For target sequences possessing known folds but low sequence identities (less than 30 percent), the applicability of comparative modeling becomes uncertain. In fact, before the sequence can be properly aligned, the question of accurately detecting a remote homologous sequence must be addressed. These complications have led to the development of threading methods, an NP complete class of problems, in which the target sequence is threaded onto the backbone of the template sequence while evaluating the sequence fitness. Typically, these fitness functions represent environment based[25,26,27,28], or knowledge based potentials derived from the PDB[29,30]. Other alternative threading schemes involving one dimensional secondary structure predictions have also been proposed[31,32,25]. Although threading methods are much more reliable than traditional alignment techniques, accuracy levels for the correct detection of remote homologues is still below 40 percent. These difficulties are magnified when trying to identify correct alignments and build two and three dimensional models[33].

When analyzing a target sequence possessing an unknown fold, as is the case for most proteins, homology modeling becomes even more difficult. Since secondary structures can usually be predicted more reliably than other features of protein structure, the major efforts have focused on these one dimensional predictions. Initial attempts in the area of secondary structure prediction were based on examining stereochemical properties[34] and statistics[35,36,37]. Many studies have also focused on the development of intrinsic sets of helix propensities to give better α-helix predictions [38,39,40,41]. More recently, the benefits of multiple sequence alignments and increased database information have been instrumental in improving prediction accuracies[42]. Many methods rely on evolutionary information through an analysis of the development of protein families from both sequence and structural databases[43,44,45,46,47]. Enhancements in secondary structure prediction accuracy using evolutionary concepts have been substantial. For example, an easily implemented and standard statistical method, GOR[36], provides 60 percent accuracy for three state (α, β, coil) secondary structure prediction, with only 10 percent of these residues exhibiting reliability scores comparable to homology modeling for known folds. The PHD method[43], which uses a feed forward neural network trained by back propagation of evolutionary information, provides a sustained prediction accuracy over 70 percent with 45 percent of these residues having acceptable reliability scores. More recent neural network methods such as PSIPRED[48,49], have achieved sustained accuracies over 75 percent.

In addition to evolutionary information, other secondary structure prediction methods have exploited database information based on physical property information such as solvent accessibility. For example, reliable predictions of solvent accessibility for conserved and functional regions of the target sequence can be used to identify secondary structure by comparing accessibility patterns derived from database proteins [50]. Methods which attempt to refine the procedure for accessibility based prediction have been developed recently[51,52]. However, the extension of comparative modeling and fold recognition techniques to two and three dimensions has generally resulted in low accuracy predictions for sequences with unknown folds. Improvements will require the use of enhanced mean force potentials[53,54], or the development of more accurate ab initio techniques.

3 Outline of Prediction Approach

The proposed approach for the ab initio prediction of helical segments in polypeptides is based on the key ideas of (i) partitioning the sequence of aminoacids into oligopeptides (e.g., pentapeptides, heptapeptides) such that consecutive oligopeptides have an overlap, for instance, four aminoacids for pentapeptides; (ii) atomistic level modeling of all appropriate interactions for each oligopeptide using the ECEPP/3 force field; (iii) generation of an ensemble of low energy conformations for each oligopeptide; (iv) incorporation of the entropic contributions and free energy calculations for each oligopeptide; (v) calculations of the contributions to free energy due to the formation of cavity for selected oligopeptides; (vi) calculations of the solvation contribution to free energy using the nonlinear Poisson-Boltzmann equation for selected oligopeptides; (vii) calculations of the ionization contribution to free energy using the nonlinear Poisson Boltzmann equation for selected oligopeptides; (viii) calculation of equilibrium occupational probabilities for the helical clusters based on the free energies of the oligopeptides; and (ix) classification of residues as helical according to average propensities for each residue as calculated by the equilibrium occupational probabilities for the helical clusters. A flowchart outlining the main steps of the approach is given in Figure 1.

4 Partitioning into Oligopeptides

The concept of partitioning the aminoacid sequence into overlapping oligopeptides is based on the idea that the formation of helices relies on local interactions and the positioning of each segment within the total protein. For instance, each pair of overlapping pentapeptides has four common aminoacids, and for a single chain polypeptide with N residues this translates into an analysis of a total of $N - 4$ pentapeptides. A schematic of these overlapping subsequences for the first 12 residues of BPTI is given in Figure 2.

Note that the first aminoacid (R) participates only in one pentapeptide (denoted as 1), the second aminoacid (P) participates in two pentapeptides (denoted as 1 and 2), the third aminoacid (D) participates in three pentapeptides (denoted as 1,2,3), the fourth aminoacid (F) participates in four pentapeptides (denoted as 1,2,3,4), while the aminoacids 5-8 (C,L,K,P) each participate in five pentapeptides.

By considering such overlapping pentapeptides and performing free energy calculations based on full atomistic models for each system (see Klepeis and Floudas, 1999 [55]), the effect of the local interactions of the neighboring aminoacids is considered explicitly. As a result, situations in which the same segment of identical aminoacid sequence can adopt different conformations in different proteins, as reported by Minor and Kim[56], can be identified. This is because the local interactions extend beyond the boundaries of the helical segment, and therefore are sufficient to account for such conformational preferences, as suggested by[4]. It should also be noted that a similar partitioning can also result in overlapping heptapeptides or nonapeptides. It is also worth noting that the idea of partitioning the polypeptide into overlapping nonapeptides was first pointed out by Anfinsen and Scheraga[57] who suggested the minimization with respect to the dihedral angles of the central residue and the consideration of a five state model.

The partitioning of the aminoacid sequence into oligopeptides offers the distinct advantages that (i) the novel free energy calculation method that we have recently developed and which is based on deterministic global optimization[55] can be directly applied to a linear sequence of $N - 4$ pentapeptides or $N - 6$ heptapeptides or $N - 8$ nonapeptides, and (ii) all oligopeptide free energy calculations can be performed in parallel, where N is the number of aminoacids in the single chain polypeptide under study.

5 Atomistic Modeling

The prediction of α-helices is based on a method that includes detailed atomistic level modeling of the protein system. This modeling is based on the ECEPP/3 semi-empirical force field model. For this force field, it is assumed that the covalent bond lengths and bond angles are fixed at their equilibrium values, so that the conformation is only a function of the independent torsional angles of the system. The total force field energy, $E_{\text{forcefield}}$, is calculated as the sum of electrostatic, nonbonded, hydrogen bonded, and torsional contributions. The main energy contributions (electrostatic, nonbonded, hydrogen bonded) are computed as the sum of terms for each atom pair (i,j) whose interatomic distance is a function of at least one dihedral angle. The general potential energy terms of ECEPP/3 are shown in Figure 3, while the development of the appropriate parameters is discussed and reported elsewhere[58].

6 Ensembles of Low Energy Conformers

Locating the global minimum potential energy conformation is not sufficient because Anfinsen's thermodynamic hypothesis requires the minimization of the conformational free energy. Specifically, potential energy minimization neglects the entropic contributions to the stability of the molecule. An approximation to these entropic contributions can be developed by using information about low energy conformations. That is, once a sufficient ensemble of low energy minima has been identified, a statistical analysis can be used to estimate the relative entropic contributions, and thus the relative free energy, for each conformation in the ensemble. A variety of methods have been used to find such stationary points on potential energy surfaces. For example, periodic quenching during a Monte Carlo or molecular dynamics trajectory can be used to identify local minima[59]. In this work two algorithms are advocated for generating low energy ensembles for pentapeptide sequences. The first is based on modifications of a deterministic branch and bound algorithm, αBB. The second, conformation space annealing (CSA), which does not provide the deterministic guarantees of the αBB, is based on the combination of genetic algorithms and simulated annealing[60].

Our previous work has shown that the generation of ensembles of low energy conformers is possible through algorithmic modifications of the general αBB procedure [55]. The original implementation of the deterministic αBB global optimization algorithm requires the minimization of a convex lower bounding function in each domain. The unique solution for each lower bounding minimum can then used as a starting point for the minimization (or function evaluation) of the original energy function in the current domain. In the case of local minimization, each partitioned region provides a single minimum energy conformation as the algorithm proceeds. Using this information, along with the global minimum energy conformation, a list of low energy conformers can be constructed.

However, this approach does not take advantage of all the information provided by the lower bounding functions. Rigorously, these functions posses a single minimum in each subdomain. Since the choice of α (convexity parameter) affects the convexity of the lower bounding functions, the α values can be modified to ensure a certain nonconvexity in these functions. In this case, the lower bounding functions possess multiple minima, and these functions can be minimized several times in each domain. In addition, since the lower bounding functions smooth the original energy hypersurface, the location of these multiple minima provide information on the location of low energy minima for the upper bounding function. Therefore, by using the location of the minima of the lower bounding function as starting points for local minimization of the upper bounding function, an improved set of low energy conformations can be identified. As before, these conformations are also localized in those domains with low energy as the subdomains decrease in size. This energy directed approach (EDA) is represented schematically in Figure 4.

The conformational space annealing method (CSA)[60] relies on stochastic measures to converge to a cluster that should include the global minimum energy conformation. Through the use of genetic algorithm updates, an ensemble of low energy minima is also produced. The first step involves the generation of a set of bank conformations, which should initially be distributed randomly throughout the conformation space. Each conformation in the bank is regarded as a representative of a group of local minima within a certain distance in conformational space. The distance measure between conformations i and j is the root mean square deviation with respect to the dihedral angles :

$$D_{ij} = \sqrt{\frac{1}{N_\theta}\sum_{i=1}^{N_\theta}(\theta_i - \theta_j)^2} \tag{1}$$

As the algorithm proceeds the parameter D_{cut}, which defines the size of a cluster in conformation space, is slowly annealed from the original bank distribution value to an arbitrarily small value.

The group representatives in the bank are updated by generating additional conformations. The generation of these conformations is based on the concepts of a genetic algorithm, so that fragments of conformation i are replaced by randomly chosen conformations from the rest of the bank. The updating rules include replacing individual dihedral angles, randomly chosen groups of correlated (small number) dihedral angles, and connected groups (large number) of dihedral angles. The newly generated conformations are minimized and compared to the set of bank conformations. If the bank conformation closest in conformational space to the new conformation exhibits a value of $D_{ij} < D_{\text{cut}}$, the bank conformation is replaced by the new conformation if the new conformation provides a lower energy value. However, if $D_{ij} > D_{\text{cut}}$ for all conformations in the bank, the new conformation defines a new cluster which will enter the bank if it provides an energy lower than the highest energy representative in the bank. In this way the number of bank conformations remains constant. The termination criteria is heuristic and is related to the total number of minimizations.

7 Free Energy and Entropic Calculations

The analysis of these pentapeptides is based on a procedure to identify the free energy probability of having the *three central residues* of the pentapeptide within the helical region of the $\phi - \psi$ space. This requires the incorporation of entropic effects to calculate free energy probabilities of individual metastable states of the system [55]. In particular, a strict interpretation of Anfinsen's thermodynamic hypothesis requires the global minimization of the conformational free energy to predict the native structure of a protein. In practice, however, most protein models include only potential and solvation effects. One reason for this neglect of including entropic effects is that a rigorous free energy model requires infinite sampling to associate accurate statistical weights with each microstate.

Other approximate calculations exist for estimating these statistical weights (and thus entropic effects). The most simplistic model would rely on only the Boltzmann weight associated with each microstate. A more sophisticated approximation, known as the harmonic approximation, utilizes second derivative information to characterize the basin of attraction. More complex schemes try to mimic the anharmonic trajectory along the energy surface. These quasi-harmonic approximations generally require the use of MC/MD simulations.

