Published in final edited form as:

Methods Mol Biol. 2014 ; 1084: 239–254. doi:10.1007/978-1-62703-658-0_13.

Towards Comprehensive Analysis of Protein Family Quantitative Stability/Flexibility Relationships using Homology Models

Deeptak Verma¹, Jun-tao Guo¹, Donald J. Jacobs^{2,*}, and Dennis R. Livesay^{1,*}

¹Department of Bioinformatics and Genomics, University of North Carolina at Charlotte, Charlotte, NC 28223, USA

²Department of Physics and Optical Science, University of North Carolina at Charlotte, Charlotte, NC 28223, USA

Summary

The Distance Constraint Model (DCM) is a computational modeling scheme that uniquely integrates thermodynamic and mechanical descriptions of protein structure. As such, quantitative stability/flexibility relationships (QSFR) that describe the interrelationships of thermodynamics and mechanics can be quickly computed. Using comparative QSFR analyses, we have previously investigated these relationships across a small number of protein orthologs, ranging from 2 to a dozen [1,2]. However, our ultimate goal is provide a comprehensive analysis of whole protein families, which requires consideration of many more structures. To that end, we have developed homology modeling and assessment protocols so that we can robustly calculate QSFR properties for proteins without experimentally derived structures. The approach, which is presented here, starts from a large ensemble of potential homology models and uses a clustering algorithm to identify the best models, thus paving the way for a comprehensive QSFR analysis across hundreds of proteins in a protein family.

Keywords

Protein flexibility; Homology modeling; Distance Constraint Model; Quantitative stability/flexibility relationships

1. Introduction

Compared to sequence space, the relatively small number of quality x-ray crystal structures limits the ability to completely characterize protein flexibility properties across families and superfamilies. In the past we have performed comparative analyses of protein flexibility for a number of protein systems, include bacterial periplasmic binding homologs [1], oxidized thioredoxin [2] and β -lactamase protein families (in review). Thus far, our largest comparative flexibility study has focused on a dataset of twelve ortholog structures; however, a truly comprehensive analysis could require hundreds of structures to describe the flexibility properties of an entire protein family. Unfortunately, there is less than one hundred out of 3900 SCOP families that have 25 or more distinct orthologs with

*To whom correspondence should be addressed: djacobs1@uncc.edu and drlivesa@uncc.edu.

experimentally solved structures. As such, a large-scale flexibility analysis on the scale of dozens to 100+ structures will require homology models to fill-in these structure gaps (see Note 1).

Our previous analyses have shown that mDCM can detect variations in QSFR properties due to subtle structural perturbations introduced by single point mutations [3,4]. Hence, the key to reproduce accurate QSFR predictions will depend critically on having good homology models so that predicted differences can be trusted to be real. Herein, we present a protocol using human C-type lysozyme as an example that achieves this goal. Starting from 65 human lysozyme homology models constructed from 13 different templates, we use a clustering/filtering algorithm to identify a subset of the models that accurately reproduce the expected flexibility properties. The key difficulty of this problem is that good homology models must be identified *a priori* without comparisons to the actual structure, which is not available. As a first step toward quantifying homology model quality, we employ a clustering approach that segregates putative structures in terms of QSFR properties. Filtering on the QSFR properties that are most physically reasonable is then applied to screen out poor models, thereby boosting confidence levels in the quality of the remaining models. We test the approach by comparing clustered QSFR properties with those from “held back” real human structures. While *a priori* identifying the best cluster has not been implemented yet, we show statistically significant results that clearly indicate that homology model structures clustered based on structure similarity, thermodynamic and dynamic properties drastically improve predictions. Moreover, average QSFR quantities calculated over all the identified good homology models successfully reproduced x-ray structures’ average QSFR properties. Consequently, this is an important step towards a comprehensive QSFR analysis for hundreds of proteins.

