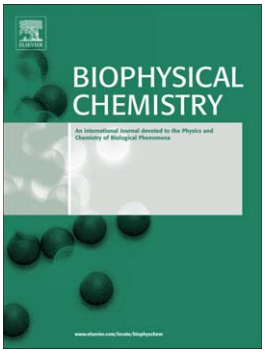
Selecting sequences that fold into a defined 3D structure: A new approach for protein design based on molecular dynamics and energetics

Giulia Morra, Chiara Baragli, Giorgio Colombo



Selecting sequences that fold into a defined 3D structure: A new approach for protein design based on molecular dynamics and energetics

Giulia Morra, Chiara Baragli, Giorgio Colombo

PII: S0301-4622(09)00217-8
DOI: doi: 10.1016/j.bpc.2009.10.007
Reference: BIOCHE 5310

To appear in: *Biophysical Chemistry*

Received date: 1 September 2009
Revised date: 7 October 2009
Accepted date: 26 October 2009

Please cite this article as: Giulia Morra, Chiara Baragli, Giorgio Colombo, Selecting sequences that fold into a defined 3D structure: A new approach for protein design based on molecular dynamics and energetics, *Biophysical Chemistry* (2009), doi: 10.1016/j.bpc.2009.10.007

Selecting Sequences that Fold into a Defined 3D Structure: A New Approach for Protein Design Based on Molecular Dynamics and Energetics.

Giulia Morra, Chiara Baragli and Giorgio Colombo*

Istituto di Chimica del Riconoscimento Molecolare, CNR, Via Mario Bianco 9, 20131 Milano, Italy

***)** Corresponding Author: Giorgio Colombo, Istituto di Chimica del Riconoscimento Molecolare, CNR; Via Mario Bianco 9, 20131 Milano, Italy. E-mail: g.colombo@icrm.cnr.it. Tel: ++39-02-28500031, Fax: ++39-02-28901239.

Abstract

The problem of finding amino acid sequences able to fold into a defined three-dimensional (3D) structure is at the basis of successful protein design efforts.

Herein, we present the results of the application of a novel, all-atom molecular dynamics based, energy decomposition approach to the selection of sequences able to fold into a given 3D conformation. First, the energy decomposition approach is applied to natural sequences associated to a well-defined structure to identify the principal energetic coupling interactions necessary to stabilize it, defining the specific energetic signature for the fold. Then, several different sequences are threaded on the defined 3D structure and only those sequences whose energetic signature (pattern) is close to that of the natural sequence, according to a similarity criterion, are selected as able to populate the specific fold. Furthermore, it is possible to evaluate the fitness of a certain sequence for a fold by combining the information provided by the energetic signature to that contained in the contact map, which recapitulates the fold topology. The results show that the better fit between the energetic properties of a sequence and the topology corresponds to a better stabilization of the protein fold by that sequence.We applied this approach to a library of natural and artificial WW domain sequences, previously developed by the Ranganathan group, containing sequences that are experimentally known to be able and unable to fold into native structures. The results show that our approach can correctly identify 70% of the sequences known to populate the typical WW domain fold.

Keywords: Protein Folding, Protein Design, Molecular Dynamics, Self-organization.

Introduction.

Successful protein design relies on the correct identification of sequences that fold into defined three-dimensional (3D) structures. This problem, also known as the "inverse protein folding" problem, can be tackled effectively by specifying what information in the sequence is necessary and sufficient to determine a certain fold.

Since the seminal work of the Eisenberg group [1], considerable progress has been made the development of computational methods for identifying amino acid sequences compatible with a target structure [2-6]. Mayo and coworkers reported one of the most notable examples, namely the complete redesign of a zinc finger protein [5]. Using different atomistic energy functions that mimic the physical interactions between aminoacids, several other groups have achieved outstanding successes in redesigning natural folds, in the *de novo* construction of novel folds, or in the re-design of enzymes [7-9].

The basic principle in atomistic design is the optimization of a target potential function providing sequences with defined thermodynamic minima corresponding to the native configurations, well separated from alternative conformational states. Having a deep free-energy minimum for the native state conformation ensures the production of a sequence with high thermal stability. However, the native states of natural proteins reside in shallow free energy minima corresponding to partially stable and dynamic folds, characterized by a multiplicity of (similar) conformations.

Starting from these concepts, Ranganathan et al. proposed a different strategy based on applying Statistical Coupling Analysis (SCA) to multiple sequence alignments of a protein family, to identify the mutual inter-residue dependencies evidenced by conserved statistical correlations between amino acid distributions at specific sites [10-14]. The application of this approach showed that a small set of residues at specific positions (in a certain protein family) coevolves among a majority which are largely uncoupled, and that the strongly coevolving residues are organized into spatially connected networks stabilizing their respective structures through packing interactions. If used in the design and selection of new sequences folding to a certain target-structure, this purely statistical, mechanism-free method should in principle produce sequences with the same marginal stabilities and biological functions as those of natural proteins. Ranganathan and coworkers were actually able to design artificial WW domains showing thermodynamic and structural properties in excellent agreement with the ones of their natural counterparts [11,14]. Moreover, the authors showed that the artificial sequences could

perform the same function as the native ones, showing class-specific recognition of proline-containing target peptides [11,14].

In this paper, we aim to analyze the energetic determinants of the sparse architecture of residue-residue interactions necessary to stabilize a certain fold, based on the analysis of the conformational dynamics and interactions in the native state of a natural protein, and to use this information for the selection of other, non-natural sequences able to fold to the same target 3D structure.

This approach is based on a recently introduced Energy Decomposition Method, aimed at identifying the key residues (*interaction hot spots*) for the stabilization and folding of the protein to a defined 3D structure [15-19]. Previously, we showed the ability of this method to capture the essential changes occurring in the energetics of a protein upon single amino acid mutation [15-19]. These changes are mainly related to stability variations in an ensemble of single-point mutants of a certain protein [19]. The main obstacle in trying to define what properties of a sequence are necessary to define a certain fold and how structural constraints impact on the selection of a certain sequence, using atomic level resolution for the study of interactions, is represented by the vast complexity of the energetic interactions between amino acids. The Energy Decomposition Method alleviates this problem by providing a simplified view of stabilizing interactions, extracting the major contributions to energetic stability of the native structure from all-atom molecular dynamics (MD) simulations. In this method, for a protein of *N* residues, the matrix of average non-bonded interactions between pairs of residues is built from an MD trajectory. The energy map is then simplified through eigenvalue decomposition (Principal Component Analysis) [15-19].

The eigenvector associated with the lowest eigenvalue is made of *N* components, each one describing the contribution to stabilization energy provided by the corresponding protein residue. Each of the components describes the contribution of the respective aminoacid to the stabilization energy of the protein. Analysis of the *N* components of the eigenvector associated with the lowest eigenvalue was shown to single out those residues (hot sites) behaving as strongly interacting and possible stabilizing centers. In general, these residues constitute a network of strongly coupled interactions typical for a certain fold. This vectorial representation of the sequence (sequence eigenvector, SE) may be thought of as the “energetic” signature of that fold [19].