In this work, entropic effects are included via the harmonic approximation[61,62,63]. The development of this model can be understood physically by first considering the partition function for the system :

$$Z = \exp^{-\frac{(E-TS)}{k_B T}} = \exp^{-\frac{E}{k_B T}} \exp^{\frac{S}{k_B}} \tag{2}$$

In Equation (2) the partition function is the product of the Boltzmann ($\exp[-E/k_B T]$) factor and the number of states available to the system ($\exp[S/k_B]$). At a given stationary point, the harmonic approximation is equivalent to :

$$E(\theta) = \mathbf{E}(\theta_\gamma) + \frac{\mathbf{1}}{\mathbf{2}}\left(\theta - \theta_\gamma\right) \mathbf{H}(\theta_\gamma) \left(\theta - \theta_\gamma\right) \tag{3}$$

Here γ identifies the stationary point, and the stationarity condition ($\nabla E(\theta_\gamma) = \mathbf{0}$) is used to eliminate the gradient term. In this way, each basin of attraction is characterized by properties of its corresponding minima, which include the local minimum energy value, $E(\theta_\gamma)$, and the convexity (Hessian) information around the local minimum, $\mathbf{H}(\theta_\gamma)$. An analogous representation of this system is N_θ independent harmonic oscillators, each with its own characteristic vibrational frequency. The minimum can then be characterized by the occupation of each normal mode.

To develop an expression for the entropic effect, Equation (3) can be substituted into Equation (2). By summing over all energy states, the partition function becomes :

$$Z_\gamma^{\text{har}} = \exp^{-\frac{E(\theta_\gamma)}{k_B T}} f(T) \prod_i^{N_\theta} \frac{1}{\lambda_i} \tag{4}$$

In Equation (4), $f(T)$ is a function dependent only on temperature, while λ_i represent the eigenvalues of $\mathbf{H}(\theta_\gamma)$. Comparison of Equations (4) and (2) implies that :

$$\exp^{\frac{S}{k_B}} \propto \prod_i^{N_\theta} \frac{1}{\lambda_i} \tag{5}$$

Equation (5) can be rewritten in terms of the harmonic entropic contribution, S_γ^{har} :

$$S_\gamma^{\text{har}} \quad \propto \quad -k_B \ln\left[\text{Det}\left(H(\theta_\gamma)\right)\right] \tag{6}$$

A more rigorous derivation of the harmonic approximation leads to the following expression for the harmonic entropy :

$$S_\gamma^{\text{har}} \quad = \quad -\frac{k_B}{2} \ln\left[\text{Det}\left(H(\theta_\gamma)\right)\right] \tag{7}$$

This can be used to calculate relative free energies via the following equation :

$$F_\gamma^{\text{har}} \quad = \quad E(\theta_\gamma) + \frac{\mathbf{k_B T}}{\mathbf{2}} \ln\left[\text{Det}\left(\mathbf{H}(\theta_\gamma)\right)\right] \tag{8}$$

Finally, each microstate can be assigned a statistical weight (p_γ^{har}) by considering the ratio of the partition function for that microstate (Z_γ^{har}) to the total partition function :

$$p_\gamma^{\text{har}} \quad = \quad \frac{\left[\frac{1}{\left[\text{Det}(H(\theta_\gamma))\right]}\right]^{1/2} \exp\left(-\frac{E(\theta_\gamma)}{k_B T}\right)}{\sum_{i=1}^{N_\gamma} \left[\frac{1}{\left[\text{Det}(H(\theta_{\text{i}}))\right]}\right]^{1/2} \exp\left(-\frac{E(\theta_{\text{i}})}{k_B T}\right)} \tag{9}$$

To develop a meaningful comparison of relative free energies, the total partition function (denominator of Equation (9)) must include an adequate ensemble of low-energy local minima, as well as the global minimum energy conformation. Therefore, efficient methods for identifying low energy ensembles, as outlined in the previous section, must be employed. It should also be noted that the harmonic approximation does not require the explicit inclusion of a contribution based on the density of states because each local minimizer is accounted for only once (in contrast to counting methods).

Relative free energies can also be calculated for clusters of low energy conformers. This analysis is useful because it is difficult to capture the true accessibility of individual structures based on a point-wise approximation of entropic effects. That is, the harmonic free energy approximation does not provide a continuous free energy landscape. Typically, structures are clustered by calculating and comparing root mean squared deviations. In the case of determining α helical structure, a conformer is said to belong to the α-helical cluster if the torsional angles of three central residues belong to the α-helical region of the $\phi - \psi$ space (denoted as AAA). The relative free energy of the α-helical cluster can be calculated by the following equation :

$$F_{\text{AAA}} \; = \; -k_B T \ln \sum_{i \in \text{AAA}} p_i^{\text{har}} \tag{10}$$

In Equation (10) the individual p_i^{har}, which refers to the statistical weight based on the harmonic approximation, are summed for the set of conformations belonging to the AAA cluster. These individual probabilities are calculated by normalizing each probability with respect to the overall probability at a given temperature :

$$p_i^{\text{har}} = \frac{\exp[-\beta(F_o^{\text{har}} - F_i^{\text{har}})]}{\sum_j \exp[-\beta(F_o^{\text{har}} - F_j^{\text{har}})]} \tag{11}$$

A reference free energy, F_o^{har}, is used to normalize the probabilities at each temperature point. All free energies, F_o^{har}, F_i^{har} and F_j^{har}, refer to the harmonic approximation of the free energy as calculated using Equation (8). The denominator, which represents the total probability at a given temperature, is calculated by summing over the set of all conformers.

8 Electrostatic Contributions to Free Energy

Initially, the overlapping pentapeptides are modeled as neutral peptides surrounded by a vacuum environment using the ECEPP/3 force field. The incorporation of solvation effects requires additional energetic modeling, as well as considering the role of ionizable side chains. These contributions can be included through explicit or continuum based hydration models.

Explicit methods include solvation effects by actually surrounding the peptide with solvent molecules. Energetic evaluations require the calculation of both solvent–peptide and solvent–solvent interactions. Although these methods are conceptually simple, explicit inclusion of solvent molecules greatly increases the computational time needed to simulate the peptide system. Therefore, most simulations of this type are limited to local conformational searches.

Continuum models use a simplified representation of the solvent environment by neglecting the molecular nature of the water molecules. Many models estimate free energies of solvation as a function of geometric quantities, such as surface areas and volumes. More rigorous calculations of solvation free energies include electrostatic continuum models, which rely on numerical solutions to the Poisson–Boltzmann equation, and from which dielectric and ionic strength effects are obtained[64].

In this work, solvation and ionization free energies are calculated through the solution of the nonlinear Poisson Boltzmann equation, for which both finite difference and multigrid boundary element solution methods are available[65,66]. In particular, the finite difference solution of the Poisson Boltzmann equation as implemented in the DELPHI package is adopted[67,68]. In addition, the approach includes a procedure for effectively evaluating both the solvation and ionization free equilibria of the peptide conformations[69,70]. The resulting total free energies can then be used to identify equilibrium occupational probabilities for the α-helical clusters.

The overall methodology encompasses the following steps:

1 Using the ECEPP/3 forcefield, an ensemble of low potential energy oligopeptide (e.g., pentapeptide) conformations, along with the global minimum potential energy conformation, are identified using deterministic global optimization based techniques.

2 Determine a set of unique conformers ($\mathcal{Z}$) by removing all duplicate and symmetric minima, as well as those that do not differ by more than 50 degrees for at least one dihedral angle (disregarding the first and last backbone dihedral angles and the last dihedral angle in each side chain).

3 For the set $\mathcal{Z}$ calculate the vibrational entropic component using the harmonic approximation.

4 Model cavity formation in an aqueous environment using a solvent accessible surface area correlation :

$$F_{\text{cavity}} = \gamma A_{sa} + b \tag{12}$$

This macroscopic free energy term is based on a fit to the experimental free energy of transfer of alkane molecules into water. The values for the γ and b parameters are taken to be 0.005 kcal/mol $\mathring{A}$ and 0.860 kcal/mol, respectively.

5 Rank the set ($\mathcal{Z}$) according to the energies given by ($F_{\text{vac}}^{\text{har}} + F_{\text{cavity}}$).

6 For a subset of conformations belonging to ($\mathcal{Z}$) calculate the total energy according to :

$$F_{\text{total}} = F_{\text{vac}}^{\text{har}} + F_{\text{cavity}} + F_{\text{solv}} + F_{\text{ionize}} \tag{13}$$

Here F_{solv} represents the difference in the polarization energies when moving from a vacuum to an aqueous environments, and $F_{\text{ionization}}$ represents the ionization energy (see below). The thermodynamic process that captures this transition is given in Figure 5.

7 F_{total} is subsequently used to calculate equilibrium occupational probabilities of the α-helical cluster.

8.1 Solvation Free Energy

Calculating the polarization of the environment as an aqueous phase is based on the difference between electrostatic polarization free energies of the peptide in the vacuum and water environments. The change in going from a vacuum to aqueous environment is given by :

$$F_{\text{solv}} = F_{\text{polar}}(\epsilon = 80) - F_{\text{polar}}(\epsilon = 1) \tag{14}$$

This involves finding the induced surface charge (solving the Poisson-Boltzmann equation) for two systems in which the only difference is the dielectric constant (ϵ) of the surrounding medium; that is 80 and 1 for the aqueous and vacuum phases, respectively.

Finding F_{polar}, which corresponds to the reaction field energy, requires solving the Poisson-Boltzmann equation when the neutral protein (zero ionization) is in the aqueous and vacuum phases. The reaction field energy is determined by calculating the induced surface charge at the surface of the molecule (due to point charges) and then summing the potential at every charge :

$$F_{\text{polar}} = \frac{1}{2}\sum_i \sum_s \frac{q_i \sigma_s}{|\mathbf{r_i} - \mathbf{r_s}|} \tag{15}$$

Reaction field energies can be obtained as a special application of the solution of the Poisson-Boltzmann equation. In particular, the distribution of charges and dielectric boundaries is first used to solve the Poisson-Boltzmann equation through finite difference for all points of a three-dimensional grid. This provides a potential at each grid point. In order to calculate the surface charge density, the proximal grid point potentials are combined for a patch of the constructed Conolly surface. The reaction field energy is calculated by determining the effect of the charge density at each surface patch for each charge point.

8.2 Ionization Free Energy

For ionizable residues additional calculations must be made for the ionization of these groups in the aqueous phase at a given pH. The determination of this quantity depends on the calculation of the partition function for single or multiple titration sites :

$$F_{\text{ionize}}(\text{pH}) = kT \ln Z \tag{16}$$

The partition function includes contributions from all 2^N ionization states of the system, where N is the number of ionizable groups :

$$Z = \sum_{i=1}^{2^N} \exp\left[-\Delta G_i / kT\right] \tag{17}$$

The free energy of the ith state is given by ;

$$\Delta G_i = \sum_{j=1}^{N} \left(x_j 2.303kT \left(\text{pH} - \text{pK}_j\right) + \delta_j \sum_{1 \le k < j} \delta_k \Delta G_{jk} \right) \tag{18}$$

Here x_j is the charge on the group in the ith state, and δ parameters are binary indicators (i.e., 0 when the group is neutral and 1 when the group is charged). pK_j is the intrinsic $\mathrm{pK_a}$ for the jth group, and pH is the current pH value. ΔG_{jk} represent coupled (multiple site) terms.

Intrinsic $\mathrm{pK_a}$ values are obtained by looking at the difference of ionizing the protein in the protein environment and in an isolated aqueous phase :

$$\mathrm{pK}_j = \mathrm{pK}_j^o - \gamma_j \Delta\Delta G_j / 2.303kT \tag{19}$$

Here γ_j is equal to -1 or +1 for acidic or basic ionizable groups, respectively. The $\Delta\Delta G_j$ term is easily related to the pK shift ($\Delta\mathrm{pK}_j$) by the following :

$$\Delta\mathrm{pK}_j = \frac{\Delta\Delta G_j}{\gamma_j 2.303kT} \tag{20}$$

The thermodynamic cycle for $\Delta\Delta G_j$ involves the introduction of the ionizable group into the protein system and the difference in free energy when going between the neutral and protonated form of that group. This is represented by Figure 6.