2. Methods

2.1. A Brief Overview of the Distance Constraint Model

The Distance Constraint Model (DCM) is used for simultaneously calculating thermodynamic and mechanical properties of proteins. The DCM is based on a free energy decomposition scheme combined with constraint theory, such that microscopic interactions in the protein are represented as mechanical distance constraints [5,6]. Each distance constraint is associated with an enthalpic and entropic contribution. The microscopic interactions within the minimal DCM (mDCM) include: covalent bonds, hydrogen bonds and torsional-forces. Covalent bonds are quenched, whereas the other interactions fluctuate. Starting with a native protein structure, an ensemble of conformations is generated from the fluctuating constraints. However, complete enumeration of the partition function is impossible. As such, the mean field free energy of a macrostate, which is defined by the number of H-bond and torsion forces present, is computed using:

¹Within the DCM framework, flexibility and rigidity respectively quantify conformational diversity and regularity. These mechanical origins of flexibility and rigidity is linked to conformational entropy. These thermodynamic and mechanical measures combine to define QSFR. To be precise, the free energy of the protein can be expressed in terms of a global flexibility order parameter θ . This parameter defines the intrinsic flexibility of the protein and is equal to the average number of disordered torsion constraints divided by the total number of residues in a protein.

$$G(N_{hb}, N_{nt}) = U(N_{hb}) + u(N_{hb}^{\max} - N_{hb}) + v(N_{nt}) - T \{ S_{conf}(N_{hb}, N_{nt}, |\gamma, \delta_{nat}, \delta_{dis}) + S_{mix}(N_{hb}, N_{nt}) \}$$

where U is the intramolecular H-bond energy, u is an average H-bond energy to solvent, v is the energy of a native-like torsion angle, $S_c(N_{hb}, N_{nt})$ is the conformational entropy and $S_{mix}(N_{hb}, N_{nt})$ is the mixing entropy of the macrostate associated with the number of ways of distributing N_{nt} native-torsions and N_{hb} H-bonds within the protein. As a consequence of integrating mechanical and thermodynamic concepts, accurate flexibility characteristics of a given protein structure is calculated over an ensemble of possible constraint topologies that are appropriately thermodynamically weighted.

2.2. Homology Model Preparation

In this work, we focus on the ability of the mDCM to reproduce QSFR descriptions of human C-type lysozyme models. Starting with 13 different (non-human) lysozyme ortholog structures selected from SCOP [7], each is used to as a template for the human sequence. The 13 template structures have a wide range of sequence identity to the human lysozyme varying from 37.6% to 77.7%. MODELLER [8] is used to construct five models per template using otherwise default settings. Hydrogen atoms are added to the model structures and minimized followed by structure minimization using Amber99 force field. To ensure proper ionization, the H++ server [9] is used to add hydrogen atoms to the structures as expected at pH 2.7 based on calculated pK_a values. Other structural details are provided in Table 1. The same structure preparation is applied to seven human crystal structures, which are used to assess the quality of the model predictions.

2.3. Model Parameterization

Model parameter values $\{u, v, \delta_{nat}\}$ are determined by fitting to experimental heat capacity curves from differential scanning calorimetry (DSC) [5,6]. Once parameterized, the DCM can calculate a number of quantitative stability/flexibility relationship (QSFR) properties, which are thermodynamically averaged over the free energy basin. Each model and human x-ray structure is fit to the same human C_p curve obtained from differential scanning calorimetry [10]. This non-structural thermodynamic data provides empirical constraints that the mDCM leverages. As an example, best-fit curves for the five *rainbow trout* models are shown in Fig. 1. Other model structures exhibit similar heat capacity fit trends, although there are slight differences in parameters (cf. Table 1). Interestingly, the least squares fitting error is not correlated to homology model accuracy, highlighting the importance of other structural features that contribute towards prediction of accurate thermodynamic and mechanical features (see Note 2).