The lowest eigenvalue represents an effective coupling parameter: a variation in the first eigenvalue due to mutations or structural changes can be interpreted as a change (rescaling) in the strength (intensity) of all stabilizing interactions introduced by the mutation. A more detailed and quantitative description of the method is given in Materials and Methods.

A similar reasoning could be applied to the analysis of the structural properties. The native state structure, or designed target geometry, can be described in terms of the matrix of its native contacts (the contact matrix). This provides the essential geometrical definition of the topology of the native structure. It is known that the native state topology is a major determinant of the folding free-energy landscape of many (small) proteins. The vectorial representation of the topology of the native state is defined by the principal eigenvector of the native contact matrix (contact eigenvector, CE), which depends on the desired 3D structure [19].

The validity of the vectorial representation of stabilization energy was previously checked in the context of the calculation of the relative stability of single mutants of several proteins [19], showing good correlations between theoretical and experimental data. The components of the first eigenvector define the main attractive couplings that stabilize a certain folded state. In related protein mutants that can still fold properly to the native structure, mutations can either modulate the coupling intensity of these specific interactions (reflected in the value of the eigenvalue), or modulate the height of some peaks in the first eigenvector, without disrupting the overall signature of the profile [19]. We could also show that the similarity, defined in terms of the Pearson's correlation coefficient, between the sequence eigenvector (SE) of a certain sequence and the contact eigenvector (CE) of the native structure correlated reasonably well with the relative stability of the corresponding protein [19].

Building on these considerations, we set out to test the possibility of this approach to discriminate, in a large ensemble of sequences, those that are able to fold to a desired structure vs. those that are not. We selected a subset of the same natural and artificial WW-domain sequences, tested by Ranganathan in his seminal paper on SCA, as a means for sequence selection [11]. The conformational dynamics and energetics of each sequence were probed by all-atom Molecular Dynamics (MD) simulations in explicit water at 300K. The discrimination between folding vs. non-folding sequences was based on the calculation of the similarity between the SE for each simulated sequence and the equivalent vector in the natural WW domain protein, used for reference and not included in the set of Native sequences. Moreover, based on previous results demonstrating that higher correlations between SE and CE successfully identify proteins endowed with higher structural stability, we investigated the Pearson's coefficient similarity between SE's and CE's as a measure for discriminating productive folders from non-folders, independently of the previous knowledge of the SE of the native protein. This aspect may be relevant in the design or modification of novel folds where the main information available may be the geometry of the desired target, in the absence of statistically relevant information on the SE's of known structural homologues. Ideally, the knowledge of the topology of the target structure can be

used to select the sequences with the best energetic fit to it, simply by calculating the correlation between CE and different SE's, and using only the best fitting candidates in subsequent peptide-synthesis or protein production efforts. In this work, the starting structures for the sequences that are not present in the PDB, were actually built starting only from the Cα trace of the WW geometry through a general side-chain reconstruction algorithm. The same type of exercise may be extended to putative novel folds or structural modifications or known proteins: one might build an alpha-carbon trace corresponding to any desired geometry, thread sequences upon it with the reconstruction algorithm and use the Energy Decomposition Based method and Topological analysis we presented here to screen for putative suitable sequences. In this context, it is important to evaluate to what extent the similarity between CE and SE correlates with folding properties of the sequence.

No specific new free-energy function was built for the selection. Analysis of the results showed that the combination of energy decomposition and topological analysis is able to correctly identify 70% of the sequences folding to the natural WW structure.

Results

The analysis of the foldability of different sequences presented here is based on a description of the complex non-bonded energy of a protein through the approximated stabilization energy $E_{nb}{}^{app}$ [19]. The simplification is achieved by means of the first (most negative) eigenvalue and first eigenvector obtained by decomposition and diagonalization of the energy matrix from all-atom MD simulations (see Methods).

The eigenvector describes for each residue the amount of energy coupling it shares with all other residues in the native state of the protein. Upon mapping these couplings on the 3D structure of the protein, a connected network of strong interactions is revealed, involving distant residues in the sequence. These residues correspond to the most intense peaks. Most importantly, the first eigenvector reports on an organization of the energetics of the native state that is typical for a certain fold, defining an energetic signature for that fold. The information contained in the main energetic eigenvector (Sequence Eigenvector, SE) of a related set of sequences known to fold to the desired 3D structure can be used as a template to search for other sequences able to stabilize the same fold in native conditions, without limiting to single or double mutations.We applied the Energy decomposition method to the native YAP65 WW domain sequence whose X-ray structure is available in the Protein Data Bank (PDB id: 1k9r), in order to determine the energetic signature of the protein. Then, we considered four

sets of 8 sequences each, randomly extracted from the natural and artificial, folding and non folding, sequences from the four groups present in the study of Socolich *et al* .[11]. These are:

- Native sequences (N): they occur naturally and do fold into the WW domain structure.
- CC (coupled conservation) sequences: artificial sequences, created on the basis of the SCA and on the premise that *conservation of the pattern of coupled interactions* seen in natural sequences is sufficient to favor the folding to the desired structure [11]. Experimentally, a significant percentage of them proved to fold to the native structure by circular dichroism (CD) and NMR analysis. Our selected subset of CC sequences comprises only sequences that fold to a native WW structure under the considered experimental conditions in [11,14].
- IC: artificial sequences, created from multiple sequence alignments based on the hypothesis that conservation is a property of *one single site*, independently of others. Experimentally, no folding to the native WW domain geometry was observed [11].
- Random sequences (R): they were built by randomly mutating native sequences. As expected, they do not fold [11].

As shown by Socolich et al. [11] conservation of amino acid composition as inferred by a multiple sequence alignment is not sufficient to discriminate productive folders (CC) from non-folding sequences (IC).

All sequences and their labels are reported in the Supplementary Material

Correlation between sequence, energetics and folding to a given 3D structure.

The energy decomposition method was first applied to the wild type sequence of the YAP65 WW domain (PDB id: 1k9r). The resulting SE profile recapitulates the information on which residues at which positions are most important in the stabilization of the 3D native structure. Peaks in the SE correspond to those residues whose pair interactions with the rest of the amino acids contribute to the protein stability (Figure 1). In particular, pair interactions between two peak-related residues provide a significant stabilizing energy to the protein. As a natural consequence, the SE reports therefore also on the residue-residue couplings defining the network of interactions that contribute to the folding core of the protein, whose participating residues are indicated by the SE regions above a defined threshold [20].The SE of the YAP65 WW domain (Figure 2) indicates as energetically relevant residues a subset comprising the segments E8-S13 and Q17-D25 (numbering as in Socolich et al. [11]). Relevant residues correspond to positions in the principal eigenvector characterized by a high value of the

component. Graphically, most relevant residues (and the values of their components) are identified by peaks in the eigenvector profile. These segments are known to contain strongly coevolving residues (such as E8, Y21, H23) responsible for ligand binding and function in WW domain. Hence, according to our analysis these residues also provide a stabilizing contribution to the protein.Following this first step, the SE's for a set of eight natural WW domain sequences were calculated. The peaks in the profile of the SE's, indicating residues participating in stabilizing pair interactions, are rather conserved and the distribution between values over and under the threshold reflects the wild type case.The Pearson's correlation coefficient was used to define the similarity between the SE of 1k9r with each of the SE's calculated for the other native sequences (N set, Figure 2)

The results showed generally high values, in the range between 0.7 and 1 for the Pearson's correlation (see Table 1), and an average value of 0.80, indicating that the specific pattern of interactions defined by the Sequence Eigenvector (SE) profile may actually be considered important in the discovery of the energetic determinants of the folding features of the protein. Next, attention was focused on the artificial sequences proposed and tested by Ranganathan [11].