Examination of the thermodynamic cycle provides the following decomposition for $\Delta\Delta G_j$:

$$\frac{\Delta\Delta G_j}{\gamma_j} = \left(\Delta G_j(\mathrm{PS}_i^+/\mathrm{S}_i^+) - \Delta G_j(\mathrm{PS}_i^o/\mathrm{S}_i^o)\right) \tag{21}$$

$\Delta G_j(\mathrm{PS}_i^+/\mathrm{S}_i^+)$ represents the change in free energy when moving the (ionized) ionizable group from an isolated aqueous environment into the protein environment. $\Delta G_j(\mathrm{PS}_i^o/\mathrm{S}_i^o)$ represents the same transition but for the neutral form of the ionizable group.

The individual ΔG_j terms can be further decomposed :

$$\Delta G_j = \Delta G_j^{\text{rxn field}} + \Delta G_j^{\text{perm dipole}} \tag{22}$$

The first term, $\Delta G_j^{\text{rxn field}}$ refers to the reaction field effects, that is, those effects that arise due to the dielectric continuum nature of the system. For example, $\Delta G_j^{\text{rxn field}}(\mathrm{PS}_i^+/\mathrm{S}_i^+)$ is the difference in reaction field energy for group j in state i when changing the dielectric continuum from the isolated aqueous state ($\epsilon = 80$ only) to that of the protein environment ($\epsilon = 2$ in some regions). More specifically, $\Delta G_j^{\text{rxn field}}$ captures the change in free energy due to the reduced exposure to water. Since we are concerned with the effect on the ionizable group j, the rest of the protein carries zero partial atomic charges.

In order to calculate the change in reaction field energy, $\Delta G_j^{\text{rxn field}}(\mathrm{PS}_i^+/\mathrm{S}_i^+)$, the Poisson-Boltzmann equation is solved for both systems shown in Figure 7 to get the reaction field potential map $\phi^{\text{rxn field}}(\mathrm{PS}_i^+)$ and $\phi^{\text{rxn field}}(\mathrm{S}_i^+)$. This data can be used to map the surface charge distribution on the boundary between the different dielectric environments, that is, $\sigma(\mathrm{PS}_i^+)$ and $\sigma(\mathrm{S}_i^+)$, respectively. By replacing the surface integral with the appropriate summation, the change in reaction field energy becomes :

$$\Delta G_j^{\text{rxn field}}(\text{PS}_i^+/\text{S}_i^+) = \frac{1}{2}\Bigg[\sum_{s(\text{PS}_i^+)}\sum_{j+}\frac{q_{j+}\sigma_s(\text{PS}_i^+)}{|r_{j+}-r_s|} - \sum_{s(\text{S}_i^+)}\sum_{j+}\frac{q_{j+}\sigma_s(\text{S}_i^+)}{|r_{j+}-r_s|}\Bigg]$$

In this equation the set $j+$ refers to the set of partial atomic charge points (with charges q_{j+}) of the protonated ionizable group. The set of surface points are denoted as $s(\text{PS}_i^+)$ and $s(\text{S}_i^+)$ for the isolated and protein environments, respectively. The quantity $|r_{j+} - r_s|$ is the magnitude of the distance between the points defined by sets $j+$ and s.

A similar set of equations can be derived for the neutral form of the ionizable group. The systems are shown schematically in Figure 8. $\Delta G_j^{\text{rxn field}}(\text{PS}_o^+/\text{S}_i^o)$ can be calculated from the following equation :

$$\Delta G_j^{\text{rxn field}}(\text{PS}_i^o/\text{S}_i^o) = \frac{1}{2}\Bigg[\sum_{s(\text{PS}_i^o)}\sum_{jo}\frac{q_{jo}\sigma_s(\text{PS}_i^o)}{|r_{jo}-r_s|} - \sum_{s(\text{S}_i^o)}\sum_{jo}\frac{q_{jo}\sigma_s(\text{S}_i^o)}{|r_{jo}-r_s|}\Bigg]$$

The final contribution to $\Delta\Delta G_j$ is based on the difference in potential forces on the ionizable group which arise from permanent dipoles of the entire system. Rather than consider these term separately, the overall dipole change can be written as :

$$\Delta\Delta G_j^{\text{perm dipole}} = \Delta G_j^{\text{perm dipole}}(\text{PS}_i^+/\text{S}_i^+) - \Delta G_j^{\text{perm dipole}}(\text{PS}_i^o/\text{S}_i^o) \quad (23)$$

In the isolated systems, (S_i^+ and S_i^o), permanent dipole effects are not present. That is, the ionizable group is only surrounded by a uniform dielectric continuum with $\epsilon = 80$ and no permanent dipoles or ions are present. Therefore, $\Delta\Delta G_j^{\text{perm dipole}}$ collapses to :

$$\Delta\Delta G_j^{\text{perm dipole}} = \Delta G_j^{\text{perm dipole}}(\text{PS}_i^+/\text{PS}_i^o) \quad (24)$$

The calculation of this quantity requires the solution of Poisson-Boltzmann equation for two systems. For the PS_i^+ system, the potential force ($\phi_{j+}^{\text{perm dipole}}(\text{PS}_i^+)$ due to the protein dipole is calculated at the atomic centers of the protonated ionizable group (set $j+$). For the PS_i^o system, these forces ($\phi_{jo}^{\text{perm dipole}}(\text{PS}_i^o)$ are determined at the atomic centers (set jo) of the neutral form of the ionizable group. A schematic of these systems is shown in Figure 9.

The final expression for $\Delta G_j^{\text{perm dipole}}(\mathrm{PS}_i^+/\mathrm{PS}_i^o)$ is based on the sum of the effective potential at the atomic charge centers :

$$\Delta G_j^{\text{perm dipole}}(\mathrm{PS}_i^+/\mathrm{PS}_i^o) = \sum_{j+} q_{j+}\phi_{j+}^{\text{perm dipole}}(\mathrm{PS}_i^+) - \sum_{jo} q_{jo}\phi_{jo}^{\text{perm dipole}}(\mathrm{PS}_i^o) \tag{25}$$

The final step in treating multiple titration sites is the calculation of ΔG_{jk} terms. This term represents an energetic adjustment due to the permanent dipole contributions between each pair of titratable groups. In order to isolate the contributions to only those between the ionizable groups, the remaining protein is treated as uncharged. The expression for ΔG_{jk} can decomposed as :

$$\Delta G_{jk} \;=\; \Delta G_{jk}(\mathrm{PS}_i^{j+,k+}) + \Delta G_{jk}(\mathrm{PS}_i^{jo,ko}) - \Delta G_{jk}(\mathrm{PS}_i^{j+,ko}) - \Delta G_{jk}(\mathrm{PS}_i^{jo,k+}) \tag{26}$$

In total, four separate systems must be considered. The first term represents the dipole effects between the charged forms of both groups j and k. The remaining quantities, which correspond to combinations of the neutral and charged forms of groups j and k, are necessary to correct the approximations made when calculating the energies of single titration groups.

Permanent dipole calculations require the solution of the Poisson-Boltzmann equation for a distribution of permanent point charges. The solution provides the induced potential at all grid points, which can be used to calculate the effects at a subset of grid points (point charges).

9 Probabilities of α helix formation

The goal and final step of the approach is to classify the individual residues in the overall sequence as helical or non-helical. In the case of considering overlapping pentapeptides, for each residue, excluding the first and last three residues, this classification is based on information obtained from the three pentapeptides for which the residue in question maintains one of the three central positions. As a result, the combined effects of seven residues are accounted for when determining the helical propensity of each individual residue. When considering heptapeptides and nonapeptides the sphere of influence exends to eleven and fifteen residues, respectively. For each residue, the average probability of being in an helical conformation is defined by the average of the AAA probability for the aforementioned three pentapeptides. The individual AAA probability (p_{AAA}) for each pentapeptide is equivalent to the summation term shown in Equation (10) :

$$p_{\mathrm{AAA}} \;=\; \sum_{i \in \mathrm{AAA}} p_i^{\mathrm{har}} \tag{27}$$

The individual probabilities are calculated according to Equation (11), which depends on the total free energy of the system. For the case of pentapeptides without any ionizable side groups or low helical probabilities (p_{AAA}), the free energy is based on the in vacuo calculations. However, the free energy includes detailed solvation and ionization energies for those pentapeptides possessing ionizable side groups and large initial helical probabilities. A residue is defined as helical if the combined helical probabilities (p_{AAA}) of the three pentapeptides exceed an average of about 90 percent.

10 Computational Studies

10.1 Bovine Pancreatic Trypsin Inhibitor, BPTI

The approach for α-helix prediction was applied to bovine pancreatic trypsin inhibitor (BPTI), a small globular protein found in many tissues throughout the body. BPTI inhibits several of the serine protease proteins such as trypsin, kallikrein, chymotrypsin, and plasmin, and is a member of the pancreatic trypsin inhibitor (kunitz) family, which is a family of serine protease inhibitors. These proteins usually have conserved cysteine residues that participate in the formation of disulfide bonds. In particular, BPTI possesses three disulfide bonds, which are denoted as Cys5-Cys55, Cys14-Cys38, and Cys30-Cys51. The structure of the 58 amino acid residues chain of BPTI has been resolved through several methods, including X-ray crystallography (4PTI)[71] and a combination of X-ray and neutron diffraction experiments (5PTI)[72]. Basic secondary structural features include a N-terminal 3_{10} helix, a C-terminal α helix and several antiparallel β strand configurations.

Extensive experimental studies of the structural features and folding of BPTI have been conducted[73,74,75,76,77]. Many of these studies have attempted to elucidate the folding pathway of BPTI through the formation of stable intermediates, which necessarily have one or more broken disulfide bridges. Theoretical investigations of the stability and folding of BPTI have also been performed[78,79,80]. These simulations typically require information on secondary structural content and native contacts to examine the formation of the native folded state. The novel aspect of this approach is the identification of secondary structural features, including α helix prediction, through ab initio modeling.

For BPTI, the partitioning of the overall 58 residue chain into overlapping pentapeptides results in 54 pentapeptides. The individual uncharged pentapeptides are indicated in Table I. In general, the end groups for each pentapeptide are simply neutral amino groups at N termini and hydroxyl groups at C termini. For the case of N terminal proline residues the amino group is replaced by an acetyl-amino group, while C terminal proline residues require an amide-methyl group. The structural classification of each pentapeptide is based on the conformational characteristics of the three central residues. Based on the crystallographic structure, Table I indicates which pentapeptides possess core residues with full α helical structure.

For each pentapeptide, a series of free energy calculations was performed to identify low energy conformational ensembles. Energy modeling included standard potential energy components based on the ECEPP/3 forcefield, as well as configurational entropic contributions according to the harmonic approximation. The description of each conformer requires the specification of a set of independent torsion angles, and uniqueness of individual conformers was assessed based on criteria involving these variables. The total number of torsion angles and unique conformers for each pentapeptide is presented in Table II.

The free energy of each unique conformer evaluated at 298 K was used to calculate individual occupational probabilities for these metastable states. Clustering of these states was based on the classification of the backbone torsion angles of the central residues. Specifically, the probabilities of conformers exhibiting identical Zimmermann codes for the core residues were summed and used to generate a rank ordered list of conformational propensity. The first stage of the approach involves the identification of strong α helical clusters for the uncharged pentapeptides. Specifically, if the probability of the α helical cluster (AAA) is greater than 90 for more than three consecutive sets of core residues, the marked pentapeptides are considered for further analysis. The second stage involves refinement of α helix probabilities based on detailed electrostatic and ionization energy calculations obtained through the solution of the Poisson Boltzmann equation. For the set of possible α helical pentapeptides containing ionizable residues, probabilities were recalculated for a subset of conformers using a combination of the free energy at 298 K and the polarization and ionization free energy at pH 7. Finally, α helical propensity for each residue was assigned according to the average AAA probability. The results are presented in Figure 10. The prediction of an α helix corresponds to average AAA probabilities greater than 90 for more than three consecutive residues. For BPTI, α helices are predicted between residues 2 and 5 and between residues 47 and 54. These results are in excellent agreement with the experimental structure.