²The definition of a good homology model is difficult to describe, and fairly arbitrary within the field. Many scoring functions assess the quality of homology models based on statistical potentials and physics-based energy calculations [15,16]. Intrinsic errors that arise in physical properties caused by poor quality model structures are difficult to characterize and control, which are the important but problematic aspects addressed in this work.

2.4. Assessment

Instead of comparing each model structure to a single human structure that could introduce biases due to a particular conformational state, we compare the models to a background profile established from seven different human x-ray structures. We calculate average properties and define a range of likely values based on the observed fluctuations therein. That is, we are asking the question: “*When is an observed QSFR property within the range of expected values, and when is it not?*” This approach is the same as we established in an earlier work [4]. Any model QSFR metric within ± 1 standard deviation (i.e., $\pm 1\sigma$) of x-ray structures’ QSFR baseline is considered to be a “good prediction”, at a given residue position. A prediction value falling beyond $\pm 1\sigma$ defines “poor prediction” for that QSFR metric.

3. Results and discussion

3.1. The Quality of Homology Model QSFR Predictions

For each of the 65 models, Fig. 2A compares the percentage of residues within $\pm 1\sigma$ of the backbone flexibility profile to the sequence identity of the template used. Backbone flexibility is quantified by a Flexibility Index (FI), which quantifies how flexible (or rigid) backbone residues are (see Note 3). These comparisons show that better agreement between models and x-ray structures result in a better prediction of FI. Despite the overall positive relationship between model and x-ray structure similarity, there remains models that are false-positives and false-negatives. For example, the best human model prediction arises from *rainbow trout* (1LMN) resulting in highest QSFR prediction accuracy of 67%, while the accuracy of the other four 1LMN models are lower, down to 42%. Another important observation is the poor FI prediction by models derived from *house mouse* (1IVM), which is the only NMR structure template in our dataset (indicated by the blue dots). The sequence identity of this template is 78%, but the average prediction accuracy is approximately 30%, which highlight the large differences between NMR and x-ray structures. Across the whole dataset, only eight models have accuracies better than 60%, and perhaps more critically, those models come from several different templates.

The situation is similar when comparing the models to the real (held-back) structures. Structures with more similar hydrogen bond networks (**panel B**), TM-Scores (**panel C**), and overall RMSD (**panel D**) also enrich good FI predictions. Unfortunately, the number of structurally similar models giving poor FI predictions is still too large. These results are not unexpected since subtle changes in the model structures can cause drastic differences in hydrogen bond interactions and strength. Note that the upper boundary in Fig. 2 (accuracy = 82%) is defined by the best result obtained by comparing each of the original x-ray structures to the average x-ray structures’ backbone profile. That is, 82% is the highest similarity between any two of the seven x-ray structures.

³A frequently used QSFR metric in our analyses is the flexibility index (FI) that quantifies the degree to which a given residue deviates from being isostatic. An isostatic residue is marginally rigid, meaning there are just enough mechanical constraints due to intermolecular interactions to counter act its number of degrees of freedom to keep it rigid. Positive values of the FI represent the number of excess degrees of freedom (DOF) per rotatable dihedral angle within covalent bonds within flexible regions. Negative values of the FI represent the excess number of constraints per covalent bond within a rigid region.

The above results clearly suggest that a comprehensive QSFR analysis is possible; however, this is only possible if we can filter out poor models to improve prediction correctness by boosting statistics. For example, we can use model quality assessment score that do not require comparisons to known structures to improve identification of the best models [11]. To that end, we employ QMEAN [12,13], which evaluates models based on a secondary structure interaction potential, degree of solvent exposure and other structural quantities. Fig. 3 plots prediction accuracy versus the QMEAN scores, which indicates further enrichment. The vertical green line identifies an arbitrary threshold of good QMEAN scores. While the best scoring models are above the threshold, there are an unacceptable number of models with poor accuracies as well.