First, the CC group was analyzed, following the same procedure as for native sequences: for each of the eight CC sequences SE was calculated and its similarity to SE of 1k9r was measured by Pearson's coefficient. For this set, the SE profiles still retain the modulation of the wild type (Figure 3), however they show some increased variability in the peak intensities. Pearson's coefficients' values are lower than for the Native set, but still in general good agreement with the values computed for native sequences (between 0,6 and 1; Table 1, with an average of 0.72). Finally the same analysis was performed on IC and Random sequences. The IC set comprises sequences that are not folding to the native state in spite of the native-like amino acid composition site by site, resulting from simple sequence alignment analysis. The SE profiles are as noisy as in the CC case, hence these sequences do not seem to be distinguishable from the folding ones. However, when looking at the Pearson's coefficients, they turn out to be consistently lower, with an average of 0.65. For random sequences, which are expected to produce totally uncorrelated SE's, Pearson's coefficient values of for IC and Random sequences are very low values (Table 1) with an average of 0.45.

These results show that the SE is actually capable of capturing the significant energetic features for the fold. In general, higher correlations are found between the energy couplings of the native sequence and those of proteins known to fold to the typical WW domain structure. In contrast, random sequences and also the sequences missing important coupling interactions in spite of high sequence similarities (the IC set) are characterized by a much lower correlation suggesting that the energetic signature captured

by the SE is able to discriminate between the folding propensities of different sequences. The method based on the similarity with the reference WT protein correctly identifies native sequences and to a significant extent also the sequences belonging to the CC set, thus indicating that information on pair-correlations is included in the SE. The different behavior of CC and IC sequences suggests that the foldability of a sequence appears to be encoded in more subtle sequence properties than the residue composition and single site distribution. In order to check whether the differences in the correlations between SE's of CC and IC and the SE of the Wild Type (WT) does not trivially result from a sequence similarity, we re-evaluated specifically for each sequence in our dataset the alignment score with the WT sequence [21] (Table 1). While Native sequences generally show high alignment scores with the WT, the CC and IC sequences have comparable similarities and cannot be distinguished from one another based on sequence features only. Moreover, the correlation between the results of the two methods was evaluated: the Pearson's coefficients between the SE's of IC and CC peptides with WT-SE were plotted against the respective sequence alignment scores, yielding a linear correlation coefficient of 0.17, excluding random sequences from the calculation (two-sided p-value=0.406).

Interestingly, the discrimination between folders and non-folders cannot be obtained by looking at standard global properties of the structure like RMSD (data not shown). Within 5ns of MD trajectory, no significant unfolding or global structural rearrangements are possible even for a small protein like the WW domain. Interestingly, neither the total interaction energy, that is provided for each structure by the force field parameters, nor the approximate stabilization energy calculated by the Energy Decomposition Method (as already proved in [19]), are able to distinguish between folders and non-folders.

Still, the distribution of stabilizing interactions as it is described in the SE turns out to be a well defined measure for specific sequence properties such as the ability to select favorable native contacts providing stability and cooperativity to structure formation, hence determining the sequence foldability.

Correlation between energetics and topological properties of the target 3D structure. In order to add the topological information on the fold to our analysis, the contact matrix for the WW domain was calculated, and subsequently subjected to eigenvalue and eigenvector analysis. The eigenvector associated to the highest positive eigenvalue (principal eigenvalue) was considered to be representative of fold topology, and is generally referred to as the Contact Eigenvector (CE). In analogy to SE, CE

indicates which residues constitute the essential determinants of the domain architecture [22]: in general, CE singles out residues that have a high number of contacts with other residues (Figure 4).

In a previous study [19], we demonstrated that the degree of correlation between the SE and CE is a measure of the fitness of a certain sequence to a certain fold. A suitable sequence for a certain fold places the strongly interacting residues (hot spots) where they can stabilize the structure. An example is the buried core of the protein where several residues must be tightly packed to develop energetic interactions responsible for correct folding. As a consequence, the SE of a folding sequence should be similar to the CE of the fold. Importantly, by applying this concept, it was possible to rank the differences in stability of a diverse set of mutants of a series of proteins with remarkably different folds. The comparison between SE and CE allows to shed light on the degree of compatibility between a specific sequence, which provides characteristic energetic interactions, and the 3D structure that is being evaluated. Here, we attempt to evaluate to what extent the similarity between CE and SE correlates with folding properties of the sequence.

In this context, CE was calculated for the crystal structure of 1k9r as a representative of the WW fold. Subsequently the Pearson's coefficients between the SE profiles of the native sequences from the pdb and the CE were measured: as expected, the results show great similarity (with an average of 0.87; Table 1). Slightly lower values were obtained for CC sequences (average 0.82). However, also Pearson's correlations between IC sequences SE's and WW domain CE gave similar results to those of the N sequences (0.88). Finally, the similarities between the SE's and the native CE for random sequences display in general minimal values (0.65).

Taken together, this body of results suggests that the measure of similarity of the SE of a certain sequence to that of the native sequence may effectively discriminate between sequences that are either able or unable to fold to a target structure. Moreover, they suggest that while a high degree of similarity between CE of the target structure and the SE's of different sequences is necessary for a sequence to fold, it is not a sufficient criterion to determine whether a sequence can actually populate that specific fold. According to this result, the large majority of non-random sequences in our dataset satisfy this criterion, hence they are compatible with the native structure in terms of "topological" requirements, such as the number of contacts formed (residue size and hydrophobicity) at each single site. This is reflected in the linear correlation coefficient between SE-CE similarities and sequence similarities, which results equal to 0.50, in contrast to the 0.17 calculate above (two sided p-value 0.009).

Combining energetic and topological information to select viable sequences for a defined 3D fold.
Finally, the information on the correlations between the SE's of different sequences and that of the natural sequence folding to the WW domain 3-stranded geometry (1k9r) was combined with the information on the correlations between the SE's of different sequences and the CE recapitulating the properties of the WW-domain fold. The two quantities were plotted in a graph (Figure 5). The folding sequences (native and CC) define an ensemble mainly located in the right-upper part of the graph, separated from nonfolding sequences (IC and random). The limit of 0.7 for Pearson's coefficient on both axes, defines an area that contains mostly folding sequences: 11 out of 16 folding sequences (69%) are in this part of the graph, and 13 out of 16 nonfolding sequences (81%) are out of it. Hence, by setting the acceptance threshold to 0.7 we obtain 5 false negatives and only 3 false positives. Among the false negatives, only one native sequence is not recognized (N7), possibly because of a very high peak (due to the strong relative contribution at position 23) which slightly alters the overall relative distribution of peaks (as an effect of normalization). By introducing a more restrictive threshold, such as 0.8, the number of false positives drops to zero, whereas the number of false negatives increases by two units.
The combined analysis of considering both the energetics of the sequences and the topological features of the fold proves the possibility to discriminate between folding and non-folding sequences, with nearly 70% accuracy and a limited computational cost. In this procedure, we used the SE of 1k9r as a reference, representative for the determinant residue-residue coupling interactions necessary to fold into the WW domain structure. Although the chosen WWdomain was selected randomly among all structures present in the Protein Data Bank, it might introduce some bias in the classification of folding/nonfolding sequences, which is based on the similarity to its SE.