10.2 Protein G, 1GB1

Protein G is a small globular protein produced by several Streptococcal species. The proteins are composed of two or three nearly identical domains of about 55 amino acids each. The system considered here is the immunoglobulin-binding domain from streptococcal protein G, a 56 amino acid polypeptide. The structure contains an efficiently packed hydrophobic core between a four-stranded β-sheet and a four-turn α-helix[81] with an overall secondary structure of $\beta\beta\alpha\beta\beta$. The formation of the β-sheet consists of two β hairpin turns, each connecting antiparallel strands. The first and last strands combine to form the final parallel β sheet to give the four-stranded configuration. Experimental structures have been determined using both crystallographic [82] and NMR-derived[81] data.

Analysis of the immunoglobulin binding domain of Protein G has also been the focus of theoretical studies on protein folding. In particular, the third and fourth β strands have been used to model the formation of β sheet structure through hairpin folding. Initial observations included the proposal of a simple statistical mechanical model in which the formation of hydrogen bonds, through a zipper mechanism, drives hairpin folding[83]. More recently, simulations have shown that an early step in hairpin folding is the formation of a hydrophobic cluster[84,85,86].

For Protein G, a total of 52 overlapping pentapeptides, as presented in Table III, were constructed. For all pentapeptides end groups corresponded to neutral amino groups at N termini and hydroxyl groups at C termini. The structural classification of each pentapeptide is based on the conformational characteristics of the three central (core) residues. Based on the experimentally derived structure, Table III indicates that pentapeptides 22 through 32 possess core residues with full α helical structure.

The total number of torsion angles and unique conformers for each pentapeptide is presented in Table IV. Strong α helical clusters were identified using the in vacuo free energy for the uncharged pentapeptides. Based on these results, the probabilities for charged pentapeptides 2 - 8, 14 - 35 and 44-52 were recalculated using free energies which included polarization and ionization energies. The refined probabilities were used to calculate α helical propensities for each residue according to the average AAA probability. The inclusion of rigorous solvation and ionization energies reduced the N-terminal helix to a short fragment of 4 residues, between residues 5 - 8, exhibiting average helix propensity above 90 percent. Furthermore, a depression in the helix propensity below 90 percent for both residues 49 and 50 effectively disrupts the formation of a potential C-terminal helix. One remaining extended helix survives between residues 23 and 31. To investigate the two remaining helices, additional free energy calculations were conducted for heptapeptides including residues 5 - 8 and 32 - 34 at core positions. The results indicate that the N-terminal helix does not form for the longer heptapeptides, and that the second helix extends to residue 34. These observations suggest that different oligopeptide systems may be useful for affirming the pentapeptide results. Experimentally, the immunoglobulin binding domain of Protein G exhibits one α helix between residues 22 and 35, which agrees well with the prediction of a helix between residues 23 and 34. The final results are presented in Figure 11.

10.3 Chymotrypsin Inhibitor, 3CI2

Like BPTI, chymotrypsin inhibitors are serine protease inhibitors. Chymotrypsin inhibitor 2 commonly refers to the potato I family of trypsin inhibitors, which has been resolved experimentally using both crystallographic and NMR methods. Neglecting the unstructured set of residues near the N-terminus, the truncated 63 residue chain has a morphology consisting of 6 β strands, a helix and a reactive loop. The order of these secondary structural elements follows : strand$_1$, strand$_2$, helix, strand$_3$, reactive loop, $strand_4$, $strand_5$, $strand_6$. The four largest strands combine to form a packed hydrophobic core around the helix, with $strand_1$ antiparallel to $strand_6$, $strand_6$ antiparallel to $strand_4$, and $strand_4$ parallel to $strand_3$.

The folding characteristics of chymotrypsin inhibitor 2 protein have been studied extensively through experimental methods. Important observations include its fast folding two-state type kinetics with the same folding and unfolding transition state [87,88]. Molecular dynamics simulations have failed to elucidate a single unfolding pathway, although common structural features have been identified. For example, recent simulations indicate that the nucleation of both the $\beta 3$ - $\beta 4$ hairpin and the four-turn helix rapidly form the native structure[89,90]. Like the immunoglobulin binding domain of protein G, the strongest consolidation of secondary structure is found in the α helix.

The decomposition of the truncated 63 residue chain of chymotrypsin inhibitor 2 results in a total of 59 overlapping pentapeptides, as presented in Table V. End groups corresponded to neutral amino groups at N termini and hydroxyl groups at C termini for all pentapeptides, excluding those with terminal Pro residues. Each pentapeptide is classified according to the conformational characteristics of the three central (core) residues. Based on the experimentally derived structure, Table V indicates that pentapeptides 10 through 18 possess core residues with full α helical structure.

Table VI presents the total number of torsion angles and unique conformers for each pentapeptide. Using the in vacuo free energy, the strongest α helical clusters were identified for uncharged pentapeptides 9 through 20. The C-terminal region, specifically between residues 45 and 55, also displayed relatively high AAA probabilities, although two small depressions below 75 percent disrupt the possibility for helix formation in this region. In addition, the region between residues 35 and 45, which corresponds to the reactive loop structure, does not produce helical conformers. Refined probabilities for pentapeptides 9 through 20 were used to calculate α helical propensities for each residue according to the average AAA probability. These results, as presented in Figure 12, support the prediction of a single helix between residues 12 and 21. These results agree extremely well with those found experimentally.

10.4 Comparison with Existing Methods

The results for the postulated superstructures were then compared to the PSIPRED method for secondary structure prediction[91]. PSIPRED utilizes two feed-forward neural networks to perform an analysis on output obtained from PSI-BLAST (Position Specific Iterated - BLAST)[92]. Cross validation of the method indicates that PSIPRED is capable of achieving an average Q3 score of nearly 77 percent, which is the highest result for any published secondary structure prediction methods. The predictions are based on a standard three state model to indicate the location of helix, strand and coil fragments for a given sequence.

Qualitatively, the results of the PSIPRED method and the ab initio approach agree quite well in the prediction of helical segments for the three proteins studied here. In particular, the extended helix in each of the systems is accurately predicted with a high confidence level (an average of 8 out of 9 on a 0 to 9 scale) using PSIPRED. Two disagreements are evident between the two predictions, and both represent inaccuracies in the PSIPRED results. The first is a lack of the prediction of the small N-terminal helix between residues 2 and 5 for BPTI. The second is the weak prediction (confidence level less than 1) of an additional helix between residues 17 and 20 in the immunoglobulin binding domain of protein G. Figure 11 exhibits a spike in the helix propensity for this region, although the conditions do not satisfy the criteria for assigning a helical segment.

10.5 Computational Complexity

The extension of our ab initio helix prediction approach to larger protein systems is facilitated through the use of distributed computing environments. The major expense of the overall approach involves multiple solutions of the nonlinear Poisson-Boltzmann equation for each conformation, which depends strongly on the number of ionizable groups. An estimate of the computational effort is made for a 128 processor (600 PIII) parallel machine running Linux.

For each oligopeptide, the set of in vacuo free energy calculations can be performed independently on single processors. To avoid significant idle time, the number of oligopeptides, which is on the order of the total number of residues in the sequence, should not exceed the number of available processors. As an example, a typical pentapeptide would require approximately 15 CPU hours on a single processor. On a 128 processor system, results for a full set of pentapeptides from a 68 residue sequence (or shorter) will complete in approximately 10 wallclock hours.

All additional calculations require the solution of the nonlinear Poisson-Boltzmann equation, which is carried out by finite difference routines implemented in the DELPHI package[67,68]. The calculation of F_{solv} requires two calls to DELPHI, and the number of calls is independent of the number of titratable groups in the system. For each ionizable group six additional DELPHI calls are required, four reaction field calculations (PS_i^+,PS_i^o,S_i^o, S_i^+) and two permanent dipole calculations (PS_i^+,PS_i^o). Two of the six calculations involve only single residue conformations, rather than the full protein system. When multiple titratable groups are present, four additional DELPHI calls must be made for each pair of ionizable groups. The computational effort is summarized in Table VII.

The set of DELPHI calls is performed for an ensemble of the lowest free energy conformers for each oligopeptide. For an ensemble of 5000 pentapeptide conformers the total CPU requirement is on the order about 0.5 wallclock hour on the 128 parallel processor machine. However, the computational requirements are dependent on the specific size and charge distribution of the system. When considering systems with multiple titration sites, the computational cost increases considerably. For a two titratable group pentapeptide, a system of 5000 conformers requires approximately 1.5 wallclock hours on a parallel machine, while approximately 3 wallclock hours is needed for a system with three ionizable groups.

When considering the total time to calculate helix propensities for a full protein sequence, DELPHI calculations are performed only for those segments with oligopeptides exhibiting strong helix propensities in the vacuum state are considered. For BPTI, 17 pentapeptides with ionizable side chains were included in this set. Overall, 1 day of wallclock time was required to perform the DELPHI calculations, in addition to about 1 day for the initial in vacuo free energy runs.

These values can also be used to estimate the total time to calculate free energies for oligopeptides of larger protein systems. For the in vacuo free energy calculations the total wallclock time will always be 1 day as long as the number of processor exceeds the total number of oligopeptides, since each oligopeptide is run sequentially. When considering the DELPHI calculations, although the dependence is approximately linear, the actual result varies according to the number of residues with titratable side chains and their occurrence in the set of oligopeptides. If we consider a 100 residue sequence with a composition similar to BPTI, the number of wallclock days required for the DELPHI calculations will double for a 128 processor machine (2 days instead of 1). The time can be easily reduced to 1 wallclock day by doubling the number of available processors.

11 Conclusions

A general method has been developed for true ab initio prediction of helix propensity for residues in a given protein sequence. An important component of the approach is that some information regarding helix formation is retained locally, which is evidenced by experimental observations regarding the strong nucleation characteristics of helices. In order to capture local interactions and the unique positioning of each residue in the overall protein, the protein sequence is decomposed into overlapping oligopeptides. The analysis also involves detailed atomistic level modeling, and the refinement of helix propensities according to polarization and ionization energies calculated through the solution of the nonlinear Poisson Boltzmann equation. The end result is the prediction of helical segments according to the average helix propensity assigned to each residue.

The approach has been applied the location of α and $3-10$ helices for three benchmark proteins which have been studied both experimentally and through simulation : bovine pancreatic trypsin inhibitor, immunoglobulin binding domain of protein G and chymotrypsin inhibitor 2. For BPTI and chymotrypsin inhibitor 2, the ab initio study based on overlapping pentapeptides and the experimental results have excellent agreement. For the immunoglobulin binding domain of protein G, the study of overlapping heptapeptides and pentapeptides provides very good agreement with the experimental results.

Acknowledgments

The authors gratefully acknowledge financial support from the National Science Foundation, Air Force Office of Scientific Research, and the National Institutes of Health (R01 GM52032).