3.2. Expectation Maximization Clustering

The above results indicate that good flexibility predictions using models are possible; however, many models give unsatisfactory results. Moreover, there appears to be no systematic organization of the good models based on model quality or sequence/structure similarity to the target. Therefore, we employ a new clustering/filtering procedure over QSFR data consisting of a large number of heterogeneous metric types. Additionally, structure quality, sequence identity and other thermodynamic information are considered. Taken together, the key assumption of this strategy is that models with similar properties would tend to cluster together.

Clustering is done in two steps, both based on the expectation maximization (EM) algorithm (see Note 4). The first step focuses on non-QSFR properties discussed above (i.e., percent sequence identity between and QMEAN structure quality score), plus QSFR quantities that characterize the thermodynamic properties related to structure quality (i.e., the free energy barrier height between the native and unfolded basins, and the global flexibility of the native, transition and unfolded states). EM identifies three clusters, whose mean and standard deviation are summarized in Table 2.

Fig. 4 plots the QMEAN scores versus θ_{nat} , which describes the global flexibility of the native structure. The three clusters are color-coded. The cluster-2 models (colored red) have the best average QMEAN score and also the smallest variation therein (0.68 ± 0.03). As such, it was selected for further analysis, resulting in 23 models the advanced to the second round of clustering.

The second step starts by calculating an all-to-all correlation for each QSFR metric. For example, the correlation between all FI vectors for all $23 \times 22/2 = 253$ pairs. This process is repeated for all of the other QSFR quantities as well (see Note 5). The whole set of QSFR correlation coefficients, plus pairwise structure similarities calculated using RMSD and TM-scores [14], are again clustered using the EM algorithm. Fig. 5 shows the resultant clusters from one view of the data, where TM-scores are plotted versus the probability of a backbone

⁴Expectation maximization (EM) assigns a probability distribution to every data instance, which defines the probability of it belonging to each of the clusters. The algorithm can create its own clusters and does not require *a priori* information regarding the expected number of clusters. To find the optimum number of clusters the EM algorithm cross validates and calculates the average log-likelihood. Starting with one cluster, the numbers of clusters are increased if the average log-likelihood continuously increases at each step.

torsion angle to rotate. Cluster-9 (shown in red) has the highest average TM-score, and also the QSFR quantities are well conserved therein. As demonstrated next, these 18 models constitute a significant enrichment of models that reproduce the flexibility profiles of the known x-ray structures. Fig. 6 summarizes the clustering workflow described above.

3.4. Model Enrichment by Clustering

To compare how well the above cluster of models performs, we compare the average accuracy therein with how well we could have done without clustering. That is, we could have simply used QMEAN to identify the best model, or some set of best models. We compare the accuracy of the QSFR descriptions provided by the EM cluster to: (1) the best QMEAN model and (2) to the group of five best QMEAN models. As before, accuracy is determined by comparison to the profile developed from the seven human x-ray structures.

Fig. 7 plots the results of the three different model sets for eleven different QSFR metrics describing backbone properties. Interestingly, the single-best QMEAN model fails to give an accurate prediction in most cases, but averaging over five best QMEAN models significantly improves accuracy. In all cases, averaging over the various EM clusters further improves accuracy, by as much as 20% in some instances. Taken together, these results suggest that clustering significantly improves the description of backbone flexibility profiles using homology models. Excitingly, note that the EM-set averaged FI accuracy is 78%, which is very close to the 82% upper bound identified from the seven x-ray structures (see Note 6).

Fig. 8 compares the five cooperativity correlation plots for QMEAN-5 and EM sets, which are compared to the real plots derived by averaging over the seven x-ray structures. Both sets do a reasonable job of reproducing the experimental structures, but the EM set appears more similar. A key problem with this visual analysis is it neglects variation within real structures. As such, we make profile comparisons, although this time each pixel data point represents a residue pair (vs. a single residue). These results are shown in Fig. 9, which confirms EM clustering improves prediction accuracy in all cases. The accuracy in all of the EM clusters is greater than 70% accuracy, whereas none of the QMEAN-5 sets reach that level. A scatter plot of CC metric values provides insight on correlation distribution (Fig. 10). The EM set is clearly much sharper and more accurate than the QMEAN-5 set.