Therefore, in order to be independent of previous knowledge, a final test was performed without any previous assumption on the similarities with a certain reference sequence. In this line of thought, for each sequence, the correlation between its SE and SE's of all other sequences were measured by Pearson's coefficient. The resulting values were clustered by means of a cluster analysis. This procedure highlights the presence of a distinct, dominant cluster containing sequences that are similar to one another in terms of SE's. Strikingly, most of the sequences in this cluster were proved experimentally to fold to the required WW-domain geometry. In detail: 13 out of 17 folding sequences, including the wild type, (76%) belong to this cluster and 10 out of 16 non-folding sequences (63%) do not. However, the number of false positives also increases to 7, and includes only IC sequences.These

results confirm that folding sequences show similar energetic features, so they can be clustered together and distinguished from non-folding sequences; moreover, SE captures with good approximation the most significant energetic characteristics of each sequence. Therefore, the energy decomposition method can be a useful tool to investigate whether a sequence is likely to fold on a required structure. In this context, the CE-SE similarity criterion should be considered only as a necessary, but not sufficient, condition for a sequence to fold into a required geometry.

Discussion and Conclusion.

In this paper we applied a Molecular Dynamics based strategy to predict, given a protein fold and a set of sequences with very similar chemical features (composition, single site distribution etc), which ones are able to fold to the given structure and which ones are not, based on the hypothesis that foldability requires cooperativity and therefore correlation among different sites. The analysis was based on purely physico-chemical and structural properties; hence no previous knowledge of the folding abilities of sequences was used. The procedure entails the evaluation of the Sequence Eigenvector (SE) of each sequence on the target fold by means of the Energy Decomposition Method developed in our group [15-19]. The Energy Decomposition Method was applied to a set of 32 WW domain sequences, including 8 native sequences (N) and two groups of 8 designed sequences, one obeying both site conservation statistics and a set of selected residue-residue correlations (CC) and the other satisfying only single site conservation (IC). 8 random sequences were also considered for comparison. The similarity between each sequence's SE and the template SE from native WW domain protein 1k9r was shown to correlate with the folding capability of the sequence under examination. Around 70% of the effectively folding sequences has a Pearson's correlation coefficient of at least 0.70 with the template, whereas not folding sequences reach lower values. A threshold of 0.70 for the Pearson's correlation coefficient could therefore be used to classify between folders and non-folders to the WW domain. We also tested a classification criterion based on the comparison of a sequence's SE to the structure CE, recapitulating the structural properties of the fold. The similarity of the sequence's SE and the WW domain structure CE is not a sufficient criterion to distinguish between folding and not folding sequences, since it yields equivalent scores for N, CC and IC sequences. Apparently, the similarity to a structure's CE guarantees the correct placement of hydrophobic and polar residues in the highly buried and exposed sites, respectively. The natural occurring frequency of a given type of residue with specific chemical properties at a given site (such as polar or non polar amino acids) corresponds to the conservation statistics obtained in the multiple sequence alignment. The natural single site distribution,

where site correlations are neglected, is the information used the to build the artificial IC sequences, which hence satisfy the required constraints in terms of local burial as the N and CC sequences do. This observation correlates with the experimental finding pointed out by Ranganathan [11], that IC sequences are generally soluble, which is not the case for the majority of R sequences. This fact could indicate that the hydrophobic collapse resulting in protein solubility is accounted for by the correct hydrophobic-polar pattern along the chain, while the unique shape of the native fold requires a network of specific cooperative interactions, also involving distant sites, somehow not present in the IC sequences and whose absence is not detected when comparing their SE to the protein's CE.

This is confirmed by the statistically significant correlation we found between sequence similarity and SE-CE similarity. The SE-CE similarity evaluates the basic agreement of a candidate sequence with a target structure. In light of the results shown here, the SE-CE similarity should be considered only as a necessary, but not sufficient, condition for a sequence to fit to a certain topology. Equivalent information could be obtained by comparing the sequence to a multiple alignment of natural sequences folding to the given structure. In this respect, our approach offers the advantage of requiring in principle only the structure of the natural template protein, or information on the topology of a design-target structure recapitulated by the CE. In a typical design application the exact native structure is not available. In this context, we speculate that at the present stage the SE-CE correlation of our method might be used as mainly as a preliminary screen for sequences that can be efficiently threaded on the structure.

While the CE-SE comparison can identify sequences having a single site distribution that fit the requirements of the structure, the specific energy-related details of the cooperative interactions seem to be captured by the comparison of each sequence's SE with the template WT-SE, which in fact proves able to discriminate between CC and IC sequences. Moreover, the set of native sequences obtains the highest average similarity score with the template. The natural sequences (N) are likely to have evolved a more complex cooperativity not entirely described by the correlation data used to design the CC set, which nevertheless were shown to be essential and also sufficient for the domain to fold.

SE reflects the energetic properties of the native state. In a previous paper, we could show that the approximate stabilization energy recapitulated by the SE can be used as an effective approximation of the enthalpic part of the folding free energy. This was applied for estimating the stability difference between two sequences differing for only one mutation on a total of about 60 residues. In that case, in light of the small perturbation, it was reasonable to assume that the energetic properties of the unfolded

state are essentially identical for the two sequences, allowing a good approximation of the folding enthalpy [19]. This hypothesis may not hold here, considering sequences differing at many positions. The number of perturbations may actually alter the distribution of conformations and interactions defining the unfolded states for each of the sequences.

Given the impossibility to fully characterize the unfolded states for many different sequences, a more accurate calculation of the free energy differences between sequences could be obtained with free energy perturbation methods and thermodynamic cycles connecting different mutants. However, free energy perturbations require long equilibration times at intermediate values of the Hamiltonian coupling parameters and result in time-consuming efforts. Moreover, limitations due to the accuracy of the force field parameters and sampling issue may impact on the final outcome of the calculations. These are out of the scope of this paper.

The present analysis and data set is based on a randomly and arbitrarily chosen template, namely the native YAP65 WW domain 1k9r, whose specific properties might introduce some bias in the sequence classification. With the aim of removing this bias, we attempted the clustering strategy outlined at the end of the Results section, considering the similarity between all sequence pairs and not only with respect to the template sequence and then clustering the data into two sets, in order to distinguish between folders and non folders. The performance of the classification method improves slightly, but the template structure still might have some influence on the results. In the future, one may think of optimizing the selection of the template structure, for instance by means of a clustering procedure on the PDB data entries. Such a procedure might be suitable for inverse protein folding applications, where the target structure is known and new sequences can be selected or optimized by means of the present classification scheme. The acceptance threshold of 0.80 set for the SE, WT-SE correlation proves able to discriminate between folders and non-folders in the case of WW domain, a small protein for which very similar sequences are analyzed. In spite of the significant similarities among the sequences in the data set, both the less and the more restrictive thresholds of 0.70 and 0.80 chosen, allow the discrimination of productive folders from non-folding sequences based only on a mathematical and numerical descriptor. This constitutes a valid prerequisite to the general applicability to different systems.