References

1. V. Munoz and L. Serrano, *Nat Struct Biol*, 1, 399–409 (1994).
2. R. Srinivasan and G. D. Rose, *PNAS*, 96, 14258–14263 (1999).
3. R. L. Baldwin and G. D. Rose, *TIBS*, 24, 26–33 (1999).
4. R. L. Baldwin and G. D. Rose, *TIBS*, 24, 77–83 (1999).
5. K. A. Dill, *Prot Sci*, 8, 1166–1180 (1999).
6. R. L. Baldwin, *Biophys Chem*, 55, 127–135 (1995).
7. W. A. Lim and R. T. Sauer, *J Mol Biol*, 219, 359–376 (1991).
8. C-K Chan, Y. Hu, S. Takahashi, D. L. Rousseau, W. A. Eaton, and J. Hofrichter, *PNAS*, 94, 1779–1784 (1997).
9. D. Hamada, S. Segawa, and Y. Goto, *Nat Struct Biol*, 3, 868–873 (1996).
10. B. Honig and F. E. Cohen, *Fold Des*, 1, R17–R20 (1996).
11. B. Honig and A. S. Yang, *Adv Prot Chem*, 46, 27–58 (1995).
12. G. J. Barton, In M. J. E. Sternberg, editor, *Protein Structure Prediction*, pages 31–64, Oxford, 1996. Oxford University Press.
13. R. Schneider, A. de Daruvar, and C. Sander, *Nucleic Acids Res.*, 25, 226–230 (1997).
14. T. Niermann, K. Kirschner, and I. P. Crawford, *Biol. Chem.*, 368, 1087–1088 (1987).
15. D. F. Feng and R. F. Doolittle, *Methods Enzymol.*, 266, 368–382 (1996).
16. M. J. Thompson and R. A. Goldstein, *Proteins*, 25, 28–37 (1996).
17. M. J. Thompson and R. A. Goldstein, *Proteins*, 25, 38–47 (1996).
18. R. Hughey and A. Krogh, *Comput. Applic. Biosci.*, 12, 95–107 (1996).
19. M. A. McClure, C. Smith, and P. Elton, In D. States, P. Agarwal, T. Gaasterland, L. Hunter, and R. F. Smith, editors, *Fourth Internatioinal Conference on Intelligent Systems for Molecular Biology*, pages 155–164, St Louis, MO, 1996. AAI Press.
20. C. Notredame and D. G. Higgins, *Nucleic Acids Res.*, 24, 1515–1524 (1996).

21. A. Sali and T. Blundell, In H. Bohr and S. Brunak, editors, *Protein Structure by Distance Analysis*, pages 64–87, Amsterdam, 1994. IOS Press.

22. M. S. Johnson, A. C. W. May, M. Rodionov, and J. P. Overington, *Methods Enzymol.*, 266, 575–598 (1996).

23. T. Cardozo, M. Totrov, and R. Abagyan, *Proteins*, 23, 403–414 (1995).

24. A. Sali, L. Potterton, F. Yuan, H. Vljmen, and M Karplus, *Proteins*, 23, 318–326 (1995).

25. D. Fischer and D. Eisenberg, *Protein Sci.*, 5, 947–955 (1996).

26. J. U. Bowie, R. Luthy, and D. Eisenberg, *Science*, 253, 164–169 (1991).

27. J. U. Bowie, K. Zhang, M. Wilmanns, and D. Eisenberg, *Methods Enzymol.*, 266, 598–616 (1996).

28. D. Eisenberg, R. Luthy, and A. D. McLachland, *Proteins*, 10, 229–239 (1991).

29. M. J. Sippl, *Curr. Opin. Struct. Biol.*, 5, 229–235 (1995).

30. C. Lemer, M. J. Rooman, and S. J. Wodak, *Proteins*, 23, 337–355 (1995).

31. B. Rost, In H. Bohr and S. Brunak, editors, *Protein Folds : A Distance Based Approach*, pages 132–151, Boca Raton, FL, 1995. CRC Press.

32. B. Rost, In C. Rawlings, D. Clark, R. Altman, L. Hunter, and T. Lengauer, editors, *Third International Conference on Intelligent Systems for Moleuclar Biology*, pages 314–321, Cambridge, UK, 1995. AAAI Press.

33. T. J. P. Hubbard, *Folding Design*, 1, R55–R63 (1996).

34. V. I. Lim, *J Mol. Biol.*, 88, 873–894 (1974).

35. P. Y. Chou and G. D. Fasman, *Biochem.*, 13, 211–222 (1974).

36. J. Garnier, D. J. Osguthorpe, and B. Robson, *J. Mol. Biol.*, 120, 97–120 (1978).

37. F. E. Cohen, R. M. Abarbanel, I. D. Kuntz, and R. J. Fletterick, *Biochemistry*, 22, 4894–4904 (1983).

38. A. Chakrabartty, T. Kortemme, and R. L. Baldwin, *Protein Sci.*, 3, 843–852 (1994).

39. K. T. O'Neil and W. F. DeGrado, *Science*, 250, 646–651 (1990).

40. A. Horovitz, J. M. Matthews, and A. R. Fersht, *J. Mol. Biol.*, 227, 560–568 (1992).

41. H. Qian, *J. Mol. Biol.*, 256, 663–666 (1996).

42. M. J. Zvelebil, G. J. Barton, W. R. Taylor, and M. J. E. Sternberg, *J. Mol. Biol.*, 195, 957–961 (1987).

43. B. Rost and C. Sander, *Protein Eng.*, 6, 831–836 (1993).

44. V. Di Francesco, J. Granier, and P. J. Munson, *Protein Sci.*, 5, 106–113 (1996).

45. J. Garnier, J. F. Gibrat, and B. Robson, *Methods Enzymol.*, 266, 540–553 (1996).

46. C. Geourjon and G. Deleage, *Comput. Applic. Biosci.*, 11, 681–684 (1995).

47. A. A. Salamov and V. V. Solovyev, *J. Mol. Biol.*, 247, 11–15 (1995).

48. D. T. Jones, *J. Mol. Biol.*, 292, 195–202 (1999).

49. L. J. McGuffin, K. Bryson, and D. T. Jones, *Bioinformatics*, 16, 404–405 (2000).

50. S. A. Benner and D. Gerloff, *Adv. Enzyme Regul.*, 31, 121–181 (1990).

51. S. A. Benner, I. Badcoe, M. A. Cohen, and D. L. Gerloff, *J. Mol. Biol.*, 235, 926–958 (1994).

52. B. Rost and C. Sander, *Proteins*, 20, 216–226 (1994).

53. M. J. Sippl, *J. Mol. Biol.*, 260, 644–648 (1996).

54. M. J. Sippl, M. Ortner, M. Jaritz, P. Lackner, and H. Floeckner, *Folding Design*, 1, 289–298 (1996).

55. J. L. Klepeis and C. A. Floudas, *J Chem Phys*, 110, 7491–7512 (1999).

56. D.L. Minor and P.S. Kim, *Nature*, 380, 730 (1996).

57. C.B. Anfinsen and H.A. Scheraga, *Advances In Protein Chemistry*, 29, 205 (1975).

58. G. Némethy, K. D. Gibson, K. A. Palmer, C. N. Yoon, G. Paterlini, A. Zagari, S. Rumsey, and H. A. Scheraga, *J. Phys. Chem.*, 96, 6472 (1992).

59. F. H. Stillinger and T. A. Weber, *J. Stat. Phys.*, 52, 1429–1445 (1988).

60. J. Lee, H. A. Scheraga, and S. Rackovsky, *J Comp Chem*, 18, 1222–1232 (1997).

61. P. J. Flory, *Macromolecules*, 7, 381–392 (1974).

62. N. Go and H. A. Scheraga, *J. Chem. Phys.*, 51, 4751–4767 (1969).

63. N. Go and H. A. Scheraga, *Macromolecules*, 9, 535–542 (1976).

64. B. Honig, K. Sharp, and A. Yang, *J. Phys. Chem.*, 97, 1101–1109 (1993).

65. M. Gilson, K. Sharp, and B. Honig, *J Comp Chem*, 9, 327–335 (1987).

66. Y. N. Vorobjev and H. A. Scheraga, *J Comp Chem*, 18, 569–583 (1997).

67. B. Honig and A. Nicholls, *Science*, 268, 11144–1149 (1995).

68. M. Gilson and B. Honig, *Proteins*, 4, 7 (1988).

69. D. R. Ripoll, Y. N. Vorobjev, A. Liwo, J. A. Vila, and H. A. Scheraga, *J Mol Bio*, 264, 770–783 (1996).

70. A-S Yang, M. R. Gunner, R. Sampogna, K. Sharp, and B. Honig, *Proteins*, 15, 252–265 (1993).

71. J. Deisenhofer and W. Steigemann, *Acta Crystallogr. Sect B*, 31, 238–250 (1975).

72. A. Wlodawer, J. Walter, R. Huber, and L. Sjolin, *J. Mol. Biol.*, 180, 301–329 (1984).

73. T. E. Creighton, E. Kalef, and R. Arnon, *J. Mol. Biol.*, 123, 129–147 (1978).

74. T. E. Creighton and D. P. Goldenberg, *J. Mol. Biol.*, 179, 497–526 (1984).

75. D. J. States, T. E. Creighton, C. M. Dobson, and M. Karplus, *J. Mol. Biol.*, 195, 731–739 (1984).

76. T. E. Creighton, *Biochem.*, 270, 1–16 (1990).

77. J. S. Weissman and P. S. Kim, *Science*, 253, 1386–1393 (1991).

78. A. W. Burgess and H. A. Scheraga, *Proc. Natl. Acad. Sci.*, 72, 1221–1225 (1975).

79. M. Levitt, *J. Mol. Biol.*, 170, 723–764 (1983).

80. M. H. Hao, M. R. Pincus, S. Rackovsky, and H. A. Scheraga, *Biochem.*, 32, 9614–9631 (1993).

81. A. M. Gronenborn, D. R. Filpula, N. Z. Essig, A. Achari, M. Whitlow, P. T. Wingfield, and G. M. Clore, *Science*, 253, 657–660 (1991).

82. T. Gallagher, P. Alexander, P. Bryan, and G. L. Gilliland, *Biochem.*, 33, 4721–4729 (1994).

83. V. Munoz, P. A. Thompson, J. Hofrichter, and W. A. Eaton, *Nature*, 390, 196–199 (1997).

84. V. S. Pande and D. S. Rokhsar, *PNAS*, 96, 9062–9067 (1999).

85. A. R. Dinner, T. Lazaridis, and M. Karplus, *PNAS*, 96, 9068–9073 (1999).

86. Z. Bryant, V. S. Pande, and D. S. Rokhsar, *Biophys J*, 78, 584–589 (2000).

87. S. E. Jackson and A. R. Fersht, *Biochem.*, 30, 10428–10435 (1991).

88. L. S. Itzhaki, D. E. Otzen, and A. R. Fersht, *J. Mol. Biol.*, 254, 260–288 (1995).

89. T. Lazaridis and M. Karplus, *Science*, 278, 1928–1931 (1997).

90. B. Nolting, *J. Theor. Biol.*, 197, 113–121 (1999).

91. L. J. McGuffin, K. Bryson, and D. T. Jones, *Bioinformatics*, 16, 404–405 (2000).

92. S. F. Altschul, T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman, *Nucleic Acids Res.*, 25, 3389–3402 (1997).

Table I: Pentapeptide sequences used for α helix prediction of BPTI. The second column assigns the identifier for each pentapeptide, while and (X) in the third column indicates the location, determined by experiment, of a helical core for the given sequence.