3.5. Concluding Remarks

The presented method clearly demonstrates that EM clustering based on the model output and structural details leads to an enrichment of models able to reproduce the flexibility profiles of the real structures. The sole remaining point is to determine how to select the final cluster. In the presented work, the second round of clustering identified ten clusters, and the final (Cluster-9) was chosen based on that set best reproduced the x-ray profile. That

⁵In addition to the flexibility index, FI, the mDCM calculates a number of other backbone flexibility quantities, including: (a) mechanical susceptibility, which quantifies the fluctuations within a particular residue to be rigid or flexible over the ensemble of constraint topologies; (b) the density of independent DOF; (c) the probability of a backbone torsion angle to rotate (cf. Fig. 7). The mDCM also calculates five Cooperativity Correlation (CC) metrics that quantify different types of residue-residue couplings. For example, the flexibility cooperativity correlation identifies residues pairs that are either co-rigid, flexibly correlated, or have no mechanical coupling. Examples of the five CC metrics are provided in Fig. 8.

⁶The x-ray structure upper bound varies for each of the quantities, ranging from 76% to 98%.

is, we knew the correct answer and picked the solution that was closest. While there were some indicators that this cluster was the best (e.g., highest average intra-cluster TM score), we have not yet determined a robust method to identify the best cluster. Nevertheless, the presented method represents a substantial improvement in the ability to use homology models to describe protein flexibility properties, thus making our goal of a comprehensive analysis of a 100+ proteins with a given family that much closer.

Acknowledgments

This work has been partially supported by NIH R01 GM073082, S10 SRR026514, and the Department of Bioinformatics and Genomics. Key to the distance constraint model is the use of graph-rigidity algorithms, claimed in U.S. Patent 6,014,449, which has been assigned to the Board of Trustees Michigan State University. Used with permission.

References

1. Livesay DR, Huynh DH, Dallakyan S, Jacobs DJ. Hydrogen bond networks determine emergent mechanical and thermodynamic properties across a protein family. *Chem Cent J*. 2008; 2:17. [PubMed: 18700034]
2. Mottonen JM, Xu M, Jacobs DJ, Livesay DR. Unifying mechanical and thermodynamic descriptions across the thioredoxin protein family. *Proteins*. 2009; 75(3):610–627. [PubMed: 19004018]
3. Verma D, Jacobs DJ, Livesay DR. Predicting the melting point of human C-type lysozyme mutants. *Curr Protein Pept Sci*. 2010; 11(7):562–572. [PubMed: 20887260]
4. Verma D, Jacobs DJ, Livesay DR. Changes in Lysozyme Flexibility upon Mutation Are Frequent, Large and Long-Ranged. *PLoS Comput Biol*. 2012; 8(3):e1002409. [PubMed: 22396637]
5. Jacobs DJ, Dallakyan S. Elucidating protein thermodynamics from the three-dimensional structure of the native state using network rigidity. *Biophys J*. 2005; 88(2):903–915. [PubMed: 15542549]
6. Livesay DR, Dallakyan S, Wood GG, Jacobs DJ. A flexible approach for understanding protein stability. *FEBS Lett*. 2004; 576(3):468–476. [PubMed: 15498582]
7. Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *Journal of Molecular Biology*. 1995; 247(4):536–540. [PubMed: 7723011]
8. Sanchez R, Sali A. Evaluation of comparative protein structure modeling by MODELLER-3. *Proteins*. 1997; (Suppl 1):50–58. [PubMed: 9485495]
9. Gordon JC, Myers JB, Folta T, Shoja V, Heath LS, Onufriev A. H++: a server for estimating pKas and adding missing hydrogens to macromolecules. *Nucleic Acids Res*. 2005; 33(Web Server issue):W368–371. [PubMed: 15980491]
10. Takano K, Yamagata Y, Fujii S, Yutani K. Contribution of the hydrophobic effect to the stability of human lysozyme: calorimetric studies and X-ray structural analyses of the nine valine to alanine mutants. *Biochemistry*. 1997; 36(4):688–698. [PubMed: 9020766]
11. Cozzetto D, Kryshtafovych A, Fidelis K, Moulton J, Rost B, Tramontano A. Evaluation of template-based models in CASP8 with standard measures. *Proteins*. 2009; 77(Suppl 9):18–28. [PubMed: 19731382]
12. Benkert P, Kunzli M, Schwede T. QMEAN server for protein model quality estimation. *Nucleic Acids Res*. 2009; 37(Web Server issue):W510–514. [PubMed: 19429685]
13. Benkert P, Schwede T, Tosatto SC. QMEANclust: estimation of protein model quality by combining a composite scoring function with structural density information. *BMC structural biology*. 2009; 9:35. [PubMed: 19457232]
14. Xu J, Zhang Y. How significant is a protein structure similarity with TM-score = 0.5? *Bioinformatics (Oxford, England)*. 2010; 26(7):889–895.
15. Gniewek P, Leelananda SP, Kolinski A, Jernigan RL, Kloczkowski A. Multibody coarse-grained potentials for native structure recognition and quality assessment of protein models. *Proteins*. 2011; 79(6):1923–1929. [PubMed: 21560165]