The method proves suitable for applications in the field of protein design, limited at the moment to the pre-selection of sequences that may fit to a certain fold, and consequently as a sequence classification tool. In this initial work, we made use of information regarding the template structure, while in a fully

ab initio protein design application, the exact native structure of the protein may not be available and this may limit the performances of our method. It is important to recognize that further theoretical developments coupled with experimental characterization of designed sequences are required in this context. From the theoretical point of view, for instance, one may benefit from efforts to characterize the unfolded states that would give a better description of the free energy components involved in the stabilization of native states. Alternatively, one could use simulation approaches allowing wider sampling to properly weight the statistical relevance of different conformations. All these efforts require parallel advances in the performances of MD algorithms and in the development of hardware. Recent reports on these subjects hold great promise with regards to the possibility to apply the methods described here on a much larger scale and with much better sampling. [23-26]

Finally, the applications presented here require relatively short MD simulations. The choice of 5-ns long MD simulations was actually an arbitrary one, which we considered a good compromise between computational efficiency and sufficient relaxation of the structures, allowing the simulation of 32 different sequences at the all-atom level of resolution. In our experience, the Molecular Dynamics simulations spanning a short time range like 5ns, while not ensuring the complete convergence to the equilibrium structure, represent a viable compromise to reach a sufficient relaxation of the side chain and local backbone conformations, yielding a well-defined pattern of non-bonded pair interactions, with a limited computational effort. Therefore, the limited size of the required Molecular Dynamics trajectories (5 ns) as well as the speed up offered by the use of implicit solvent models in the calculation makes this approach compatible with a large scale application, such as multiple sequences screening, at least in the case of small to medium size molecules.

Materials and methods.

Socolich et al. [11] studied and experimentally tested the fold of many alternative sequences of the WW domain, using 1H-NMR tests and thermal denaturation. We have applied our method to a subset of these sequences in order to test its capability in discriminating the folding sequences. The sequences can be divided into 4 groups, they are labeled as in Socolich et al (N for Native sequences, CC for Coupled conservation, IC for Independent conservation and R for Random) [11], and reported in Table 1 and Supplementary Material.

We selected 8 sequences within each group, with the only requirements that the CC seqeuences would fold to a native WW domain under the experimental conditions, and that they did not contain gaps with respect to the template sequences, to avoid discrepancies when mounting the sequence onto the template structure.

These sequences were associated to the structure of a YAP65 WW domain, PDB id: 1k9r (NMR structure), extracted from the protein data bank. The template sequence corresponds to residues 8-40 of the original protein, in the following numbered from 1 to 33 in agreement with the numbering in Socolich et al [11].

We used Reconstruction program described in [27] to thread the sequences on 1k9r fold: this program conserves only the alpha-carbon trace in 1k9r structure and adds the side chains of the residues of the sequence. Then, it determines the coordinates of the side chain atoms, using a library of rotamers. This library contains empirical data, extracted from 100 NMR structures (chosen from PDB select). Finally, the structures were minimized in order to avoid non-realistic contacts or positions of the residues.

For each structure, after a 1000 steps minimization via the Steepest Descent algorithm, 5-ns MDs NVT simulation in a octahedral water box with explicit solvent and periodic boundary conditions are run using the GROMACS package (version 3.2.1), [28] with the GROMOS96 43Al force field [29].

The simple point charge model SPC is applied, to model water molecules [30]. All bond lengths are constrained by means of the LINCS algorithm [31]. Electrostatic interactions are treated via PME implementation of the Ewald summation method. Temperature is set to 300 K and controlled by Berendsen thermostat [32]. The timestep is set to 2 fs.We neutralized the N- and C- terminals, since charged terminals form a saline bond that decreases the mobility of the protein, resulting in a less representative sampling of conformational space.

Energy Decomposition Method.

The energy decomposition method is based on the calculation of an interaction matrix M_{ij} on a representative protein structure derived from an MD trajectory [15-19]. The matrix contains the

interaction energies between residue pairs, comprising all the nonbonded interresidue atomic energy components (namely, van der Waals and electrostatic couplings between all atoms of two residues). Solvent effects such as electrostatic shielding and pair non polar terms are are implicitly taken into account by means of the generalized Born approximation (GBSA). In the following we provide a detailed explanation of the method, which was already presented previously in [15,19].

The square matrix M_{ij} of non-bonded interactions between all residues *i* and *j* in the protein can be diagonalized and re-expressed in terms of eigenvalues and eigenvectors, in the form:

$$M_{ij} = \sum_{k=1}^{N} \lambda_k w_i^k w_j^k \quad (1)$$

where N is the number of amino acids in the protein, λ_k is an eigenvalue, and w_i^k is the i-th component of the associated normalized eigenvector. λ_1 is defined as the most negative and the following eigenvalues are labelled in increasing order. In the following we refer to the first eigenvector as the eigenvector corresponding to the eigenvalue λ_1. The total non-bonded energy E_{nb} is defined as:

$$E_{nb} = \sum_{i,j=1}^{N} M_{ij} = \sum_{i,j=1}^{N} \sum_{k=1}^{N} \lambda_k w_i{}^k w^k{}_j = \sum_{k=1}^{N} \lambda_k W_k \qquad (2)$$

where $W_k = \sum_{i,j=1} w_i{}^k w^k{}_j$. If $\lambda_1 W_1$ is larger than $\lambda_k W_k$ for $k \neq 1$, the sum over *i,j* of M_{ij} is dominated by the contribution due to the first eigenvalue and eigenvector, such that the total non bonded energy can be approximated by:

$$E_{nb} \approx E^{app}{}_{nb} = \lambda_1 \sum_{i,j=1}^{N} w_i{}^1 w^1{}_j = \lambda_1 W_1 \qquad (3)$$

The hot spots, or the residues giving a more relevant contribution to the stabilization energy, are defined as those sites whose component is higher than a threshold value *t* which is calculated as the value corresponding to a normalized vector whose components provide all the same contribution for each site (flat eigenvector). These hot spot values correspond to the "peaks" in the graphical representation of the eigenvector as a function of the sequence. This corresponds to a case in which each residue contributes with the same weight to structural stability. In this approximation the threshold value depends only on the number *N* of residues in the protein and is calculated as: $w_i{}^1 = \frac{1}{\sqrt{N}}$ for each *i* [15-19].

We analyzed the MD trajectory, via a cluster analysis with the GROMOS method with 0,2 nm cutoff. We verified that the most frequented cluster was significantly more populated than the others, not to neglect significant structural deviations captured from other clusters. The Energy Decomposition method was applied on the representative structure of this cluster.