Pentapeptide	ID	PDB	Pentapeptide	ID	PDB
Arg Pro Asp Phe Cys	1	X	Gly Leu Cys Gln Thr	28	
Pro Asp Phe Cys Leu	2	X	Leu Cys Gln Thr Phe	29	
Asp Phe Cys Leu Glu	3		Cys Gln Thr Phe Val	30	
Phe Cys Leu Glu Pro	4		Gln Thr Phe Val Tyr	31	
Cys Leu Glu Pro Pro	5		Thr Phe Val Tyr Gly	32	
Leu Glu Pro Pro Tyr	6		Phe Val Tyr Gly Gly	33	
Glu Pro Pro Tyr Thr	7		Val Tyr Gly Gly Cys	34	
Pro Pro Tyr Thr Gly	8		Tyr Gly Gly Cys Arg	35	
Pro Tyr Thr Gly Pro	9		Gly Gly Cys Arg Ala	36	
Tyr Thr Gly Pro Cys	10		Gly Cys Arg Ala Lys	37	
Thr Gly Pro Cys Lys	11		Cys Arg Ala Lys Arg	38	
Gly Pro Cys Lys Ala	12		Arg Ala Lys Arg Asn	39	
Pro Cys Lys Ala Arg	13		Ala Lys Arg Asn Asn	40	
Cys Lys Ala Arg Ile	14		Lys Arg Asn Asn Phe	41	
Lys Ala Arg Ile Ile	15		Arg Asn Asn Phe Lys	42	
Ala Arg Ile Ile Arg	16		Asn Asn Phe Lys Ser	43	
Arg Ile Ile Arg Tyr	17		Asn Phe Lys Ser Ala	44	
Ile Ile Arg Tyr Phe	18		Phe Lys Ser Ala Glu	45	
Ile Arg Tyr Phe Tyr	19		Lys Ser Ala Glu Asp	46	
Arg Tyr Phe Tyr Asn	20		Ser Ala Glu Asp Cys	47	X
Tyr Phe Tyr Asn Ala	21		Ala Glu Asp Cys Met	48	X
Phe Tyr Asn Ala Lys	22		Glu Asp Cys Met Arg	49	X
Tyr Asn Ala Lys Ala	23		Asp Cys Met Arg Thr	50	X
Asn Ala Lys Ala Gly	24		Cys Met Arg Thr Cys	51	X
Ala Lys Ala Gly Leu	25		Met Arg Thr Cys Gly	52	X
Lys Ala Gly Leu Cys	26		Arg Thr Cys Gly Gly	53	
Ala Gly Leu Cys Gln	27		Thr Cys Gly Gly Ala	54	

Table II: Summary of pentapeptide information. The second column provides the number of dihedral angles for the peptide, while the third column indicates the total number of unique conformers identified during the ensemble generation of the uncharged pentapeptides.

ID	DA	Unique	ID	DA	Unique
1	33	16047	28	33	13306
2	32	15176	29	35	15591
3	35	15537	30	34	14253
4	32	13700	31	36	17216
5	29	9503	32	32	13296
6	30	12065	33	29	11578
7	29	13228	34	28	9785
8	27	8945	35	32	14398
9	28	11222	36	30	9998
10	27	11004	37	35	11704
11	29	9783	38	42	16344
12	27	8202	39	44	17904
13	36	15528	40	40	13700
14	39	13817	41	41	15364
15	42	14838	42	41	17997
16	44	17851	43	36	18261
17	46	20289	44	34	15945
18	41	17768	45	35	14153
19	40	14571	46	36	14136
20	39	14240	47	32	14248
21	33	14692	48	34	12565
22	35	15915	49	40	12630
23	34	13479	50	39	16798
24	31	11169	51	37	9503
25	32	10419	52	36	16426
26	32	12468	53	32	13802
27	31	12116	54	26	13127

Table III: Pentapeptide sequences used for α helix prediction of Protein G. The second column assigns the identifier for each pentapeptide, while and (X) in the third column indicates the location, determined by experiment, of a helical core for the given sequence.

Pentapeptide	ID	PDB	Pentapeptide	ID	PDB
Met Thr Tyr Lys Leu	1		Glu Lys Val Phe Lys	27	X
Thr Tyr Lys Leu Ile	2		Lys Val Phe Lys Gln	28	X
Tyr Lys Leu Ile Leu	3		Val Phe Lys Gln Tyr	29	X
Lys Leu Ile Leu Asn	4		Phe Lys Gln Tyr Ala	30	X
Leu Ile Leu Asn Gly	5		Lys Gln Tyr Ala Asn	31	X
Ile Leu Asn Gly Lys	6		Gln Tyr Ala Asn Asp	32	X
Leu Asn Gly Lys Thr	7		Tyr Ala Asn Asp Asn	33	
Asn Gly Lys Thr Leu	8		Ala Asn Asp Asn Gly	34	
Gly Lys Thr Leu Lys	9		Asn Asp Asn Gly Val	35	
Lys Thr Leu Lys Gly	10		Asp Asn Gly Val Asp	36	
Thr Leu Lys Gly Glu	11		Asn Gly Val Asp Gly	37	
Leu Lys Gly Glu Thr	12		Gly Val Asp Gly Glu	38	
Lys Gly Glu Thr Thr	13		Val Asp Gly Glu Trp	39	
Gly Glu Thr Thr Thr	14		Asp Gly Glu Trp Thr	40	
Glu Thr Thr Thr Glu	15		Gly Glu Trp Thr Tyr	41	
Thr Thr Thr Glu Ala	16		Glu Trp Thr Tyr Asp	42	
Thr Thr Glu Ala Val	17		Trp Thr Tyr Asp Asp	43	
Thr Glu Ala Val Asp	18		Thr Tyr Asp Asp Ala	44	
Glu Ala Val Asp Ala	19		Tyr Asp Asp Ala Thr	45	
Ala Val Asp Ala Ala	20		Asp Asp Ala Thr Lys	46	
Val Asp Ala Ala Thr	21		Asp Ala Thr Lys Thr	47	
Asp Ala Ala Thr Ala	22	X	Ala Thr Lys Thr Phe	48	
Ala Ala Thr Ala Glu	23	X	Thr Lys Thr Phe Thr	49	
Ala Thr Ala Glu Lys	24	X	Lys Thr Phe Thr Val	50	
Thr Ala Glu Lys Val	25	X	Thr Phe Thr Val Thr	51	
Ala Glu Lys Val Phe	26	X	Phe Thr Val Thr Glu	52	

Table IV: Summary of pentapeptide information. The second column provides the number of dihedral angles for the peptide, while the third column indicates the total number of unique conformers identified during the ensemble generation of the uncharged pentapeptides.

ID	DA	Unique	ID	DA	Unique
1	40	11506	27	40	16465
2	40	13998	28	40	17392
3	41	17160	29	38	14052
4	41	6373	30	36	15207
5	36	13446	31	37	15127
6	37	14089	32	35	14520
7	36	10365	33	34	13436
8	36	15659	34	31	11423
9	38	15058	35	33	14832
10	38	14563	36	33	14796
11	37	13340	37	30	13140
12	37	14899	38	31	11266
13	36	14534	39	33	14842
14	34	13604	40	33	15295
15	38	16686	41	33	16180
16	35	13942	42	36	16356
17	35	13027	43	35	17710
18	35	13508	44	34	13052
19	33	11803	45	34	14942
20	30	7966	46	36	15524
21	32	9743	47	36	15206
22	30	9296	48	35	14521
23	31	9297	49	37	13870
24	35	13341	50	37	14857
25	37	13055	51	35	10579
26	36	14895	52	36	14015

Table V: Pentapeptide sequences used for α helix prediction of Chymotrypsin Inhibitor. The second column assigns the identifier for each pentapeptide, while and (X) in the third column indicates the location, determined by experiment, of a helical core for the given sequence.

Pentapeptide	ID	PDB	Pentapeptide	ID	PDB
Lys Thr Glu Trp Pro	1		Val Leu Pro Val Gly	31	
Thr Glu Trp Pro Glu	2		Leu Pro Val Gly Thr	32	
Glu Trp Pro Glu Leu	3		Pro Val Gly Thr Ile	33	
Trp Pro Glu Leu Val	4		Val Gly Thr Ile Val	34	
Pro Glu Leu Val Gly	5		Gly Thr Ile Val Thr	35	
Glu Leu Val Gly Lys	6		Thr Ile Val Thr Met	36	
Leu Val Gly Lys Ser	7		Ile Val Thr Met Glu	37	
Val Gly Lys Ser Val	8		Val Thr Met Glu Tyr	38	
Gly Lys Ser Val Glu	9		Thr Met Glu Tyr Arg	39	
Lys Ser Val Glu Glu	10	X	Met Glu Tyr Arg Ile	40	
Ser Val Glu Glu Ala	11	X	Glu Tyr Arg Ile Asp	41	
Val Glu Glu Ala Lys	12	X	Tyr Arg Ile Asp Arg	42	
Glu Glu Ala Lys Lys	13	X	Arg Ile Asp Arg Val	43	
Glu Ala Lys Lys Val	14	X	Ile Asp Arg Val Arg	44	
Ala Lys Lys Val Ile	15	X	Asp Arg Val Arg Leu	45	
Lys Lys Val Ile Leu	16	X	Arg Val Arg Leu Phe	46	
Lys Val Ile Leu Gln	17	X	Val Arg Leu Phe Val	47	
Val Ile Leu Gln Asp	18	X	Arg Leu Phe Val Asp	48	
Ile Leu Gln Asp Lys	19		Leu Phe Val Asp Lys	49	
Leu Gln Asp Lys Pro	20		Phe Val Asp Lys Leu	50	
Gln Asp Lys Pro Glu	21		Val Asp Lys Leu Asp	51	
Asp Lys Pro Glu Ala	22		Asp Lys Leu Asp Asn	52	
Lys Pro Glu Ala Gln	23		Lys Leu Asp Asn Ile	53	
Pro Glu Ala Gln Ile	24		Leu Asp Asn Ile Ala	54	
Glu Ala Gln Ile Ile	25		Asp Asn Ile Ala Gln	55	
Ala Gln Ile Ile Val	26		Asn Ile Ala Gln Val	56	
Gln Ile Ile Val Leu	27		Ile Ala Gln Val Pro	57	
Ile Ile Val Leu Pro	28		Ala Gln Val Pro Arg	58	
Ile Val Leu Pro Val	29		Gln Val Pro Arg Val	59	
Val Pro Arg Val Gly	30				

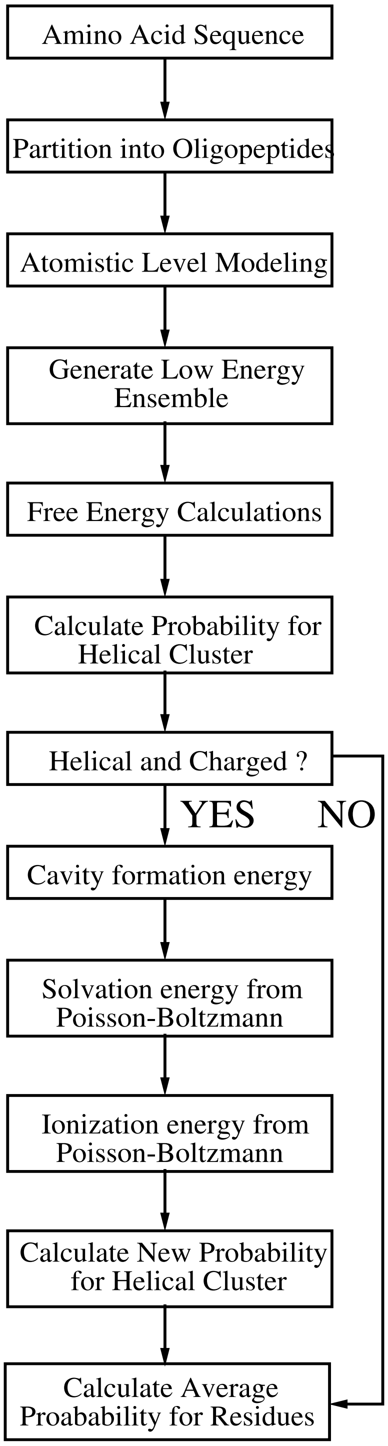
Table VI: Summary of pentapeptide information. The second column provides the number of dihedral angles for the peptide, while the third column indicates the total number of unique conformers identified during the ensemble generation of the uncharged pentapeptides.