16. Gopal SM, Klenin K, Wenzel W. Template-free protein structure prediction and quality assessment with an all-atom free-energy model. *Proteins*. 2009; 77(2):330–341. [PubMed: 19422063]

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Fig. 1. Five homology models constructed from *rainbow trout* template are fit to human lysozyme heat capacity curve. The curve represented by red-dots describes the data from DSC experiment, and the five dashed-blue curves show typical fits.

Fig. 2.

Comparison of the predicted flexibility index accuracy for all 65 homology models against (a) pairwise sequence identity, (b) hydrogen bond network similarity, (c) TM-scores and (d) structure RMSD. The latter properties that depend on structure are averaged over 7 known x-ray crystal structures. Model structures in close agreement with x-ray structures reproduce the flexibility index with higher accuracy, although false-negatives and false-positives are also present. Blue data points represent NMR structures. Note that 82% is the best score that occurred when comparing a QSFR property associated with each of the original x-ray structures against the average QSFR property taken over all 7 structures.

Fig. 3.

Comparison of FI percent accuracy prediction. Models with high quality may not necessarily result in higher FI accuracy. Blue data points represent NMR structures. The green dashed line represents an arbitrary QMEAN threshold value at 0.7.

Fig. 4. Clustering using structural and thermodynamic quantities. Clusters are represented by different colors. Models clustered in red (cluster-2) are selected for further analyses.

Fig. 5. Results obtained from a second round of EM clustering using structural and mechanical quantities. Clusters are represented by different colors. Model pairs clustered in red (cluster-9) are the final filtered models.

Fig. 6.
Expectation maximization clustering workflow to filter best homology models.

Fig. 7. Comparison of protein backbone QSFR metric accuracy levels for different model sets. Average QSFR quantities obtained from EM 18 best filtered models have higher agreement with x-ray structures' QSFR quantities as compared to QMEAN best and 5-best models.

Fig. 8. Comparison of residue-residue coupled QSFR metrics. A good qualitative resemblance is observed between EM 18-best models and original x-ray structures' average QSFR cooperativity metrics.

Fig. 9. Comparison of residue-residue coupled QSFR metric accuracy levels for different model sets. Average QSFR quantities obtained from EM 18 best filtered models have better precision levels.

Fig. 10.