The energy pair decomposition is obtained with the MM-PBSA adapting the algorithm implemented in AMBER8 using the GB approximation (GB model of Onufriev et al. [33]). The pairwise energy contributions are calculated as sum of gas phase interactions, solvation free energy with GB and hydrophobic term obtained with the LCPO method [34]. The contact map C_{ij} of a structure is a matrix that describes which residues are in contact in the starting conformation and is defined by looking at Cα atom pairs.

If the distance between any two Cα atoms is below a cutoff value, the corresponding matrix entry is set to l, otherwise it is set to 0. The distance cutoff is set to 0.65 nm. For the sake of homogeneity with the energy matrix, also contacts between nearest neighbors *i,j+1* are included. Therefore:

$$C_{ij} = \begin{cases} 1 & r_{ij} \le 6.5 \\ 0 & r_{ij} > 6.5 \end{cases}$$

Upon diagonalization of the matrix C_{ij}, we obtain a set of eigenvalues and eigenvectors. The principal eigenvector, corresponding to the most positive eigenvalue, has all components c_i of equal sign, which is also true for the first eigenvector of matrix M_{ij}.

The similarities among energy patterns and between contact map and energy patterns are calculated using the Pearson's correlation coefficient:

$$r(c,w) = \frac{\sum_i (c_i - < c >)(w_i - < w >)}{\sqrt{\left[\sum_i (c_i - < c >)^2\right]\left[\sum_i (w_i - < w >)^2\right]}}$$

To evaluate patterns similarities we also performed a cluster analysis with Matlab 7 release 14. In order to compute the distances among patterns, we used the function PDIST, method CORRELATION: given an *m*-by-*n* data matrix X, which is treated as *m* (1-*by*-*n*) row vectors x_1, x_2, ..., x_m, the various distances between the vector x_r and x_s are defined as follows:

$$d_{rs} = 1 - \frac{\left(x_r - \overline{x_r}\right)\left(x_s - \overline{x_s}\right)'}{\left[\left(x_r - \overline{x_r}\right)\left(x_r - \overline{x_r}\right)'\right]^{1/2}\left[\left(x_s - \overline{x_s}\right)\left(x_s - \overline{x_s}\right)'\right]^{1/2}}$$

We then clustered the patterns with the LINKAGE function (default method), which creates a hierarchical cluster tree.

References.

[1] Bowie, J.U., R. Luthy, and D. Eisenberg, A method to identify protein sequences that fold into a known three-dimensional structure, Science 253 (1991) 164-170.

[2] Watters, A.L., P. Deka, C. Corrent, D. Callender, G. Varani, T. Sosnick, and D. Baker, The highly cooperative folding of small naturally occurring proteins is likely the result of natural selection, Cell 128 (2007) 613-624.

[3] Kuhlman, B., G. Dantas, G.C. Ireton, G. Varani, B.L. Stoddard, and D. Baker, Design of a novel globular protein fold with atomic-level accuracy, Science 302 (2003) 1364-1368.

[4] Lee, M.R.T., Jerry; Baker, David; Kollman, Peter A, Molecular dynamics in the endgame of protein structure prediction, J.Mol.Biol. 313 (2001) 417-430.

[5] Dahiyat, B.I. and S.L. Mayo, De novo protein design: Fully automated sequence selection, Science 278 (1997) 82-87.

[6] Desjarlais, J.R. and T.M. Handel, De novo design of the hydrophobic cores of proteins, Protein Sci. 4 (1995) 2006-2018.

[7] Kortemme, T., L.A. Joachimiak, A.N. Bullock, A.D. Schuler, B.L. Stoddard, and D. Baker, Computational redesign of protein-protein interaction specificity, Nature Struct. Mol. Biol. 11 (2004) 371-379.

[8] Kraemer-Pecore, C.M., J.T. Lecomte, and J.R. Desjarlais, A de novo redesign of the ww domain, Protein Sci. 12 (2003) 2194-2205.

[9] Harbury, P.B., J.J. Plecs, B. Tidor, T. Alber, and P.S. Kim, High resolution protein design with backbone freedom, Science 282 (1998) 1998.

[10] Lockless, S.W. and R. Ranganathan, Evolutionarily conserved pathways of energetic connectivity in protein families, Science 286 (1999) 295-299.

[11] Socolich, M., S.W. Lockless, W.P. Russ, H. Lee, K.H. Gardner, and R. Ranganathan, Evolutionary information for specifying a protein fold, Nature 437 (2005) 512-518.

[12] Suel, G.M., S.W. Lockless, M.A. Wall, and R. Ranganathan, Evolutionary conserved networks of residues mediate allosteric communication in proteins. , Nat. Struct. Biol. 10 (2003) 59-69.

[13] Shulman, A.I., C. Larson, D.J. Magelsdorf, and R. Ranganathan, Structural determinants of allosteric ligand activation in rxr heterodimers., Cell 116 (2004) 417-429.

[14] Russ, W.P., D.M. Lowery, P. Mishra, M.B. Yaffe, and R. Ranganathan, Natural-like function in artificial ww domains, Nature 437 (2005).

[15] Tiana, G., S. Simona, G.M.S. De Mori, R.A. Broglia, and G. Colombo, Understanding the determinants of stability and folding of small globular proteins from their energetics, Protein Science 13 (2004) 113-124.

[16] Ragona, L., G. Colombo, M. Catalano, and H. Molinari, Determinants of protein stability and folding: Comparative analysis of beta-lactoglobulins and liver basic fatty acid binding protein, Proteins: Structure, Function and Bioinformatics 61 (2005) 366-376.

[17] Colacino, S., G. Tiana, R.A. Broglia, and G. Colombo, The determinants of stability in the human prion protein: Insights into the folding and misfolding from the analysis of the change in the stabilization energy distribution in different condition, Proteins: Structure, Function and Bioinformatics 62 (2006) 698-707.

[18] Colacino, S., G. Tiana, and G. Colombo, Similar folds with different stabilization mechanisms: The cases of prion and doppel proteins, BMC Struct. Biol. 6 (2006).

[19] Morra, G. and G. Colombo, Relationship between energy distribution and fold stability: Insights from molecular dynamics simulations of native and mutant proteins, Proteins: Struct. Funct. and Bioinf. 72 (2008) 660-672.

[20] Tiana, G., F. Simona, G.M.S. De Mori, R.A. Broglia, and G. Colombo, Understanding the determinants of stability and folding of small globular proteins from their energetics, Protein Science 13 (2004) 113-124.

[21] Larkin, M.A., G. Blackshields, N.P. Brown, R. Chenna, P.A. McGettigan, H. McWilliam, F. Valentin, I.M. Wallace, A. Wilm, R. Lopez, J.D. Thompson, T.J. Gibson, and D.G. Higgins, Clustalw and clustalx version 2., Bioinformatics 23 (2007) 2974-2948.

[22] Bastolla, U., M. Porto, H.E. Roman, and M. Vendruscolo, Principal eigenvector of contact matrices and hydrophobicity profiles in proteins, Proteins: structure, function and bioinformatics 58 (2005) 22-30.

[23] Klepeis, J.L., K. Lindorff-Larsen, R.O. Dror, and D.E. Shaw, Long-timescale molecular dynamics simulations of protein structure and function, Curr. Op. Struct. Biol. 19 (2009) 120-127.

[24] Shaw, D.E., M.M. Deneroff, R.O. Dror, J.S. Kuskin, R.H. Larson, J.K. Salmon, C. Young, B. Batson, K.J. Bowers, J.C. Chao, M.P. Eastwood, J. Gagliardo, J.P. Grossman, C.R. Ho, D.J. Ierardi, I. Kolossvary, J.L. Klepeis, T. Layman, C. McLeavey, M.A. Moraes, R. Mueller, E.C. Priest, Y.B. Shan, J. Spengler, M. Theobald, B. Towles, and S.C. Wang, Anton, a special-purpose machine for molecular dynamics simulation, Communications of the Acm 51 (2008) 91-97.

[25] van der Spoel, D., E. Lindahl, B. Hess, A.R. van Buuren, E. Apol, P.J. Meulenhoff, D.P. Tieleman, A.L.T.M. Sijbers, K.A. Feenstra, R. van Drunen, and H.J.C. Berendsen, Gromacs user manual version 3.2. Www.Gromacs.Org (Editor,2004)

[26] Hess, B., C. Kutzner, D. van der Spoel, and E. Lindahl, Gromacs 4: Algorithms for highly efficient, load-balanced, and scalable molecular simulation, J. Chem. Theory and Computation 4 (2008) 435-447.

[27] De Mori, G.M.S., G. Colombo, and C. Micheletti, Study of the villin headpiece folding dynamics by combining coarse-grained monte carlo evolution and all-atom molecular dynamics. , Proteins: Struct. Funct. and Bioinf. 58 (2005) 459-471

[28] Lindahl, E., B. Hess, and D. van der Spoel, Gromacs 3.0: A package for molecular simulation and trajectory analysis, J. Mol. Mod. 7 (2001) 306-317.

[29] van Gunsteren, W.F., X. Daura, and A.E. Mark, Gromos force field, Encyclopedia of Computational Chemistry 2 (1998) 1211-1216.

[30] Berendsen, H.J.C., J.R. Grigera, and P.R. Straatsma, The missing term in effective pair potentials, J. Phys. Chem. 91 (1987) 6269-6271.

[31] Hess, B., H. Bekker, J.G.E.M. Fraaije, and H.J.C. Berendsen, A linear constraint solver for molecular simulations, J.Comp.Chem. 18 (1997) 1463-1472.

[32] Berendsen, H.J.C., J.P.M. Postma, W.F. van Gunsteren, A. Di Nola, and J.R. Haak, Molecular dynamics with coupling to an external bath, J. Chem. Phys. 81 (1984) 3684-3690.

[33] Onufriev, A., D. Bashford, and D.A. Case, Modification of the generalized born model suitable for macromolecules., J. Phys. Chem. B. 104 (2000) 3712-3720.

[34] Weiser, J., P.S. Shenkin, and W.C. Still, Approximate atomic surfaces from linear combinations of pairwise overlaps (lcpo), J. Comput. Chem. 20 (1999) 217-230.

Figure captions:

Figure 1. a) Structure of the YAP 65 WW domain (Pdb entry 1k9r). The terminal segment depicted in green was not considered in the calculation. The protein regions corresponding to significant contributions to stability (SE components over threshold) are depicted as blue lines. b) Same as a) showing the peaks of the SE, residues E7 M8 A9 R17 Y18 F19 L20 with VdW spheres

Figure 2. Top, SE profile of the 8 native sequences mounted on the template structure. Bottom, SE profile of the WT template sequence.

Figure 3. Top, SE profile of the 8 Random sequences, Middle: SE profile of the 8 IC sequences, Bottom: SE profile of the 8 CC sequences.

Figure 4. CE profile calculated over residues 8-40 of the experimental structure 1k9r.

Figure 5. Pearson correlation coefficients of each sequence's SE with respect to the template SE (x axis) plotted versus the Pearson correlation coefficients of each sequence's SE with respect to the CE of the X ray template structure.

Caption for Table:

Similarity values between the Sequence Eigenvector of the Native structure of YAP65 WW domain. PDB id: 1k9r and each of the sequences tested in this paper. The similarity measure is based on the calculation of the Pearson's Coefficient. The sequences are reported in the Supp. Mat.

Supplementary Material.

The Supp. Mat. File reports the list and the alignment of the sequences analyzed.

TABLE 1. Similarity values between the Sequence Eigenvector of the Native structure of YAP65 WW domain. PDB id: 1k9r and each of the sequences tested in this paper. The similarity measure is based on the calculation of the Pearson's Coefficient. The sequences are reported in the Supp. Mat.

Sequence	Pearson's coeff SE-wtSE	Clustalw alignment score Wt	Pearson coeff. SE-CE
Wt	**1**	**-**	**0.744945**
N6	0.815371	39	0.913731
N7	0.553304	51	0.825885
N8	0.90147	51	0.799291
N11	0.821357	45	0.831807
N22	0.724998	24	0.89256
N28	0.883323	57	0.896717

N33	0.834896	42	0.914411
N40	0.708581	42	0.907512
Average N	**0.78**	**43.88**	**0.87**
CC8	0.708989	45	0.901651
CC13	0.794943	18	0.722429
CC14	0.794792	48	0.878284
CC18	0.62209	27	0.767779
CC22	0.669008	24	0.700183
CC24	0.645391	51	0.879911
CC43	0.558861	48	0.837438
CC45	0.934383	39	0.854095
Average CC	**0.72**	**37.5**	**0.82**
IC1	0.68257	42	0.931625
IC5	0.613809	36	0.805095
IC10	0.671559	36	0.909656
IC16	0.59906	27	0.808594
IC23	0.627277	30	0.884833
IC25	0.597115	45	0.86078
IC35	0.688846	48	0.925319
IC41	0.742172	41	0.878261
Average IC	**0.65**	**39**	**0.88**
R2	0.308312	6	0.774204
R5	0.405056	9	0.569796
R6	0.706564	9	0.894776
R8	0.595654	6	0.559788
R10	0.096114	12	0.530359
R13	0.71476	21	0.781025
R16	0.421356	9	0.512525

R19	0.390216	6	0.576323
Average R	**0.45**	**9.75**	**0.65**

Figure 1

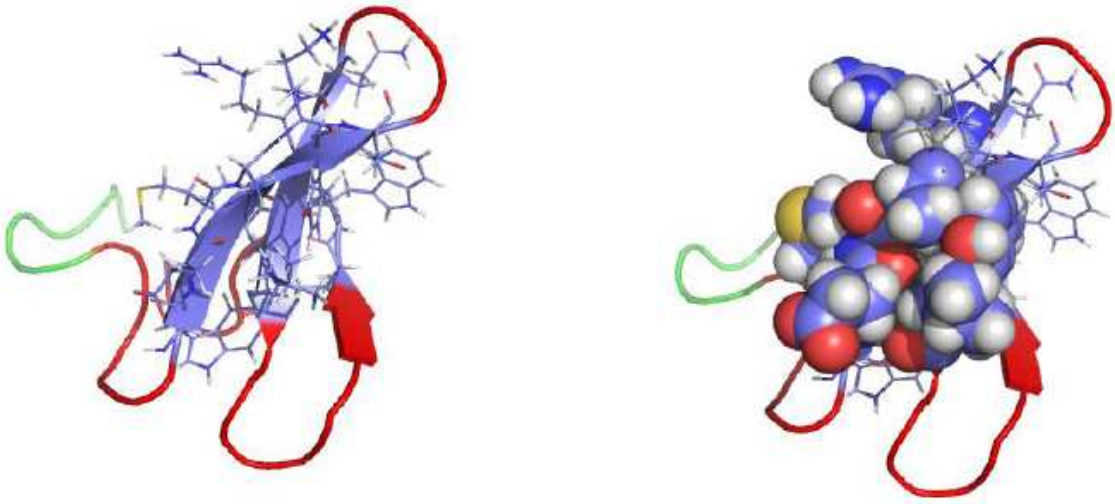


Figure 2

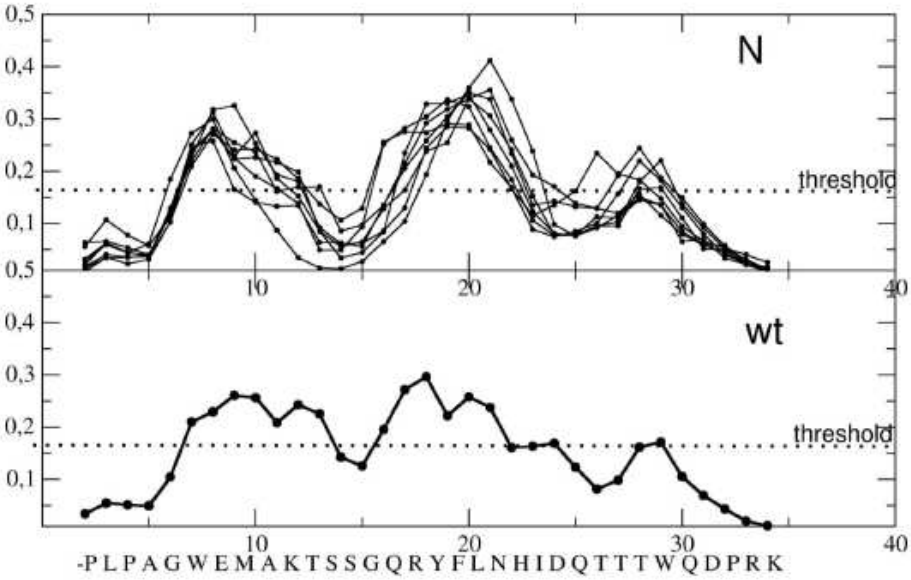


Figure 3

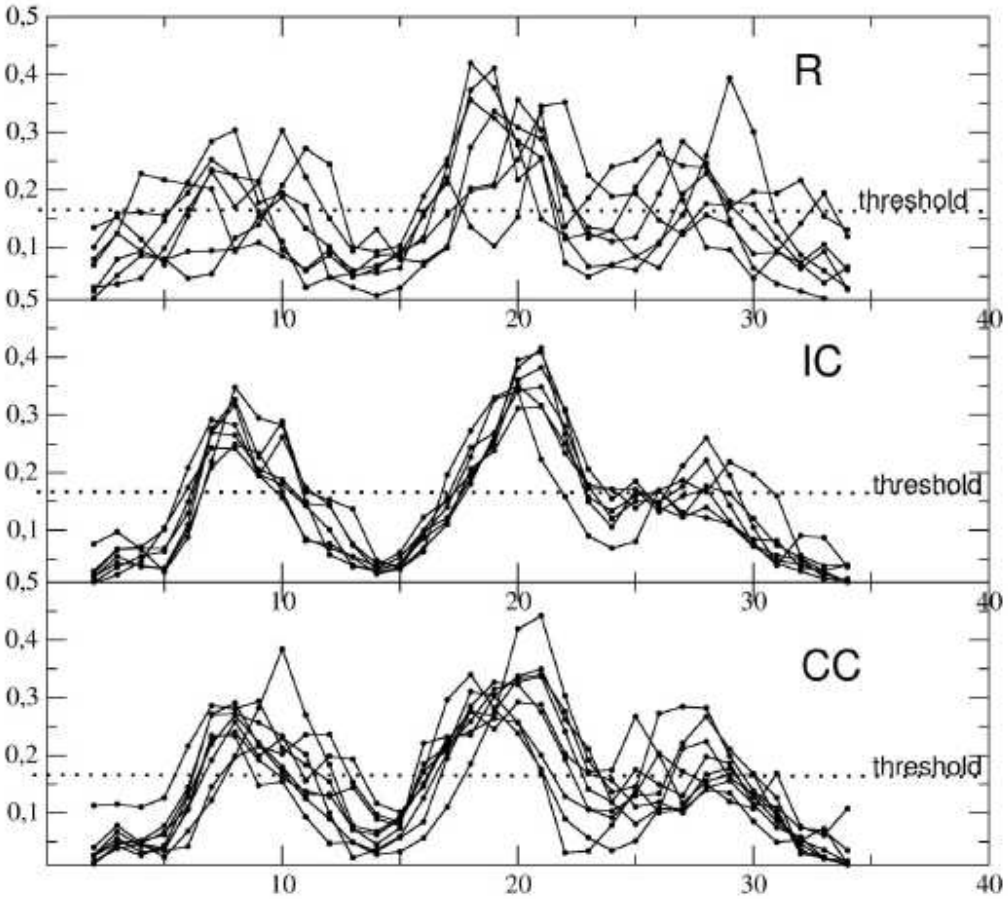


Figure 4

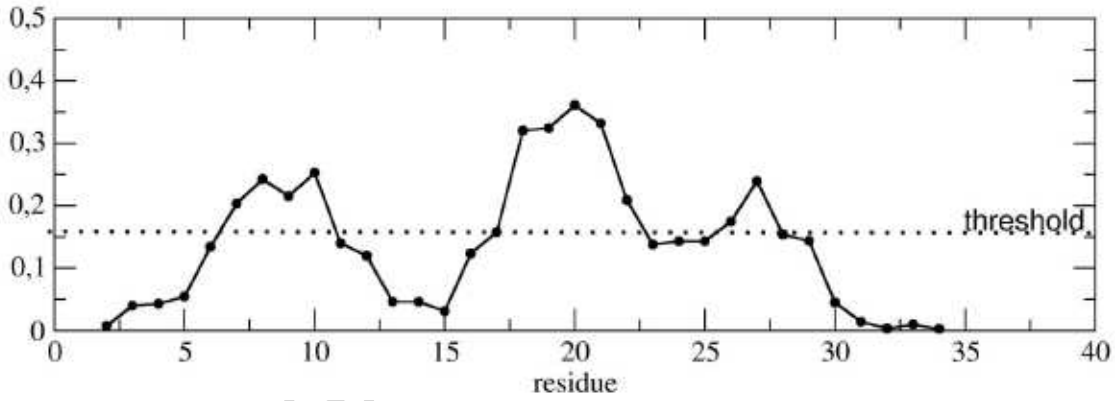


Figure 5