ID	DA	Unique	ID	DA	Unique
1	35	10298	31	30	11437
2	33	11321	32	32	11628
3	34	12697	33	34	11175
4	33	14703	34	34	10990
5	33	12990	35	38	13885
6	37	13776	36	39	15161
7	35	13654	37	38	15321
8	34	10664	38	42	17316
9	35	10727	39	43	16508
10	39	15274	40	42	16825
11	35	13712	41	45	20074
12	38	13365	42	45	17963
13	40	14972	43	45	17897
14	39	13970	44	45	17573
15	39	13485	45	44	18113
16	42	16415	46	40	14218
17	41	17093	47	40	16916
18	39	14896	48	38	14668
19	41	15837	49	38	16246
20	37	13871	50	39	14241
21	36	12371	51	39	15462
22	33	10227	52	40	16039
23	34	13767	53	36	12919
24	35	13498	54	36	14734
25	38	12992	55	36	13957
26	37	12585	56	33	9773
27	40	15102	57	35	11123
28	36	11470	58	37	12567
29	34	9755	59	33	11325
30	30	8804			

Table VII: Total number of DELPHI calls required for calculation of F_{solv}, ΔG_j and ΔG_{jk} terms.

No. ionizable groups	1	2	3	4	5	N
F_{solv}	2	2	2	2	2	2
ΔG_j	6	12	18	24	30	6N
ΔG_{jk}	0	4	12	24	40	2N(N-1)
Total	8	18	32	50	72	$2(N+1)^2$

List of figures



Amino Acid Sequence
Partition into Oligopeptides
Atomistic Level Modeling
Generate Low Energy Ensemble
Free Energy Calculations
Calculate Probability for Helical Cluster
Helical and Charged ?
YES
NO
Cavity formation energy
Solvation energy from Poisson-Boltzmann
Ionization energy from Poisson-Boltzmann
Calculate New Probability for Helical Cluster
Calculate Average Proabability for Residues

Figure 1:

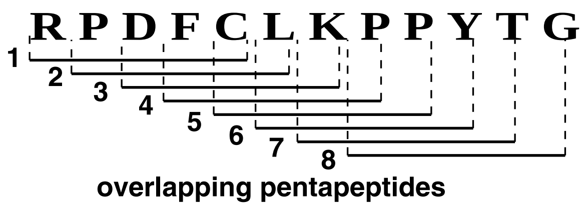


Figure 2:

$$
\begin{aligned}
E_{ECEPP/3} = & \sum_{(ij)\in ES} \frac{q_i q_j}{r_{ij}} && \text{(Electrostatic)} \\
& + \sum_{(ij)\in NB} F_{ij} \frac{A_{ij}}{r_{ij}^{12}} - \frac{C_{ij}}{r_{ij}^{6}} && \text{(Nonbonded)} \\
& + \sum_{(ij)\in HX} \frac{A'_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{10}} && \text{(Hydrogen bonded)} \\
& + \sum_{k\in TOR} \left(\frac{E_{o,k}}{2}\right)(1 + c_k \cos n_k \theta_k) && \text{(Torsional)}
\end{aligned}
$$

Figure 3:

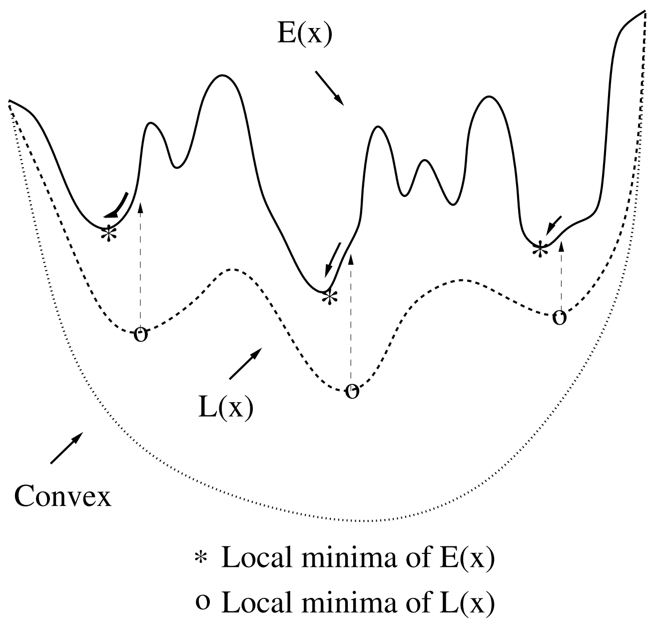


Figure 4:

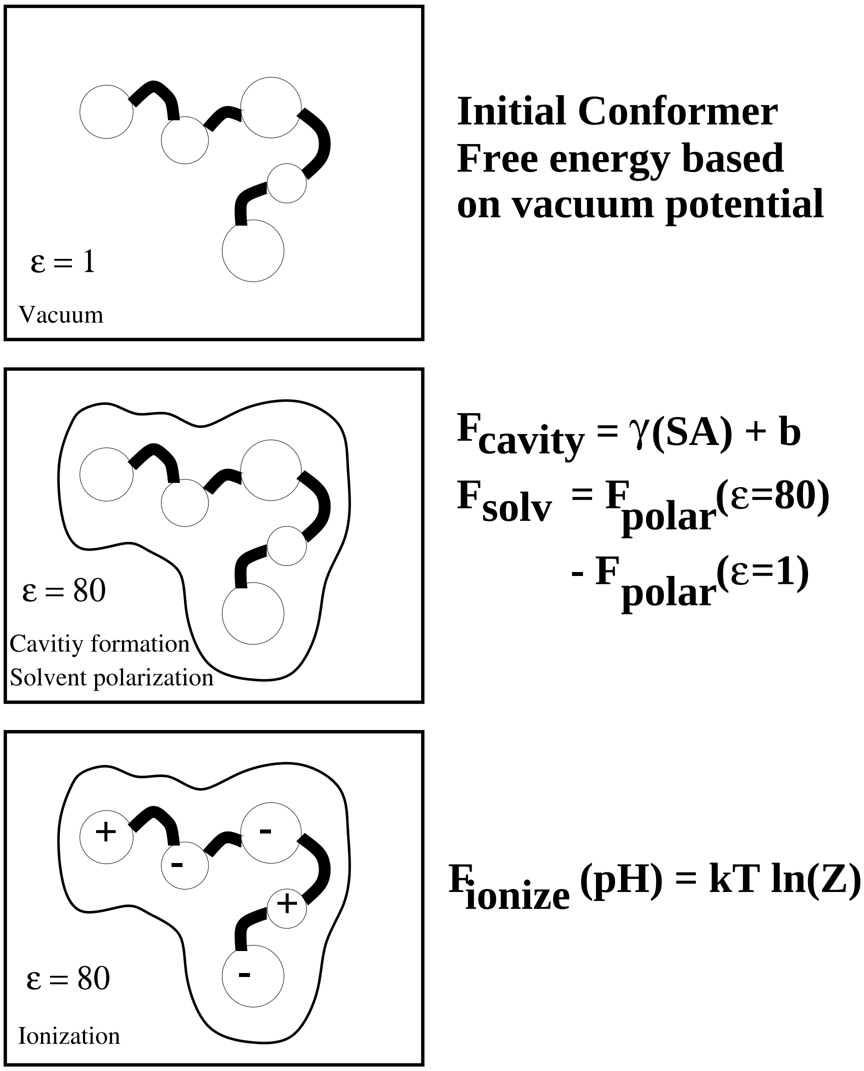

ε = 1
Vacuum
Initial Conformer
Free energy based
on vacuum potential
ε = 80
Cavitiy formation
Solvent polarization
$F_{cavity} = \gamma(SA) + b$
$F_{solv} = F_{polar}(\varepsilon=80) - F_{polar}(\varepsilon=1)$
ε = 80
Ionization
$F_{ionize}(pH) = kT \ln(Z)$

Figure 5:

$$
\begin{tikzcd}
S_i^+ \arrow[r] \arrow[d, swap, "\Delta G_j(S^+/PS_i^+)"] & S_i^o \arrow[d, "\Delta G_j(S_i^o/PS_i^o)"] \\
PS_i^+ \arrow[r] & PS_i^o
\end{tikzcd}
$$

Figure 6:

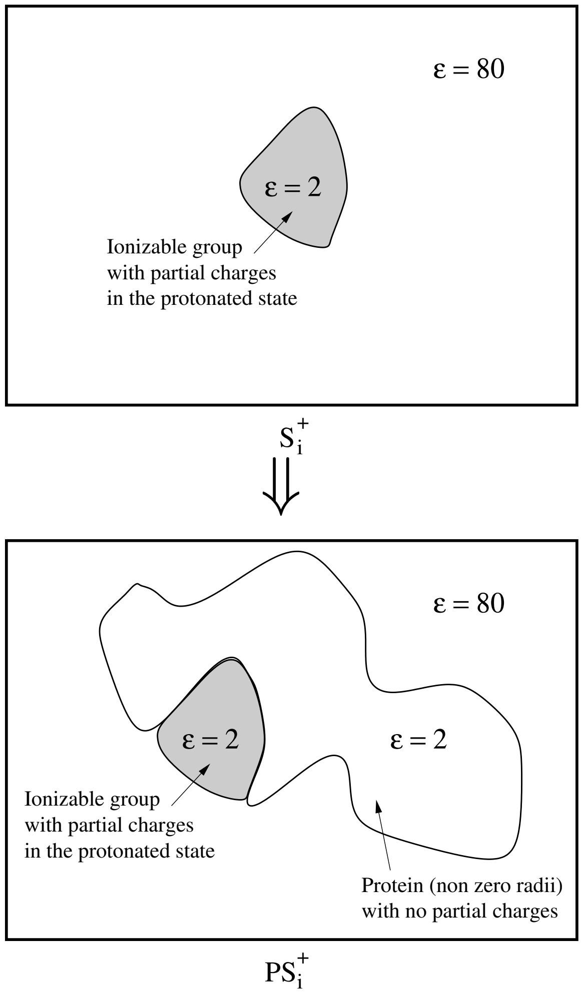

ε = 80
ε = 2
Ionizable group
with partial charges
in the protonated state
S$_i^+$
ε = 80
ε = 2
ε = 2
Ionizable group
with partial charges
in the protonated state
Protein (non zero radii)
with no partial charges
PS$_i^+$

Figure 7:

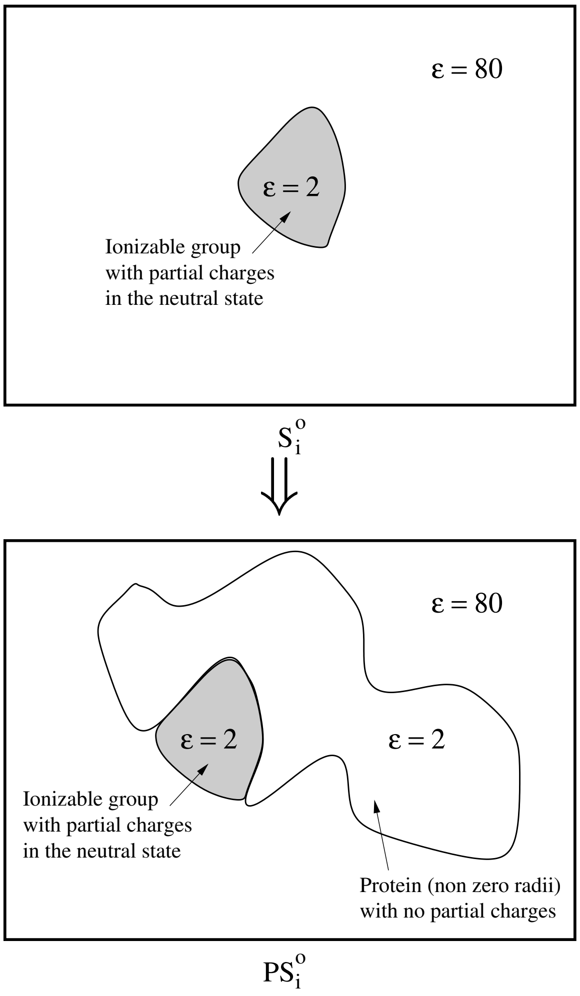

ε = 80
ε = 2
Ionizable group
with partial charges
in the neutral state
S$_i^o$
ε = 80
ε = 2
ε = 2
Ionizable group
with partial charges
in the neutral state
Protein (non zero radii)
with no partial charges
PS$_i^o$