(a and c) Flexibility cooperativity correlation (CC) and (b and d) CC_{IDF} values are compared against original x-ray structures. EM best 18 filtered models can reproduce much more accurate CC and CC_{IDF} properties of original x-ray structures as compared to QMEAN's best 5 models. QMEAN models' CC comparison with original x-ray structures exhibits wider data distribution, whereas EM models have a narrow distribution. Other metrics show similar trends. The black line shown is the best-fit regression across data points. The histograms are constructed by binning data points with equal intervals on y-axis on either side of the regression line. The interval size for binning is consistent across both, CC and CC_{IDF} plots, respectively.

Table 1

Structural template used to construct the human lysozyme homology models. Five models are built from each template. The average and percent variation in $\{u, v, \delta_{\text{hnt}}\}$ of homology models from each template are also reported.

Organism	Template PDB	Resolution (Å)	R-value	u	v	δ_{hnt}
Turkey	135L	1.30	0.189	-1.58±(13.6%)	-0.38±(69.7%)	0.86±(75.7%)
Northern bobwhite	1DKJ	2.00	0.177	-2.01±(26.2%)	-0.60±(30.0%)	1.00±(48.3%)
Domestic silkworm	1GD6	2.50	0.181	-1.85±(7.6%)	-0.50±(15.2%)	0.96±(29.9%)
Chicken	1HEL	1.70	0.152	-1.72±(13.7%)	-0.60±(18.5%)	0.50±(53.3%)
Helmeted guineafowl	1HHL	1.90	0.170	-1.69±(8.2%)	-0.40±(51.3%)	0.86±(40.8%)
Tasar silkworm	1IIZ	2.40	0.231	-2.12±(13.4%)	-0.73±(18.1%)	0.96±(44.8%)
House mouse	1IVM	-	-	-2.38±(8.4%)	-0.72±(12.8%)	1.14±(11.8%)
Ring-necked pheasant	1JHL	2.40	0.214	-1.90±(24.5%)	-0.61±(19.1%)	0.75±(54.0%)
Echidna	1JUG	1.90	0.170	-1.85±(8.7%)	-0.33±(44.0%)	1.04±(38.7%)
Rainbow trout	1LMN	1.80	0.174	-1.90±(9.7%)	-0.44±(30.0%)	1.07±(29.3%)
Dog	1QQY	1.85	0.178	-1.92±(5.4%)	-0.50±(40.1%)	0.89±(30.2%)
Horse	2EQL	2.50	0.234	-1.81±(18.0%)	-0.64±(24.8%)	0.82±(69.4%)
Japanese quail	2IHL	1.40	0.165	-1.67±(12.6%)	-0.41±(44.9%)	0.96±(59.5%)

Table 2

Clustering results from structure and thermodynamic quantities. The values represent average $\pm$ standard deviation of data points for the given quantity belonging to a defined cluster. Cluster-2 with best average QMEAN score with least standard deviation is selected. θ_{nat} , θ_{trans} and θ_{dis} correspond to protein's intrinsic flexibility in native, transition and disordered states respectively.

	Cluster-1	Cluster-2	Cluster-3
Folding free energy	2.29 $\pm$ 0.84	3.57 $\pm$ 1.23	2.79 $\pm$ 1.05
Unfolding free energy	1.84 $\pm$ 0.73	3.12 $\pm$ 1.09	2.33 $\pm$ 0.96
θ_{nat}	0.75 $\pm$ 0.11	0.92 $\pm$ 0.14	0.76 $\pm$ 0.13
θ_{trans}	1.06 $\pm$ 0.15	1.31 $\pm$ 0.20	1.10 $\pm$ 0.21
θ_{dis}	1.70 $\pm$ 0.22	2.07 $\pm$ 0.16	1.70 $\pm$ 0.30
QMEAN scores	0.67 $\pm$ 0.04	0.68 $\pm$ 0.03	0.54 $\pm$ 0.04
Sequence identity	0.60 $\pm$ 0.09	0.58 $\pm$ 0.05	0.39 $\pm$ 0.01