Figure 8:

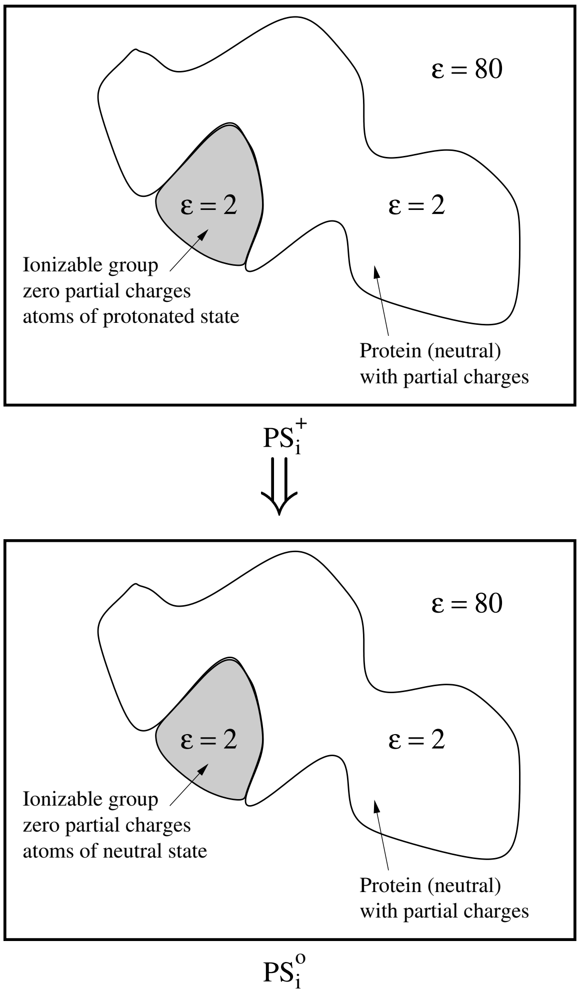
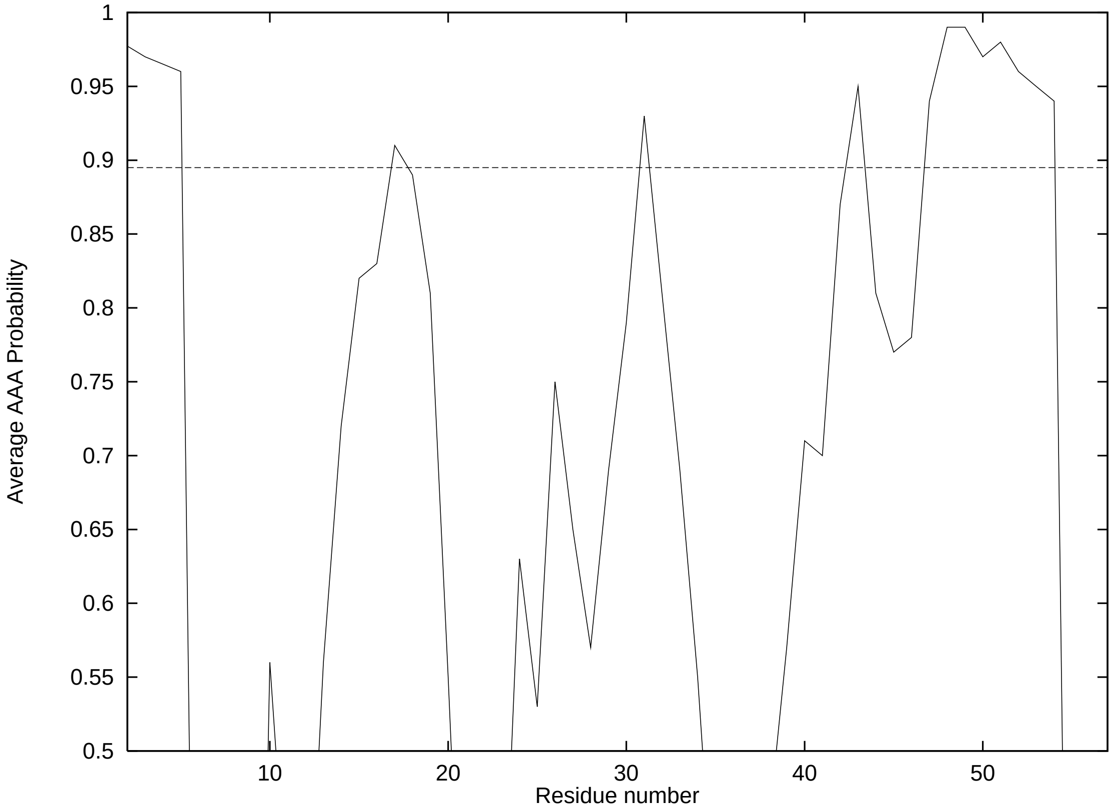
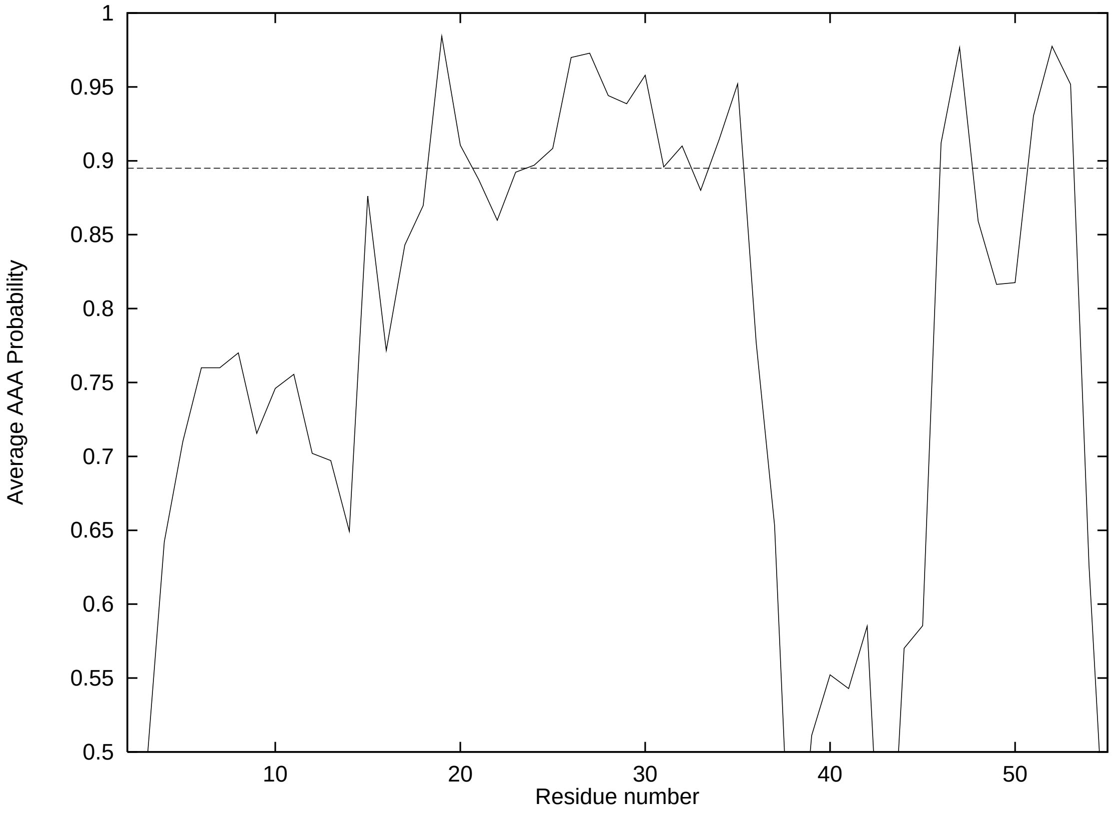

ε = 80
ε = 2
ε = 2
Ionizable group
zero partial charges
atoms of protonated state
Protein (neutral)
with partial charges
PS_i^+
ε = 80
ε = 2
ε = 2
Ionizable group
zero partial charges
atoms of neutral state
Protein (neutral)
with partial charges
PS_i^o

Figure 9:



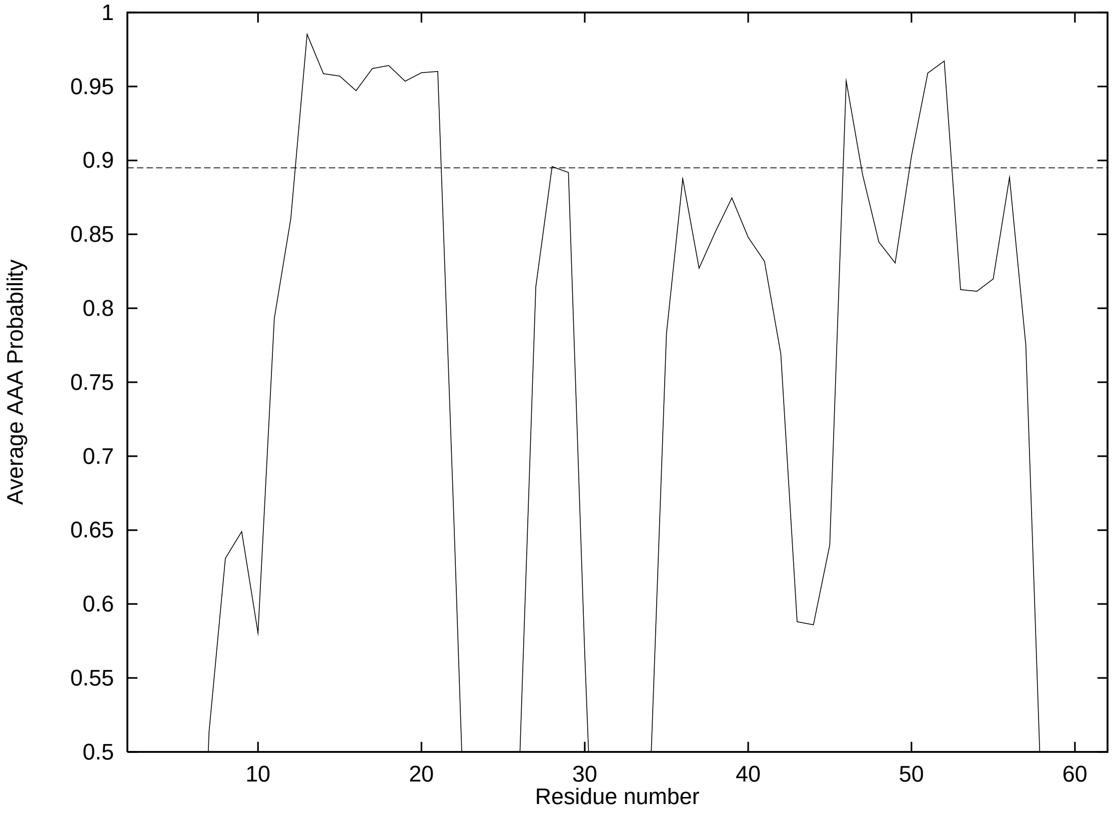
Average AAA Probability
Residue number
1
0.95
0.9
0.85
0.8
0.75
0.7
0.65
0.6
0.55
0.5
10
20
30
40
50

Figure 10:



1
0.95
0.9
0.85
0.8
0.75
0.7
0.65
0.6
0.55
0.5
Average AAA Probability
10
20
30
40
50
Residue number

Figure 11:



1
0.95
0.9
0.85
0.8
0.75
0.7
0.65
0.6
0.55
0.5
10
20
30
40
50
60
Residue number
Average AAA Probability

Figure 12: