A Computational Genome-wide Study of Protein Folding Rate

BY

SANDEEP C. GORLA
B.S. Equivalent, National Institute of Technology Durgapur, 2010

THESIS

Submitted as partial fulfillment of the requirements
for the degree of Master of Science in Bioinformatics
in the Graduate College of the
University of Illinois at Chicago, 2012

Chicago, Illinois

Defense Committee:

Jie Liang, Chair and Advisor
Yang Dai
Anjum Ansari, Physics

This thesis is dedicated to my family, whose unconditional love and support motivated me and made this possible.

ACKNOWLEDGEMENTS

I wish to thank my advisor Prof. Jie Liang from UIC for his invaluable guidance and immeasurable patience in assisting me with my research. Many thanks also go to Prof. Yang Dai and Prof. Anjum Ansari for taking time out of their busy schedules to serve on my committee.

A whole lot of thanks to past and present group members, David Jimenez Morales, Ke Tang, Hammad Naveed, Gamze Gursoy, Meishan Lin, Yingzi Li, Yun Xu, Michael Montesano, Volga Pasupuleti, Jieling Zhao, Marco Maggioni, Youfang Cao, Larisa Adamian, and Joe Dundas.

A special thanks to my friend Meghana for her invaluable suggestions made towards the writing of this thesis report.

SCG

Table of Contents

LIST OF FIGURES

LIST OF TABLES

SUMMARY

Proteins are the workhorses of living organisms and are essential in carrying out a multitude of complex processes that are critical for the survival of a cell. Proteins should be able to 'fold' to a native three-dimensional structure in a biologically relevant time to be functional. In our current study, we tried to gain insight into this complex process of protein folding by studying the folding rate distributions in a proteome.

We used a folding rate prediction model based on the evaluation of the contribution of pairs of sequentially neighboring amino acids to the folding rate. We used this model to compute the proteome-wide distributions of folding rates of several model organisms. We looked at the extreme limits of these distributions and see that many slow folding proteins in a cell require the help of the chaperone machinery to fold in a biologically relevant time. On analyzing the folding rate distributions in extremophiles we see that these halophiles have a different folding rate distribution when compared to mesophiles. We see a small increase in the fraction of the slowest folding proteins in halophiles when compared to mesophiles. We further identified the fastest and slowest folders among representative sequences of all known protein families. We analyzed certain structures among these fast/slow folders to gain valuable insight into the relationship between structure and the folding rate of a protein.

1 INTRODUCTION

1.1 Protein folding

Protein folding is the physical process by which a protein assumes it's highly structured functional conformation starting from a denatured state. This is one of the most fundamental processes in a living cell. The messenger RNA (mRNA) produced by transcription is translated by the ribosome into a polypeptide, which initially doesn't have a developed three-dimensional structure. This random coil of amino acids assembles itself into a specific three-dimensional shape (native state), which is essential to perform the correct biological function. The folding of a protein often occurs co-translationally. This means that the N-terminal of the polypeptide chain starts to fold as the rest of the protein is being synthesized. Misfolded proteins are the root cause of many neurodegenerative diseases. Anfinsen in 1961 [1] showed that a protein could spontaneously refold from denatured states in a test tube in the absence of any cellular environmental factors. This essentially means that "sequence information" alone determines the folding process. However, it is important to note that some proteins need the assistance of chaperones to fold to the correct structure. This doesn't mean that Anfinsen's Dogma is false because chaperones don't provide any information for folding; rather they prevent unproductive side-reactions like aggregation of several protein molecules.

Since Anfinsen, many researchers have worked on "the protein folding problem". The protein folding problem has two important facets: (1) the mechanism by which a protein folds to its native conformation; (2) How the amino acid sequence of a protein determines its structure? Protein folding mechanisms are usually studied via free energy landscapes

from simulations that are tested by experiments. Currently, the following mechanisms of protein folding are widely accepted. (1) Hydrophobic collapse model: characterized by a initial nucleus formation; followed by formation of secondary structure, and finally the coalescence of secondary structural elements into native three-dimensional structure (2) Nucleation-condensation model: characterized by concurrent formation of secondary and tertiary structural elements; (3) Zipping and assembly model: emphasizing zipper-like folding mechanism; (4) Funnel model: emphasizing folding as involving parallel pathways forming a funnel-shaped energy landscape rather than a single microscopic pathway.

1.2 Protein folding kinetics

Levinthal, in 1969 [41], noted that for a protein to fold by sequentially sampling the vast possible conformational space, it would take a time longer than the age of universe. How then are proteins able to converge to their native states so fast? Understanding this relationship between a protein sequence and its folding rate is a fundamental and challenging problem. Folding speeds of different proteins vary significantly. Small proteins usually fold faster with simple 2-state kinetics. They have no visible intermediates in the course of folding. Larger proteins generally fold at a lower rate via a 3-state folding kinetics and metastable intermediates are often observed.

Studying protein folding rates is very important because of the following reasons: (1) It gives us a valuable insight into the protein folding mechanism. (2) Understanding protein folding kinetics helps us design proteins with desired folding speeds. (3) Protein aggregation is directly related to the rate of protein folding. The failure of a protein to

fold properly is the root cause of a range of diseases like Cystic Fibrosis and Alzheimer's disease.

Protein folding rates can be calculated and analyzed using a range of biochemical experiments [2-7]. These methods are both expensive and time consuming. This coupled with the rapidly increasing gap between newly discovered sequences and available experimental folding rate data requires a fast and efficient computational model to predict protein folding rates.

1.3 Protein folding rate prediction models

Plaxco et al. [8] in 1998 made the important observation that the average relative contact order (RCO), a measure of relative fraction of local vs. non-local non-covalent contacts, correlates well with the folding rates of two-state folding proteins. Subsequently, many variations of this idea have been studied, indicating that folding rates also correlate with long-range order (LRO) [9], the effective contact order (ECO) [10], the total contact distance (TCD) [11], a chain topology parameter (CTP) [12] and the effective length of protein, L_{eff} [13]. However, these results were obtained with a relatively small data set and often require the knowledge of the native structure of the protein. Ouyang and Liang [14], using a large set of both single-state and multi-state folders, showed that folding rates correlate well with the number of residues that form geometric contacts. Other sequence-based methods [15-19] have been proposed recently. These methods do not consider the influence of sequence order on the protein folding rate. Xu et al. [20] used amino acid sequence order to derive a method, based on an extended version of pseudo-amino acid composition, for predicting protein folding rates.

In our current analysis, we used a novel, highly accurate folding rate prediction model proposed by Ouyang [21]. This model was developed by evaluating the contribution of amino acid pairs in a sequence to the folding rate. In this model an optimization procedure based on simulated annealing was used to divide all possible nearest neighbor pairs (NNPs) of residues in a protein into three clusters, each with an optimized weight parameter, according to the contribution to the folding speed. This method predicts the folding rates with the greatest accuracy when compared with 9 other methods [11, 22-25, 49-51].

1.4 Folding rate distribution of a cell

It is important to look at the folding rate distribution of a proteome as to see what the biological implications of folding speed are. A nascent polypeptide chain should be able to fold in a biologically relevant time to be functional. In the same way we wish to see if there is an upper limit to the folding speed. These limits represent one of the most important constraints on a cell because cell growth is limited by the folding rates of its slowest folding proteins [42].

We use our prediction model to compute folding rates of all proteins with lengths between 25 to 350 amino acids from the proteomes of *E. coli, B. subtilis, S. cerevisiae, C. elegans, D. melanogaster, A. thaliana, M. musculus* and *H. sapiens*. We use only the protein sequences of lengths between 25 to 350 amino acids because the training set used in building our prediction model contains proteins only from this range; and hence we believe that our prediction model works in this range with the highest confidence. We compare and analyze the distributions of the folding rates among these different

mesophilic model organisms. We then analyze the predicted folding rate distributions of two halophiles (*H. salinarum* and *S. ruber*), two psychrophiles (*C. psychrerythraea* and *F. psychrophilum*), and four thermophiles (*A. fulgidus, M. jannaschii, P. abyssi,* and *P. horikoshii*); to see if there is a differential distribution in extremophiles compared to mesophilic model organisms. We also list and analyze the fastest and slowest folders among representative sequences of all known protein families. While doing this analysis we show how sequence order gives different and more accurate folding rate values when compared to using amino acid composition alone. We also look at the PDB structures of these special families to gain insight into the structure-folding rate relationship. For doing all the above comparisons we need to do a length normalization as the size of a protein does affect the folding rate of a protein [26-31]. Once normalized, we can freely compare the folding rates of proteins having different lengths.

2 DATASET

2.1 Proteomic data

For comparing proteome-level protein folding rate distributions we have downloaded all protein sequences between the lengths of 25 and 350 amino acids from the proteomes of *E. coli, B. subtilis, C. elegans, D. melanogaster, S. cerevisiae, A. thaliana, M. musculus, H. sapiens, H. salinarum*, *S. ruber*, *C. psychrerythraea* , *F. psychrophilum*, *A. fulgidus, M. jannaschii, P. abyssi,* and *P. horikoshii* from UniprotKB [32].

2.2 Data on representative Sequences of All Known Protein Families

To compare folding rates of proteins at the family level, we have downloaded sequences between the lengths of 25 and 350 amino acids which represent each family in SCOP [33] from the ASTRAL [34] database.

3 PREDICTING PROTEIN FOLDING RATES FROM SEQUENCE

3.1 Summary

It is important to be able to predict folding rates from amino acid sequences alone due to the constraints on the available structure data of proteins. Devising a reliable model that can predict folding rates from sequence information alone will help us to utilize the large amount of sequence information available on public databases and also to do proteome-level analysis. There are several works describing the prediction of folding rates from primary sequence [11,22-25,49-51]. However, these methods require at least some sequence information (for e.g. structural class) or use amino-acid compositions alone; without emphasizing on the "sequence order". The position of each residue in sequence is also crucial for folding a chain to it's native three dimensional structure. Therefore, the inclusion of sequence order information in the prediction model should give us a better prediction. Here, we use the model proposed by Ouyang [21] where he evaluated the contribution of amino acid pairs in sequence to protein folding speed and developed a new folding rate prediction method with much higher accuracy.

3.2 Ouyang's folding rate prediction model

A brief description of the model is as follows. First, a nearest neighbor pair (NNP) is defined as a pair of residues, which are nearest neighbor of each other in protein sequence. Since there are total 20 amino acid types, the total number of possible NNP types is

$20\times20/2+10 = 210$ if we consider A-B and B-A as the same NNP. Then, 210 NNPs will be clustered into three groups according to the contribution to folding speed. Simulated annealing algorithm is employed to find optimal clusters. In summary, at each step, the algorithm heuristic considers some neighbor solutions of the current clustering, and probabilistically decides between moving the system to one of neighbor solution or staying in current state. The probabilities are chosen so that the system ultimately tends to move to states of lower energy state (better clustering). Typically this step is repeated until the system reaches a state that is good enough for the application, or until a given computation budget has been exhausted.

In this model, the optimization procedure includes following steps: (1) randomly divide 210 NNPs into three clusters, each cluster initially has 70 NNPs. (2) optimize weights of three clusters by SVD with leave-one-out testing. If there is improvement of correlation coefficient r between calculated $\ln(k_f)$ and experimental data, the current cluster solution will be kept. Otherwise, the acceptance is specified by an acceptance probability function, which is depended on the difference between current r and previous r and current temperature. (3) If the current solution is accepted then new clusters will be created from current ones. To efficient generate candidate clusters, three move sets are used. Each time, two of three clusters are randomly selected. Then 1, 3, or 5 NNPs will be selected from one cluster and move to another cluster. (4) Decrease temperature and repeat step 2 and 3 till reaching the lowest temperature limitation. The final outputs are three clusters of 210 NNPs and three optimized weight parameters. The same dataset is tested and the result is listed in Table 1.

Table 1: Optimized clustering solution of 210 NNP types

Cluster	Weight	Cluster Members (NNP Types)
Cluster 1	**0.2991**	AA AF AH AI AP AT AW CF CH CP CS CV DD DE DN DQ DS EF EK EL EM EQ ER ES FF FK FN FR FY GH GI GK GM GN GQ GR HK HM HR HY II IN IY KM KR KT LL LP LQ LS MQ MR MW NP NQ NR NY PP PS PW QW RT RV SW YY
Cluster 2	**-0.5136**	AC AE AQ AV CE CG CI CM CR CT CW CY DG DI DM DP DR DV DY EH EV EW EY FI FP FQ FS FV GS GT GV GW HH HI HL HP HQ HV HW IK KK KQ KV KY LT MP MS NS NT NV NW PR PY QT QY RS RW ST SV TW VV VW VY WW WY
Cluster 3	**-0.0656**	AD AG AK AL AM AN AR AS AY CC CD CK CL CN CQ DF DH DK DL DT DW EE EG EI EN EP ET FG FH FL FM FT FW GG GL GP GY HN HS HT IL IM IP IQ IR IS IT IV IW KL KN KP KS KW LM LN LR LV LW LY MM MN MT MV MY NN PQ PT PV QQ QR QS QV RR RY SS SY TT TV TY

The positive weight (0.2991) of cluster 1 denotes NNP types in this group can facilitate protein folding. On the other hand, cluster 2 contains NNP types, which may retard folding. And cluster 3 doesn't show significant effect. Consequently, we build a predictive model using following linear equation:

$$\ln\left(k_f\right) = 11.6418 + 0.2991N_{cluster1} - 0.5136N_{cluster2} - 0.0656N_{cluster3}$$

Here, *N* is the number of NNPs which belong to one of above three clusters. The excellent performance (r=0.98) of leave-one-out test shows that this model can be used as accurate protein folding prediction (Figure 1). We further collected 10 folding rate data which are not used in building model. The real prediction result also shows high correlation (r=0.95 Figure 2) between predicted data and experimental data. The average difference of all test cases is only 0.73 unit (Table 2).

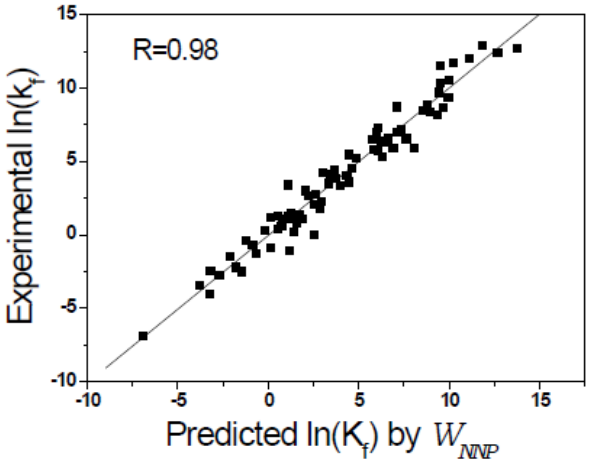


Figure 1: Jack-knife test of the predicted folding rate from weighted NNP vs. experimentally measured values

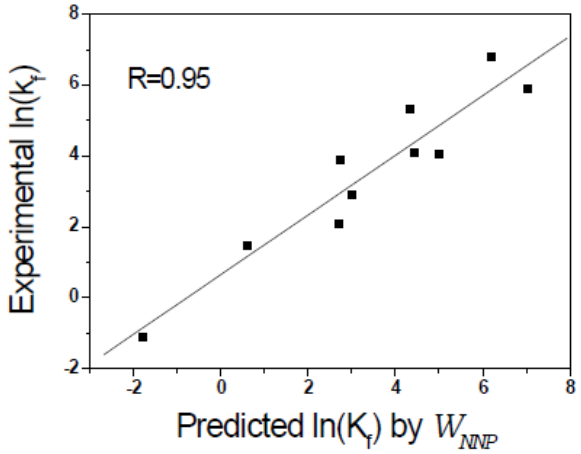


Figure 2: Folding rate prediction results of 10 test cases

Table 2: List of predicted and experimental folding rates

PDB	Predicted	Exp.
1AEY	2.704	2.09
1AYE	6.187	6.79
1COA	2.759	3.87
1D6O	0.595	1.45
1HNG	2.994	2.89
1HZ6	4.450	4.10
1PNJ	-1.789	-1.1
1RIS	7.044	5.89
1SRL	4.992	4.04
3MEF	4.338	5.30

Compared with other methods [11, 22-25, 49-51] (Table 3), the empirical relationships derived for weighted NNP predict the folding rates with greatest accuracy. We systematically analyze the sequence information which can be used for folding and find both amino acid composition and relative residue position in sequence determines protein folding kinetics.

Table 3: Comparison of protein folding rate predictive ability of 9 methods

Method	**Parameter**	**Used information**	**Size of testing dataset**	**R**	**Reference**
linear single regression	contact order	3D structure	32	0.74	Plaxco et al.[8]
linear single regression	long-range order	3D structure	32	0.81	Gromiha and Selvaraj[23]
linear single regression	total contact distance	3D structure	32	0.88	Zhou and Zhou[11]
linear multiple regression	secondary structure content	3D structure	32	0.91	Gong et al.[24]
linear multiple regression	amino acid properties	1D sequence + structure class info.	32	0.97	Gromiha[25]
linear single regression	Composition Index	Sequence only	62	0.72	Ma et al.[49]
linear multiple regression	amino acid properties	Sequence only	77	0.96	Gromiha et al.[50]
quadratic response surface model	amino acid properties	Sequence only	77	0.9	Huang and Gromiha[51]
linear multiple regression	***Weighted NNP***	***Sequence only***	***80***	***0.98***	

4 LENGTH CORRECTION FOR THE MODEL

We need to correct for length in our model in order to compare folding rates of proteins having different lengths. We plotted the predicted folding rates of all organisms used in this study against their lengths and fitted the data to several simple equations. We found that the simple linear equation

$$\textit{Predicted Folding Rate} = 11.08 - 0.08 \times (\textit{Length})$$

fits the data with an r- square value of 0.8114 and standard error of 4.156 (Fig. 3). We use this equation for length normalization of predicted folding rates in our further analysis.

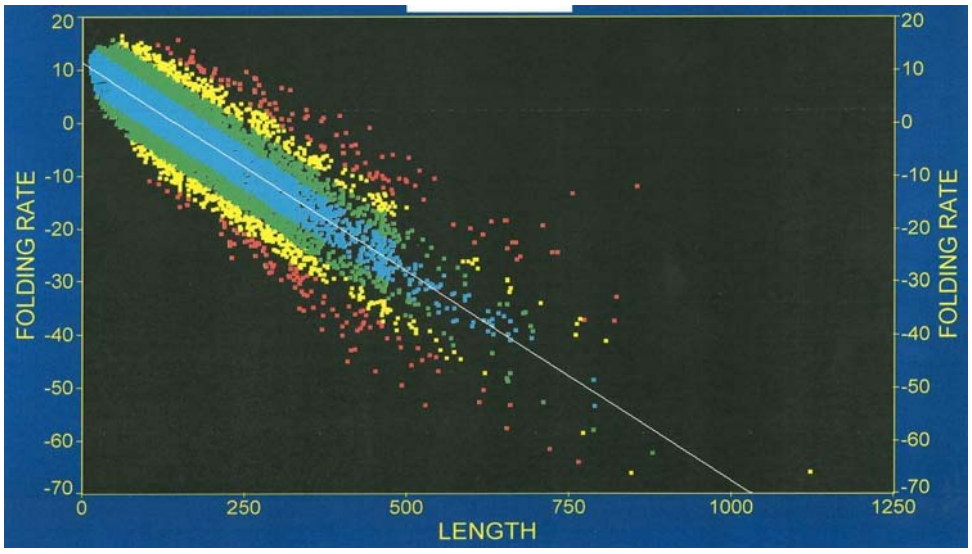


Figure 3: Curve fitting predicted folding rates of all organisms to linear equation y = a + bx; where 'y' is the predicted folding rate, 'x' is the length of the protein, a = 11.08, b = 0.08. We find that this simple equation fits the data with an r- square value of 0.8114 and standard error of 4.156.

5 RESULTS AND DISCUSSION

5.1 Cellular proteomes have a broad distribution of folding rates

Protein synthesis rates and folding rates hugely impact cell growth rates. In the present study we have calculated folding rate distributions across the whole proteome that give us more information than what we get by just calculating the average folding rate of a proteome. We downloaded whole proteomes of *E. coli, B. subtilis, S. cerevisiae, C. elegans, D. melanogaster, A. thaliana, M. musculus* and *H. sapiens* from UniprotKB. Using our model, we calculated the predicted folding rates of all the proteins between the lengths of 25 and 350 amino acids in a proteome. Figure 4 below contains the boxplot of predicted folding rates plotted against length for all the proteomes. Figure 5 illustrates the proteome-wide distribution of predicted folding rates for each proteome. We then do the length normalization and look at the new distributions (Fig. 6). By doing this we are basically removing the dependency of protein folding rate on length and looking to see if there are other factors influencing a given protein's folding rate.

The general shapes of the boxplot in Fig. 4 and histogram in Fig. 5 are conserved across different organisms. Cellular proteomes have a broad distribution of folding rates varying across 45 units of $ln(k_f)$ in almost all the model organisms. This is significant because 45 units on logarithmic scale translates to a huge range on a normal scale. The means, standard deviations and quantile ranges are given in Table 5. From Fig. 6 we see that the folding rate distributions have a characteristic bell shape and are almost similar after length normalization for all the mesophilic model organisms. The little difference

that was present in the original distributions was almost completely lost after length normalization. Table 6 gives the means, standard deviations and quantile ranges after length normalization. This essentially means that length of a protein is the major determinant of protein folding rate for proteins in the range of 25-350 amino acids. At the same time, we should also note that length is not the only determinant of a given protein's folding rate. If length was the only determinant we would have observed that all proteins after length normalization have the same folding rate; instead we see a characteristic normal-like distribution.

In all the distributions in Fig. 5, the extreme limits are +15 and -30 of $ln(k_f)$. 15 units of $ln(k_f)$ translated to 3.27 $(\mu sec)^{-1}$. This is equal to a folding time of ~0.3 μsec. Estimates from polymer collapse theory [43] and reaction rate theory [43] along with experimenta observations predict a speed limit of ~N/100 μsec for a single-domain protein of length N. Our extreme limit on fast folders is very close to this estimate. On the other hand cells also require their proteins to fold in a biologically relevant time for their proper functioning. Therefore, the folding rate of a protein should be comaparable to its synthesis rate. In *E. coli,* the translation speed is about 10-15 amino acids per second. Therefore, we can estimate that a protein shouldn't fold slower than about ~1/min [46]. Our limits from distributions not in acoordance with this theoretical extreme slow limit. This means that the assistance of molecular chaperones is vital for those very slow folding proteins which otherwise wouldn't be able to fod in a biologically relevant time.

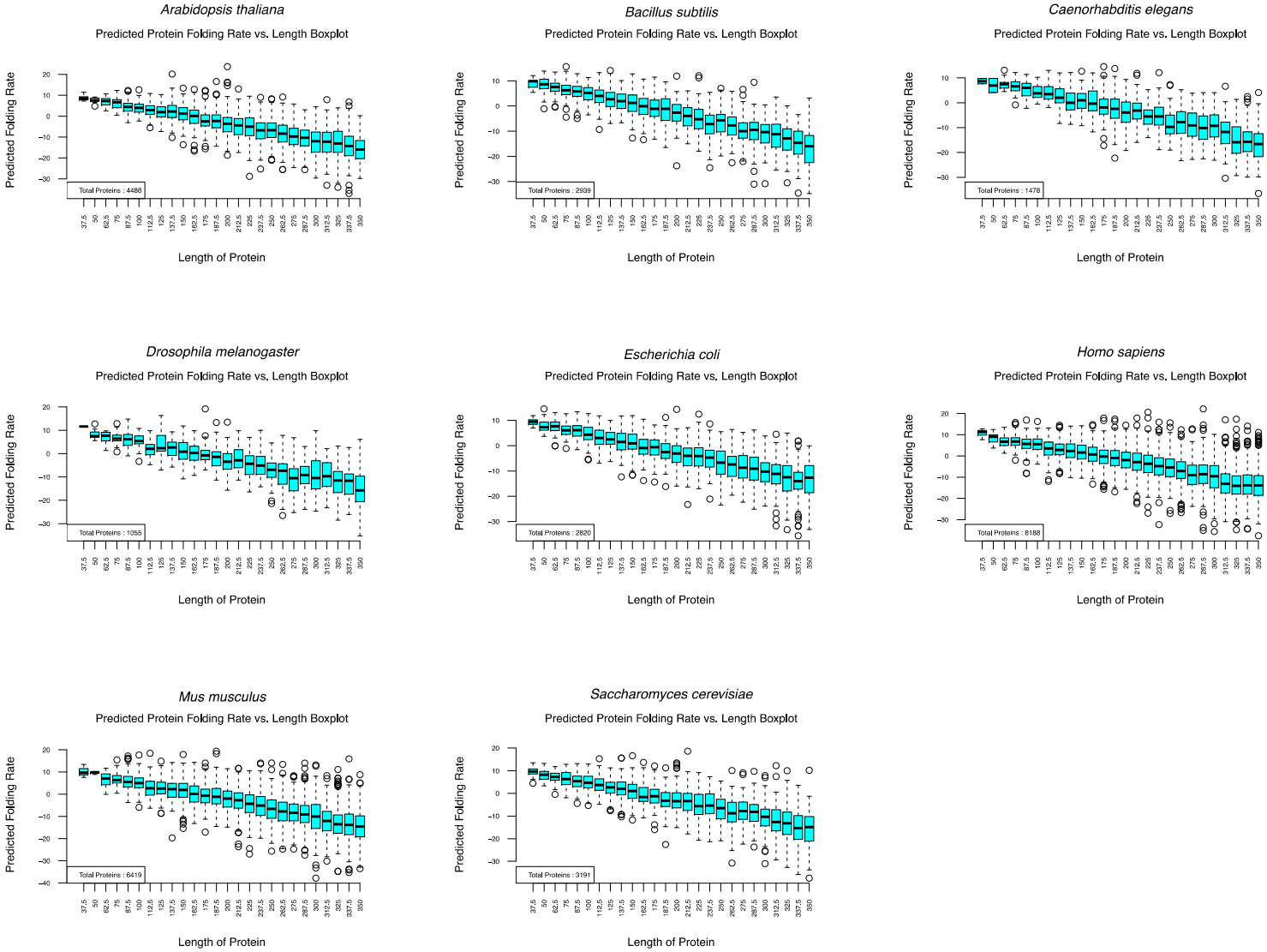


Figure 4: Boxplots of predicted folding rate vs. length for the mesophilic model organisms. Bin size is 17.5 units of length

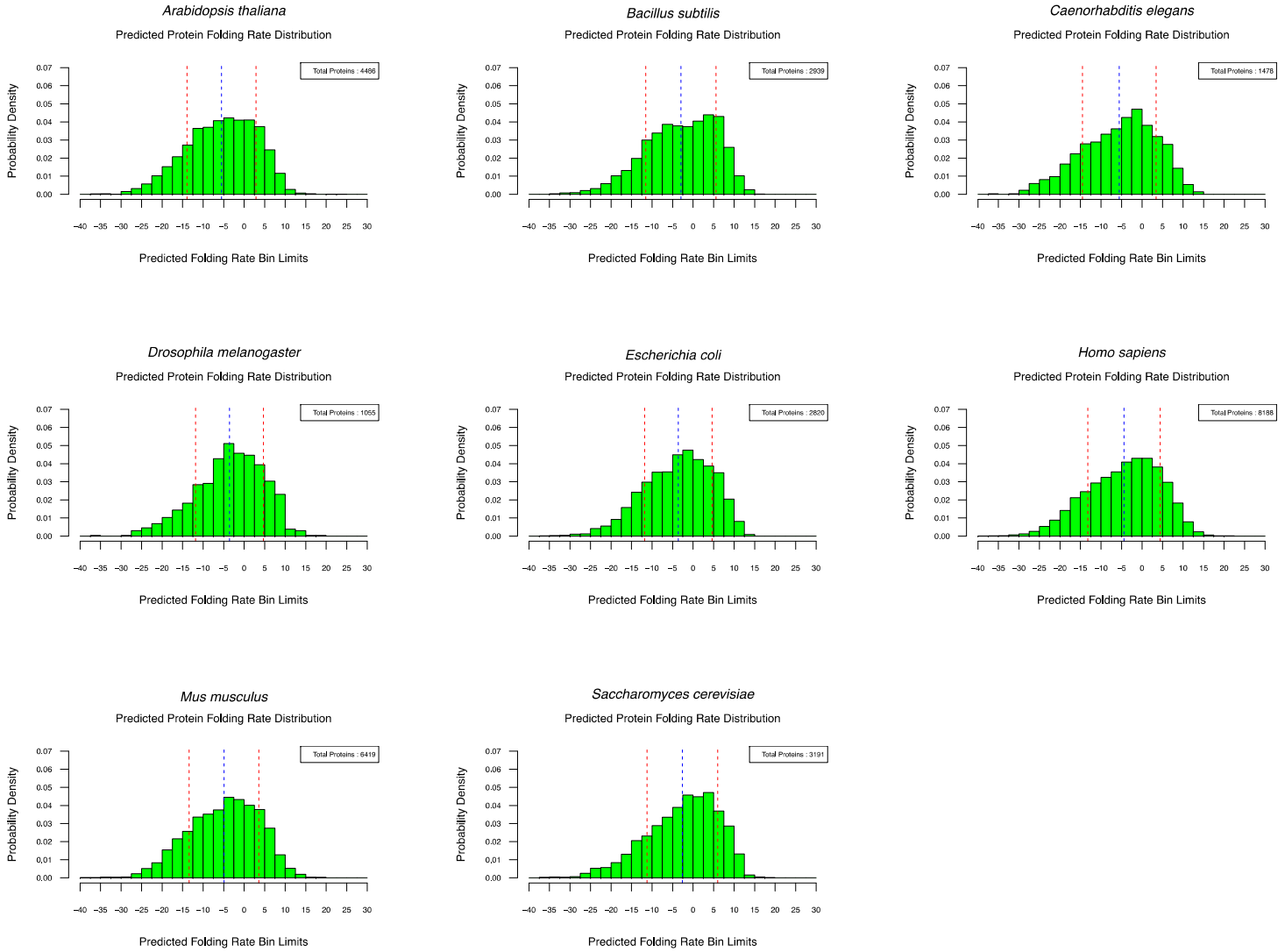


Figure 5: Predicted folding rate distributions for different mesophilic model organisms. Bin size is 2.5 units of $\ln(k_f)$. The blue line indicates the mean of the distribution; the red lines indicate the range of (mean ± standard deviation) for the distributions.

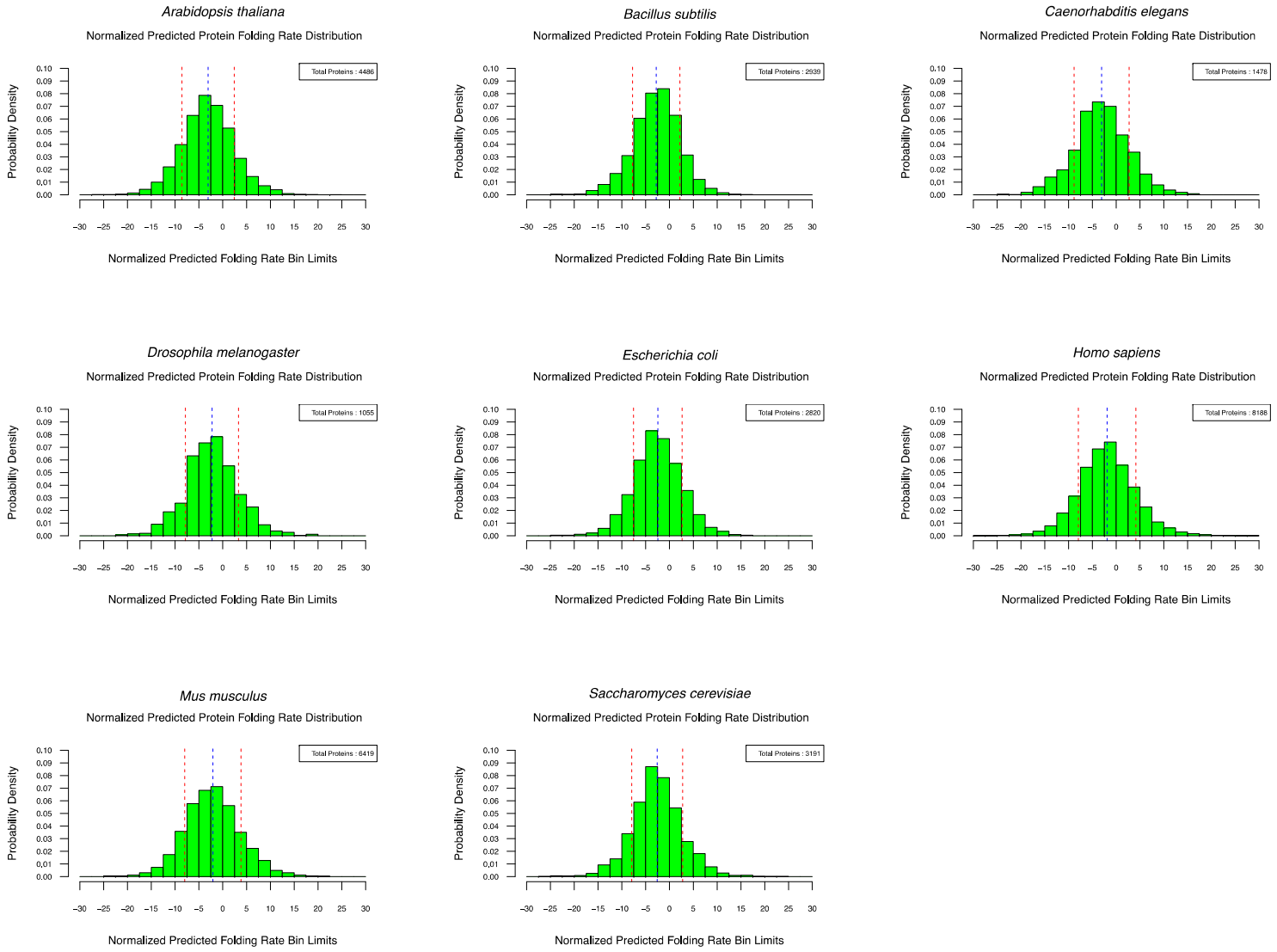


Figure 6: Predicted folding rate distributions **after length normalization** for different mesophilic model organisms. Bin size is 2.5 units of $\ln(k_f)$. The blue line indicates the mean of the distribution; the red lines indicate the range of (mean ± standard deviation) for the distributions.

Table 4: Mean, standard deviation (SD) and quantile ranges for the predicted folding rate distributions of the mesophilic model organisms

Organism	Mean	SD	0%	25%	50%	75%	100%
Arabidopsis thaliana	-5.51	8.42	-36.96	-11.37	-4.93	1.12	23.72
Bacillus subtilis	-2.96	8.57	-34.73	-8.80	-2.22	4.15	15.64
Caenorhabditis elegans	-5.56	8.97	-36.46	-11.88	-4.55	1.17	14.53
Drosophila Melanogaster	-3.58	8.28	-35.19	-8.92	-2.79	2.51	19.16
Escherichia coli	-3.61	8.22	-35.65	-9.40	-2.90	2.69	14.54
Homo sapiens	-4.40	8.82	-37.57	-10.65	-3.51	2.32	22.16
Mus musculus	-4.91	8.49	-37.78	-10.92	-4.23	1.65	19.38
Saccharomyces cerevisiae	-2.61	8.62	-37.39	-8.31	-1.41	3.91	18.62

Table 5: Mean, standard deviation (SD) and quantile ranges for the predicted folding rate distributions **after length normalization** of the mesophilic model organisms

Organism	Mean	SD	0%	25%	50%	75%	100%
Arabidopsis thaliana	-3.07	5.49	-26.06	-6.43	-3.10	0.37	23.99
Bacillus subtilis	-2.77	4.97	-23.32	-5.82	-2.54	0.41	16.88
Caenorhabditis elegans	-3.06	5.78	-24.01	-6.62	-3.11	0.61	16.16
Drosophila Melanogaster	-2.25	5.58	-22.35	-5.69	-2.31	1.13	18.67
Escherichia coli	-2.44	5.11	-24.32	-5.66	-2.55	0.78	15.19
Homo sapiens	-1.91	6.03	-28.40	-5.73	-2.03	1.73	29.94
Mus musculus	-2.06	5.89	-29.53	-5.92	-2.22	1.51	27.04
Saccharomyces cerevisiae	-2.58	5.35	-25.49	-5.75	-2.74	0.60	22.64

5.2 Folding rate distributions in extremophiles

Water molecules play a critical role in protein folding through the formation of hydrogen bonds and hydrophobic effect [44]. Hydrophobic effect is the burial of hydrophobic amino acid side-chains in the core of the protein. Water activity is hugely affected by extreme conditions like high/low temperatures and high salinity. In addition to this reduction in diffusion rates and changes in solvent viscosity [45] hugely impact protein folding in extremophiles.

We do the same analysis as mesophiles for the extremophiles. Fig. 7 gives the original folding rate distributions and Fig. 8 gives the folding rate distributions after length normalization. The shape of the distributions in the case of extremophiles is inconsistent and different from mesophilic model organisms. We observe a small increase in the percentage of the slowest folding proteins in halophiles when compared to mesophiles.

How extremophiles adapt to the extreme conditions to be able to be functional is a very complex issue. In our distributions we see a small increase in the slowest folding proteins in halophiles when compared to mesophiles. The halophiles through special adaptations seem to overcome these barriers for optimal protein folding.

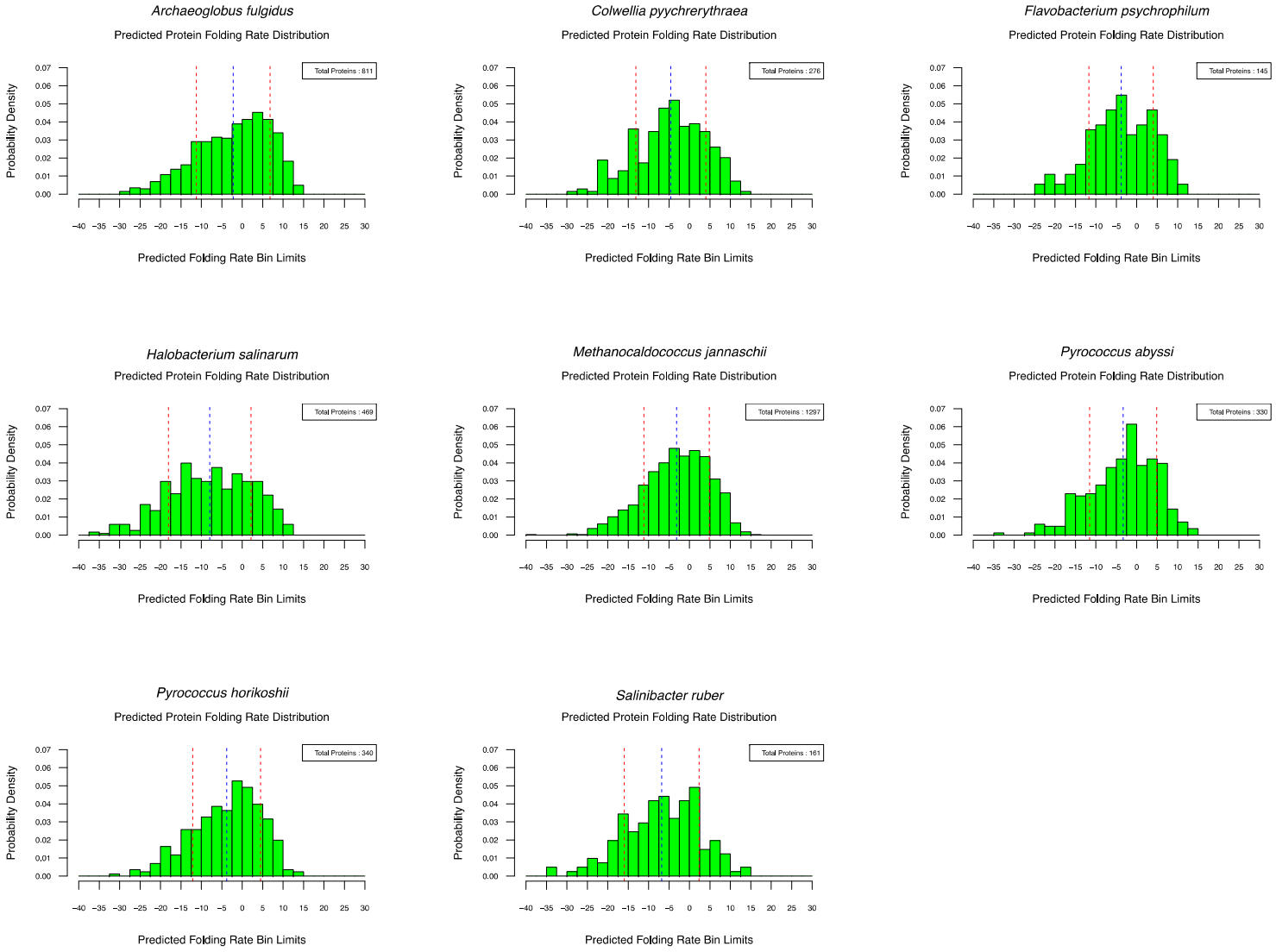


Figure 7: Predicted folding rate distributions for different extremophiles. Bin size is 2.5 units of $\ln(k_f)$. The blue line indicates the mean of the distribution; the red lines indicate the range of (mean ± standard deviation) for the distributions.

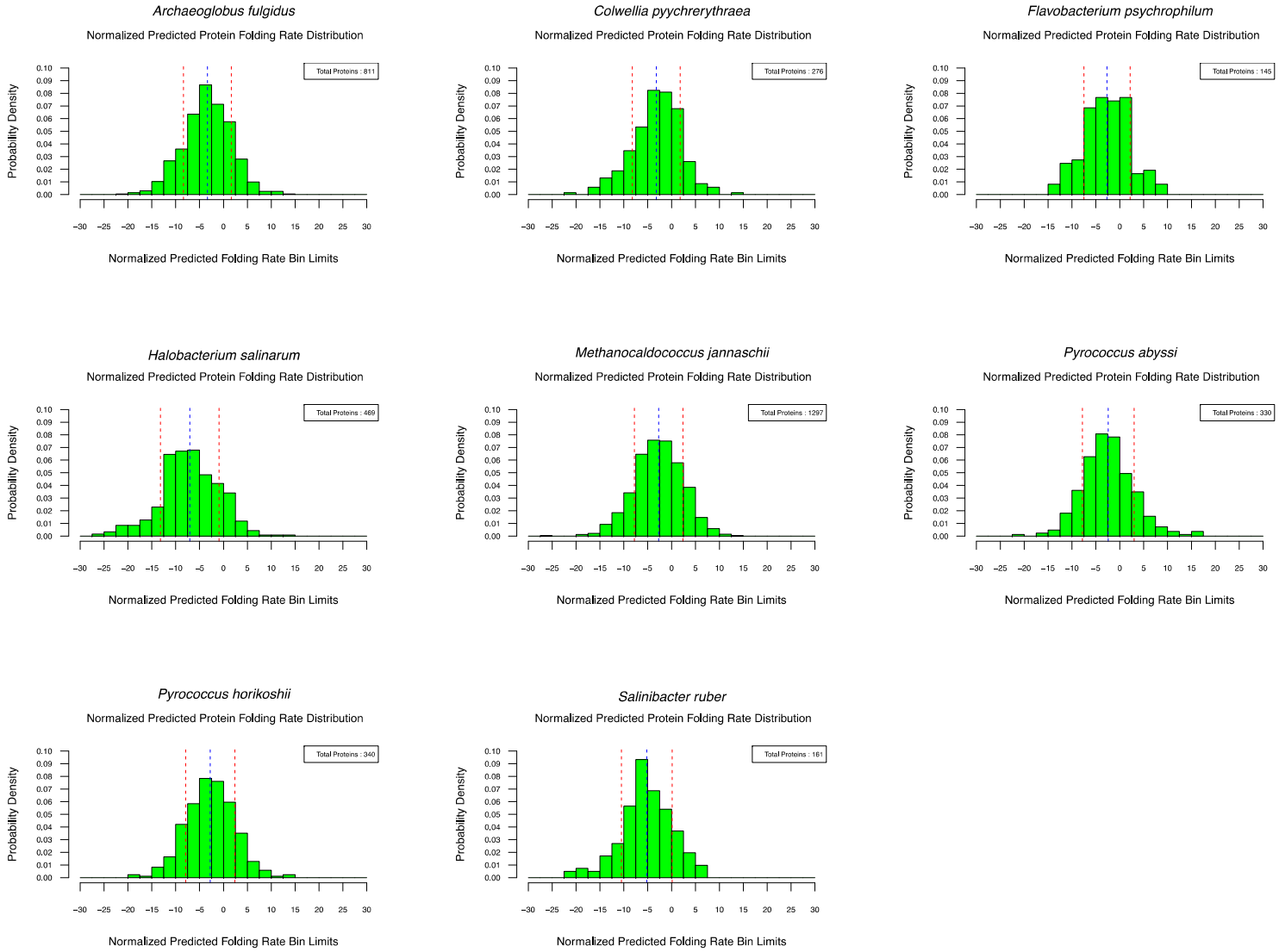


Figure 8: Predicted folding rate distributions **after length normalization** for different extremophiles. Bin size is 2.5 units of $\ln(k_f)$. The blue line indicates the mean of the distribution; the red lines indicate the range of (mean ± standard deviation) for the distributions.

Table 6: Mean, standard deviation (SD) and quantile ranges for the predicted folding rate distributions of the extremophiles

Organism	Mean	SD	0%	25%	50%	75%	100%
Archaeoglobus fulgidus	-2.22	9.02	-29.21	-8.81	-0.91	4.87	14.55
Colwellia pyychrerythraea	-4.58	8.56	-27.90	-9.65	-3.90	1.90	14.41
Flavobacterium psychrophilum	-3.84	7.86	-23.83	-9.28	-3.86	2.65	11.32
Halobacterium salinarum	-7.95	10.10	-36.41	-15.24	-7.64	0.12	12.14
Methanocaldococcus jannaschii	-3.17	8.00	-38.68	-8.44	-2.63	2.97	15.25
Pyrococcus abyssi	-3.33	8.21	-34.78	-8.22	-2.05	2.74	13.30
Pyrococcus horikoshii	-3.83	8.30	-31.73	-9.24	-2.63	2.31	13.41
Salinibacter ruber	-6.81	9.15	-33.14	-13.20	-5.98	0.17	13.62

Table 7: Mean, standard deviation (SD) and quantile ranges for the predicted folding rate distributions **after length normalization** of the extremophiles

Organism	Mean	SD	0%	25%	50%	75%	100%
Archaeoglobus fulgidus	-3.36	4.99	-20.48	-6.40	-3.37	0.03	12.93
Colwellia pyychrerythraea	-3.20	5.01	-20.98	-6.02	-2.83	0.23	13.05
Flavobacterium psychrophilum	-2.66	4.82	-13.05	-5.79	-2.63	0.77	9.48
Halobacterium salinarum	-7.06	6.12	-25.87	-10.43	-7.09	-2.90	14.38
Methanocaldococcus jannaschii	-2.73	5.08	-26.16	-6.04	-2.69	0.60	12.54
Pyrococcus abyssi	-2.41	5.40	-22.10	-5.97	-2.72	0.67	17.45
Pyrococcus horikoshii	-2.77	5.15	-19.63	-6.30	-2.87	0.72	13.04
Salinibacter ruber	-5.19	5.32	-21.57	-8.05	-5.29	-1.62	6.76

5.3 Fast and slow folders among representative sequences of all known protein families

We downloaded the representative sequences of all known protein families between lenghts of 25 and 350 amino acids from SCOP and applied our model to predict their folding rates. Then, we applied length correction on this data to analyze slow and fast folders.

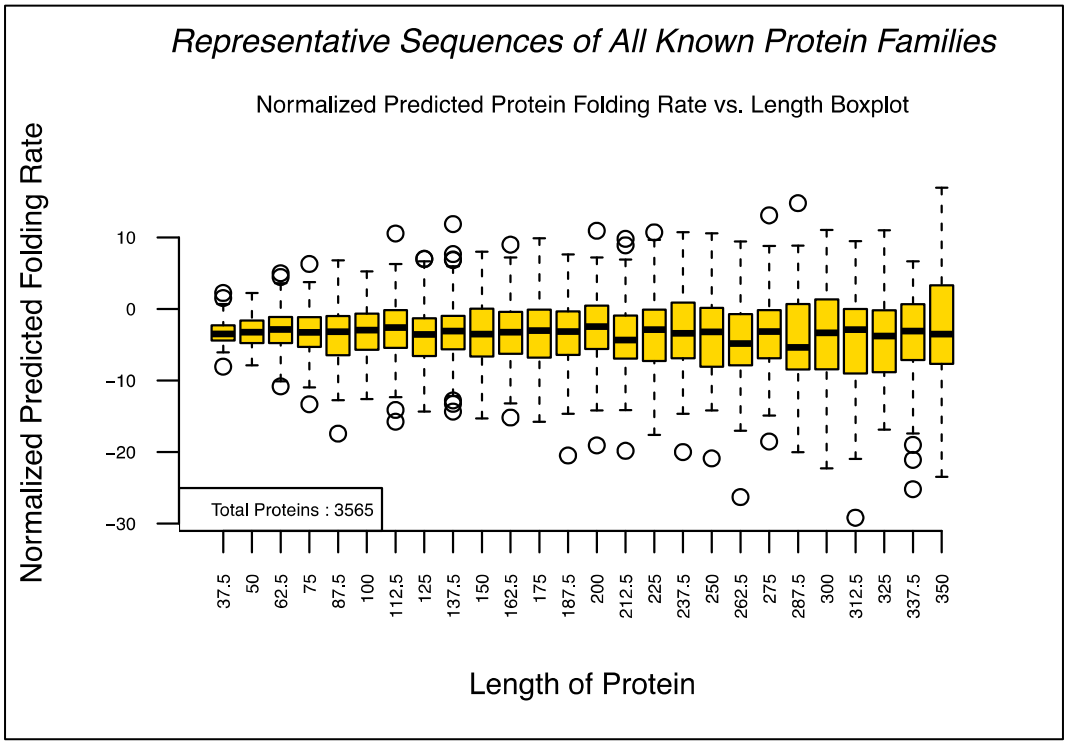


Figure 9: Folding Rate vs. Length plot after length normalization for representative sequences of all known protein families. The outliers marked as circles are used for significance analysis.

We collected 40 sequences from this dataset which have abnormally faster or slower folding rates when compared with other sequences of the same length. To confirm the significance of these 40 proteins, we do the following:

(a). We shuffle each protein sequence 1000 times and generate shuffled sequences with the same amino acid composition.

(b). We calculate predicted folding rates for these shuffled sequences.

(c). We plot the predicted folding rates' distribution of these 1000 sequences along with the predicted folding rate of the parent sequence (Figure 10).

(d). We also look at the PDB structure of the sequence to gain insight into the structure-folding rate relationship.

On analyzing the distributions we see that the folding rates of parent sequences fall near the tails of distribution as expected. This basically tells us that these sequences have a "special sequence order" which results in the abnormally faster or slower folding rates. If we had used amino acid composition alone or length to predict the folding rates in our model; these sequences would not have shown up as unique. It is interesting to observe that according to our model the fast folders among representative sequences of all known protein families have large percentage of alpha helices; while the slow folders are mainly made up of beta sheets.

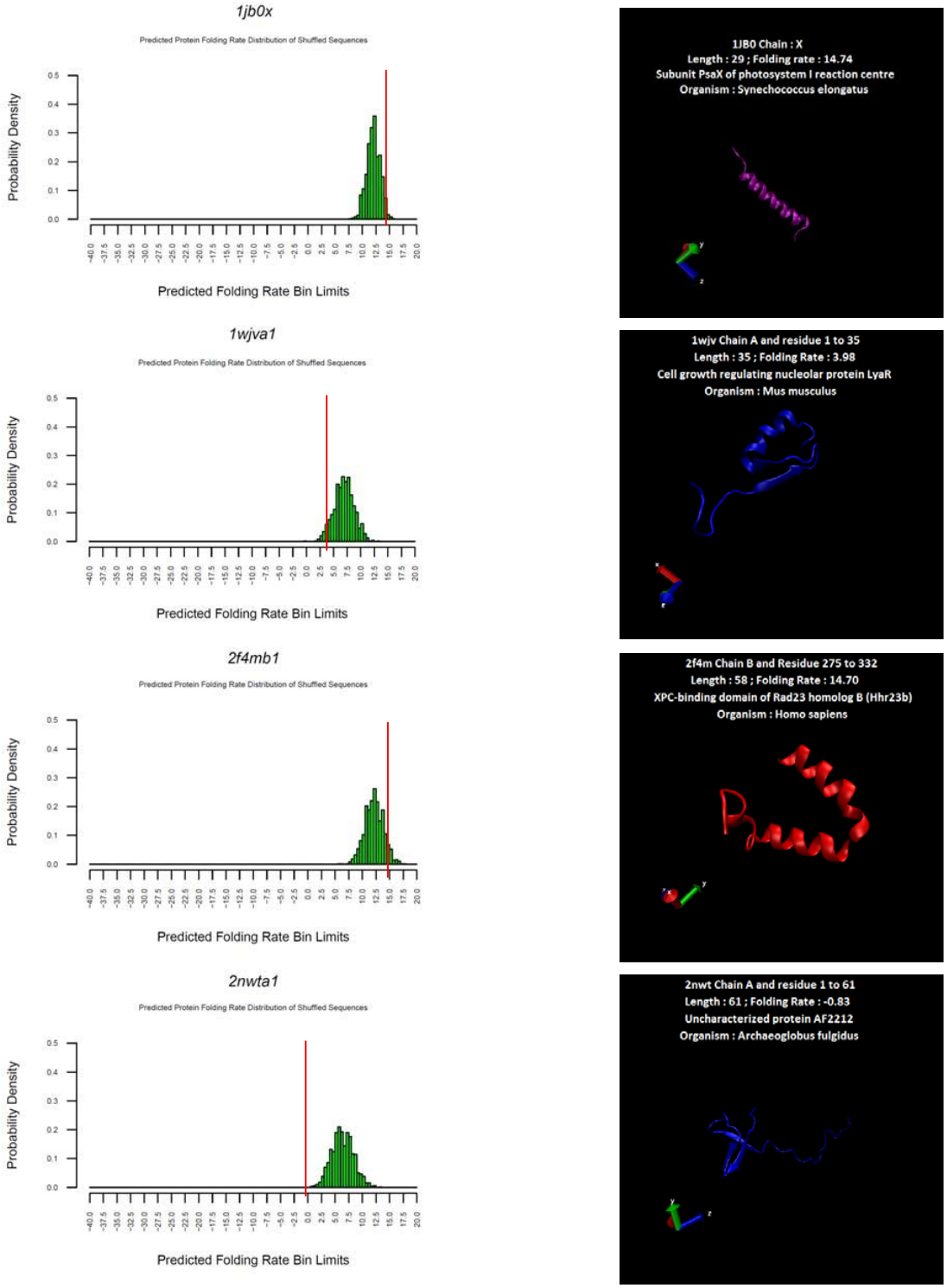
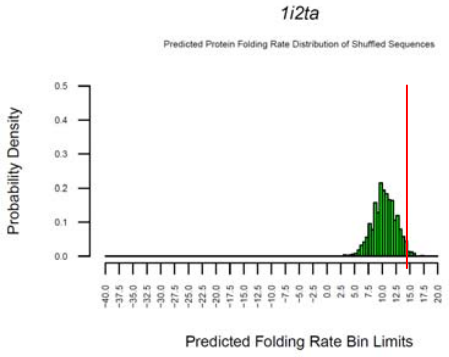
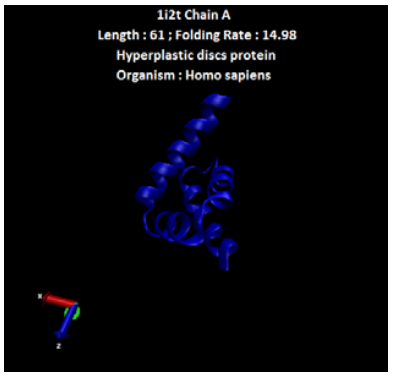
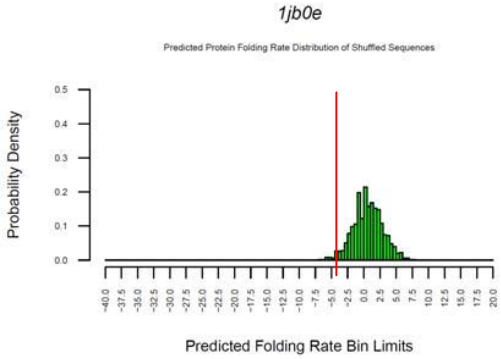
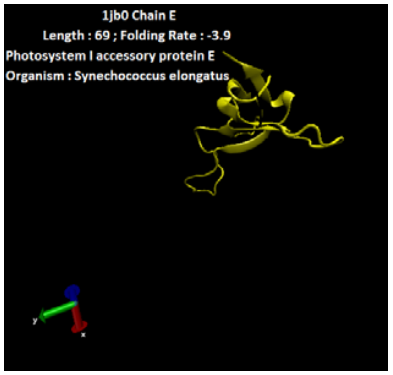


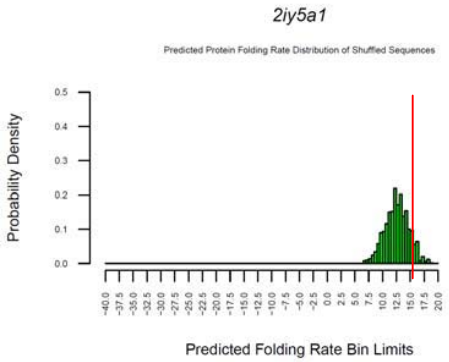
Figure 10: Predicted folding rate distributions of randomly shuffled sequences of slow and fast folding proteins among representative sequences of all known protein families. The red line indicates the value where the predicted folding rate of the actual parent sequence lies in the distribution. On the right side of each distribution we give the PDB structure of the sequence.

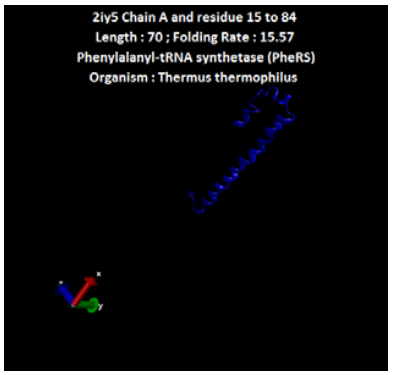


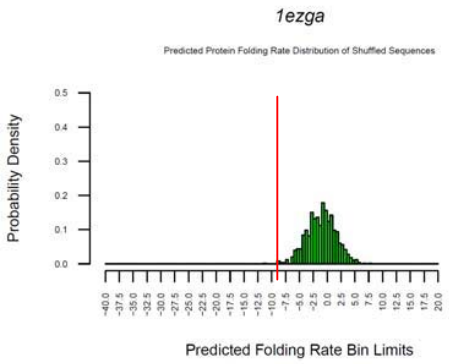












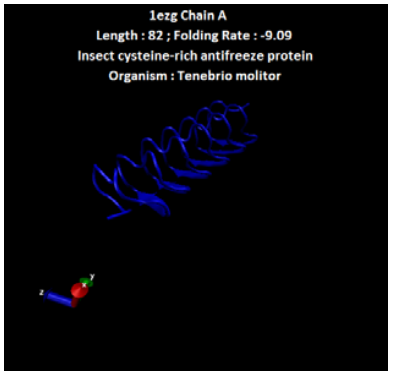


Figure 10: continued

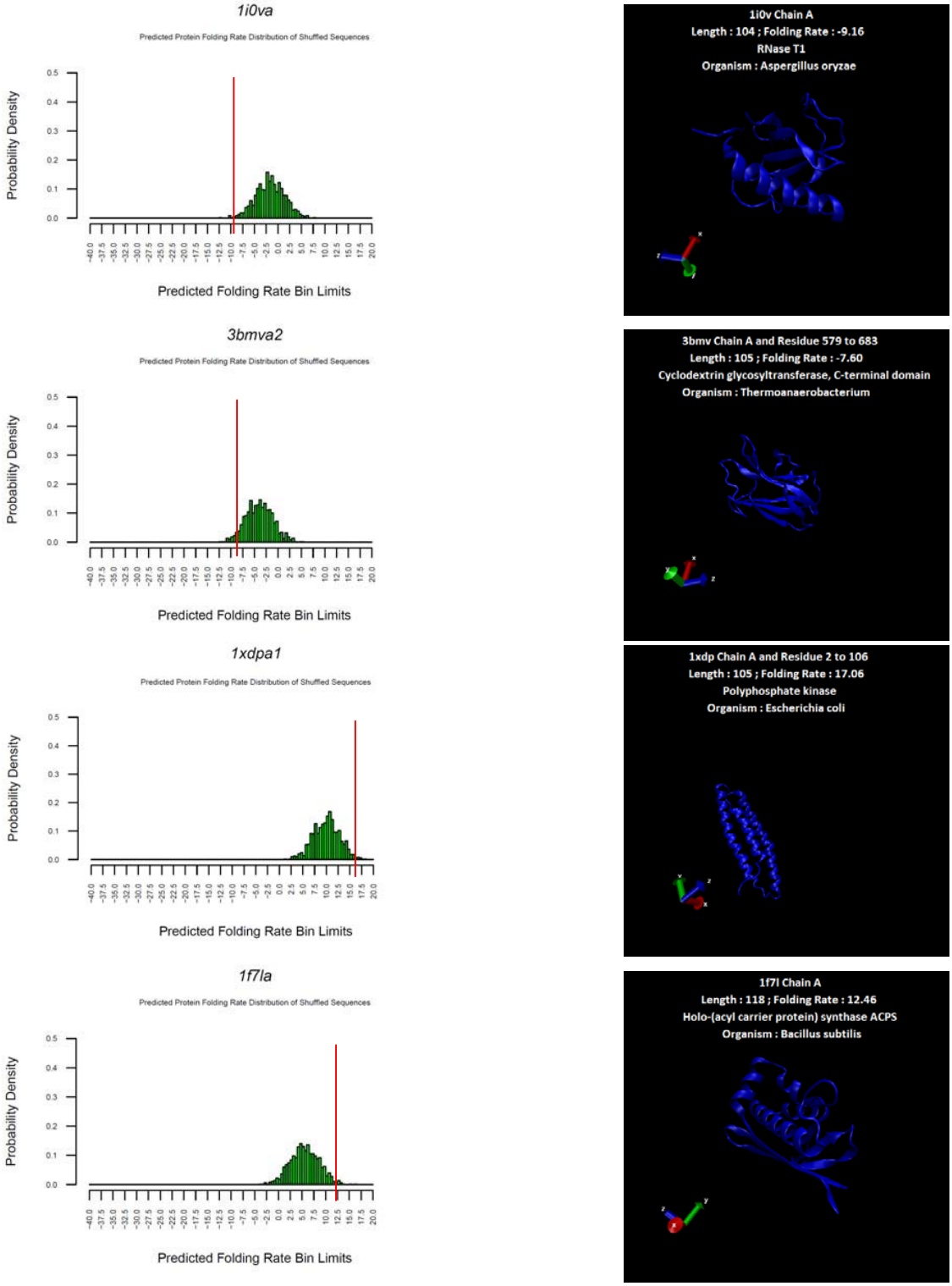
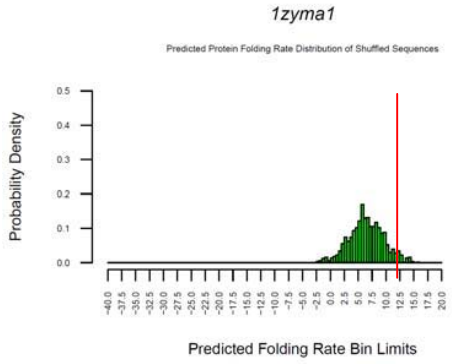
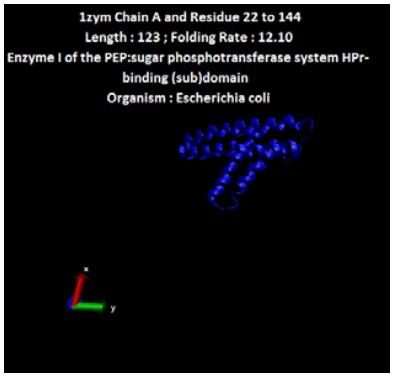
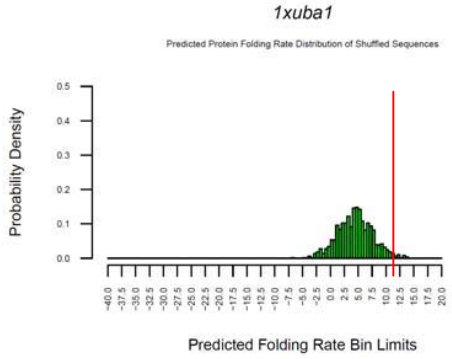
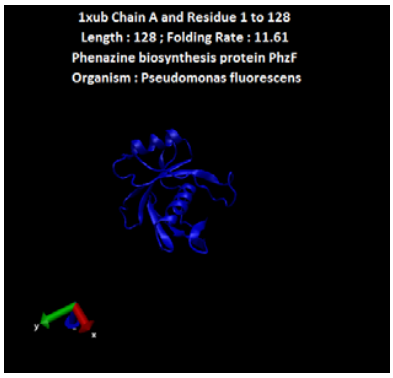


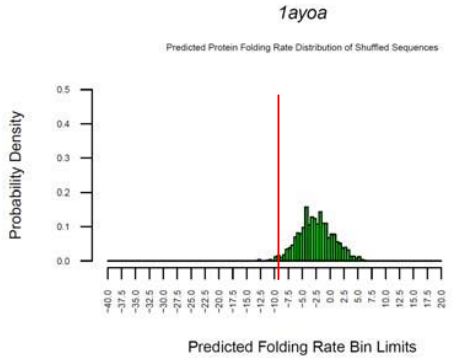
Figure 10: continued

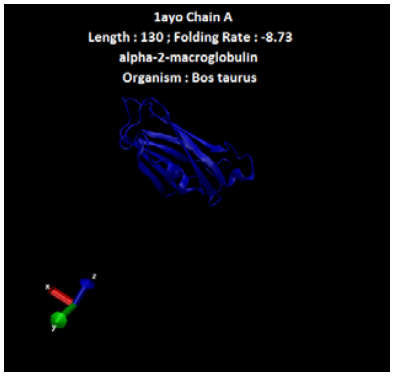


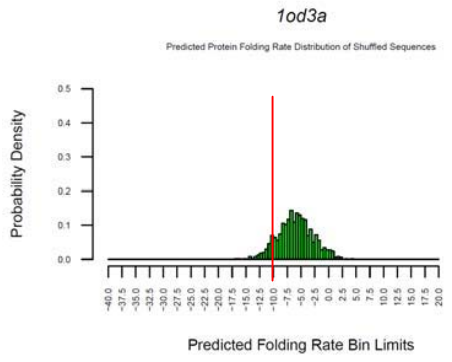












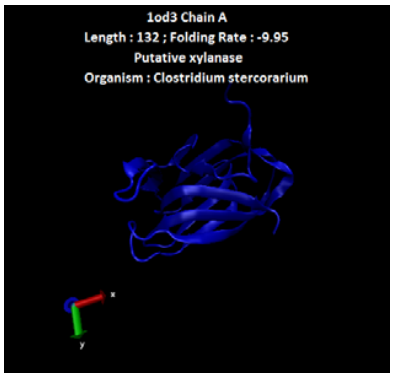
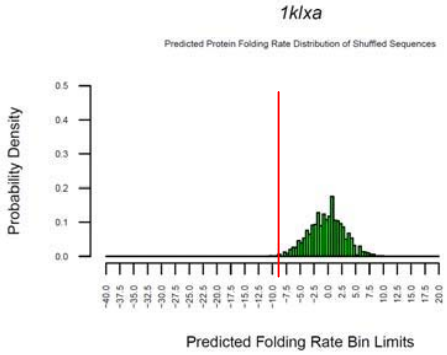
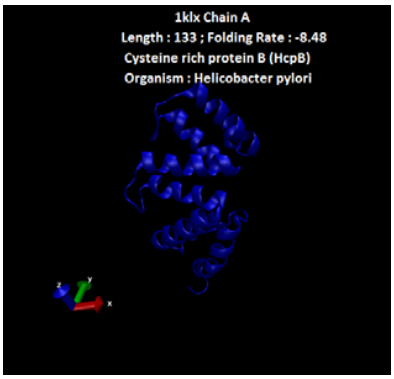
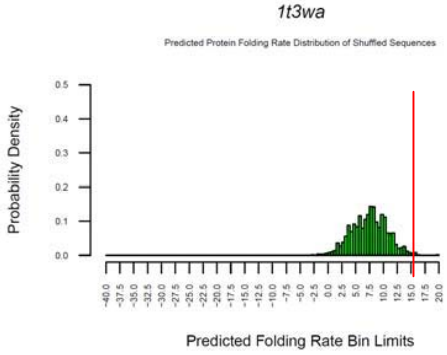
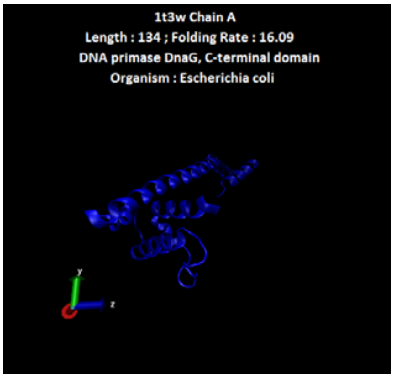


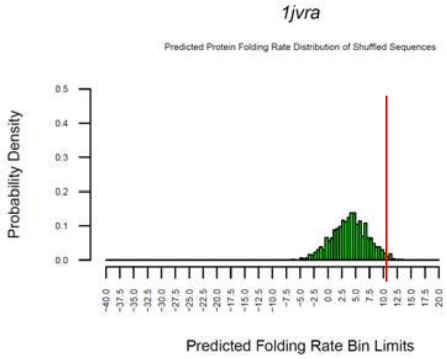
Figure 10: continued

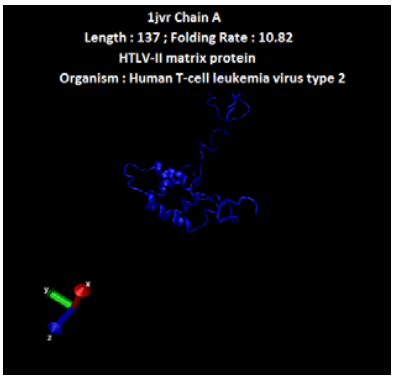


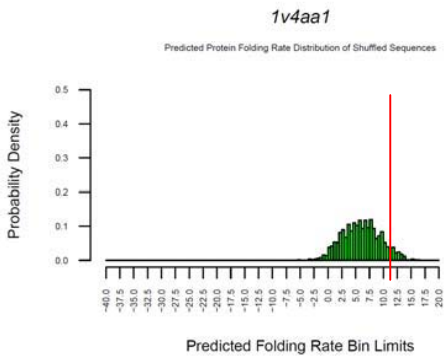












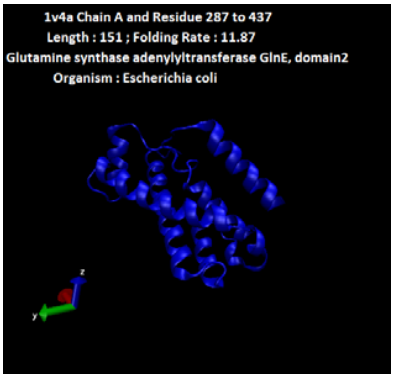
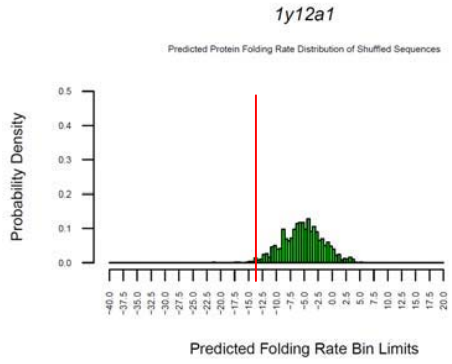
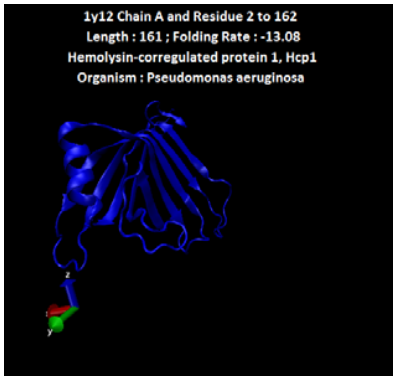
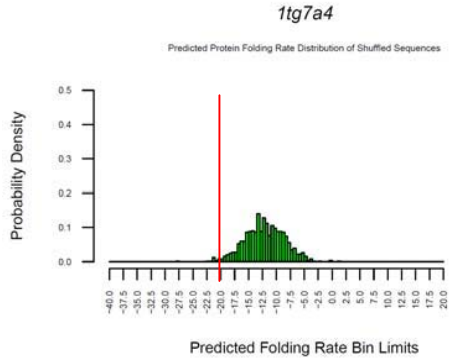
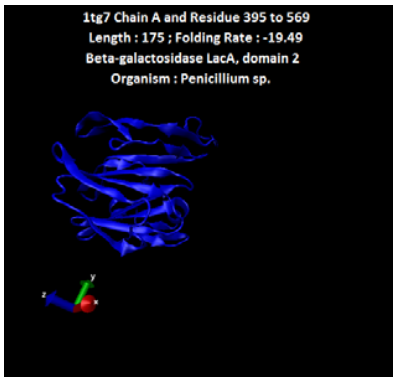


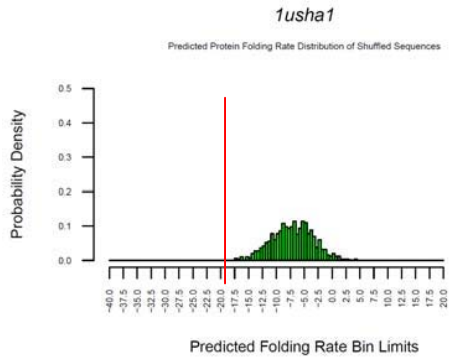
Figure 10: continued

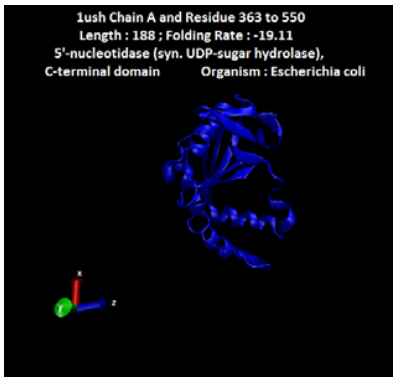


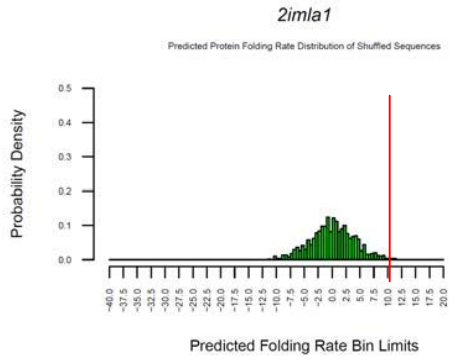












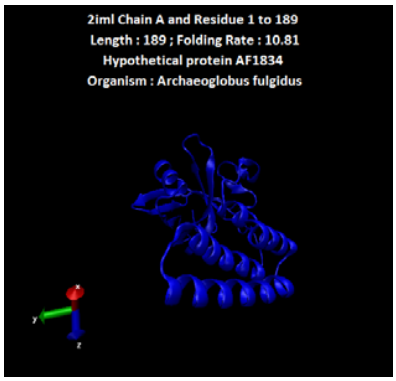
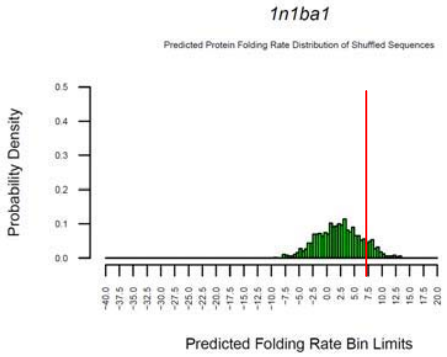
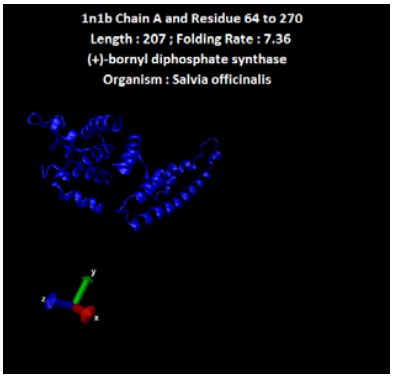
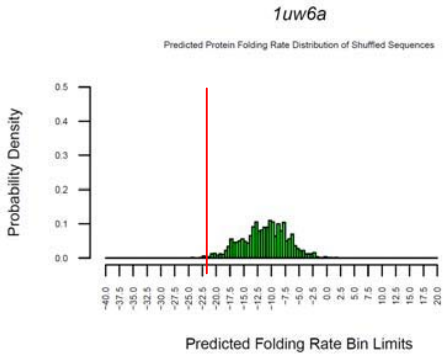
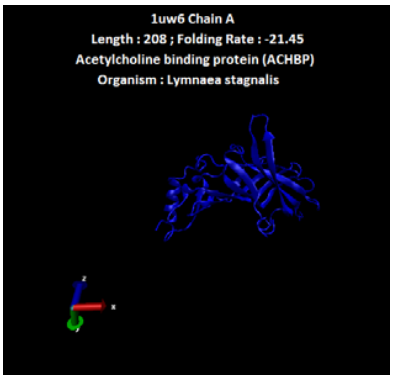


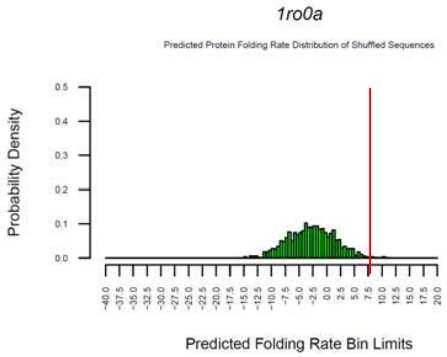
Figure 10: continued

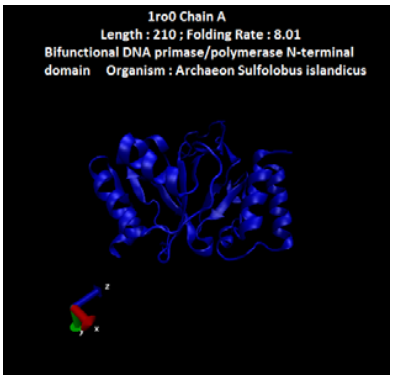


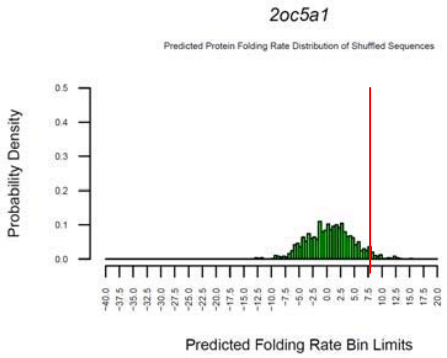












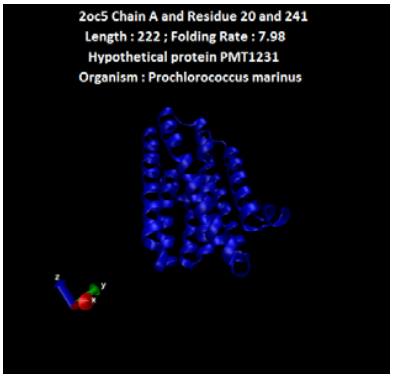
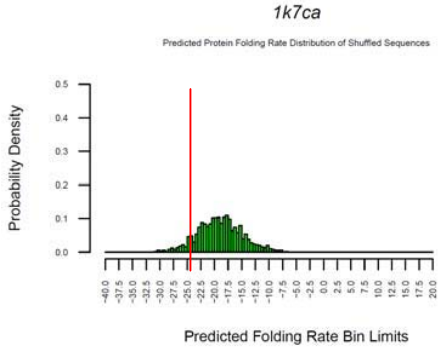
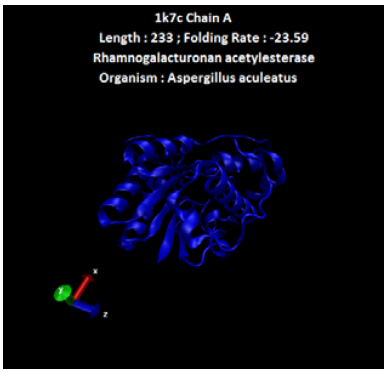
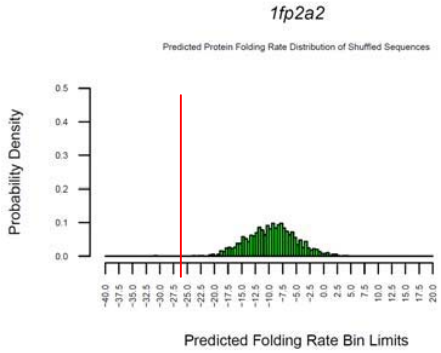
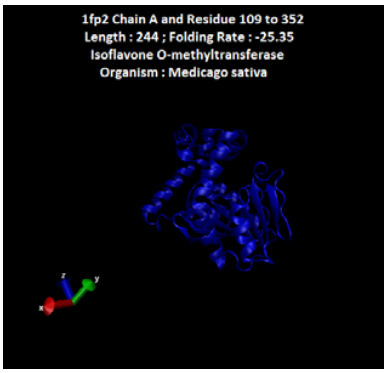


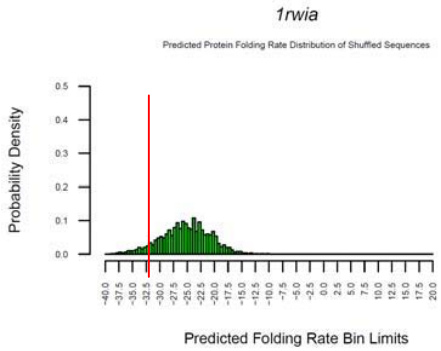
Figure 10: continued











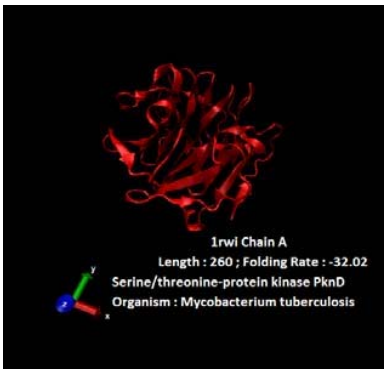


Figure 10: continued

5.4 Relationship between structure and folding rate

It has been observed that alpha helical proteins have the fastest folding kinetics when compared to all beta sheet proteins of proteins which are combinations of alpha helices and beta sheets. This is in accordance to the popular idea of topology affected folding rates [8]. Alpha helocal proteins have the smallest number of contacts per residue (structurally less complex) and thus have faster folding kinetics. It is also observed that proline isomerization and conformationally restrictive disulfide bonds [46] result in slower folding of proteins.

Our model's results [Fig. 9] are in accordnace with the above theory. The fast folders among representative sequences of all known protein families have large percentage of alpha helices; while the slow folders are mainly made up of beta sheets. Among the 40 families, we will look more deeply into the following two more structurally complex families.

5.4.1 Heliobacter pylori cysteine rich protein B

The PDB structure of the *H. pylori* cysteine rich protein Chain A (1KLX) is given in Fig. 11 (a). This is an all-alpha motif and should technically fold fast. Our model predicts that this protein folds very slowly ($ln(k_f)$ = -8.46). So, we investigate further into the sequence and three dimensional structure. It turns out that di-sulfide bridges play a vital role in the folding kinetics of this particular protein. The PDB structure with di-sulfide bridges is given in Fig. X (b). It seems that the relative slow kinetics of di-sulfide bond formation between cysteine residues is critically affecting the folding kinetics of this cysteine rich protein [47].

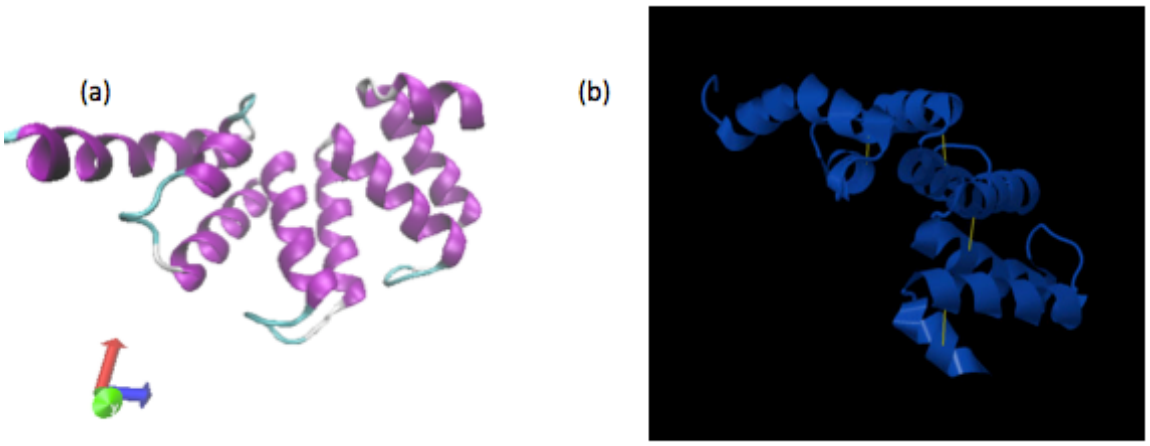


Figure 11: (a) PDB structure of 1KLX, (b) PDB structure of 1KLX with di-sulfide bridges

5.4.2 **Ribonuclease T1**

The PDB structure of Ribonuclease T1 Chain A (1I0V) is given in Fig. 12 (a). This is a relatively short protein chain and should fold fast. Our model predicts that this protein folds much slower ($ln(k_f)$ = -9.16) than other proteins of comparable length. When we look more deeply into the sequence and structure we learn that isomerization of the two cis prolyl bonds at Pro 39 and Pro 55 could be the rate-determining steps of the slow folding kinetics of Ribonuclease T1 chain A [48].

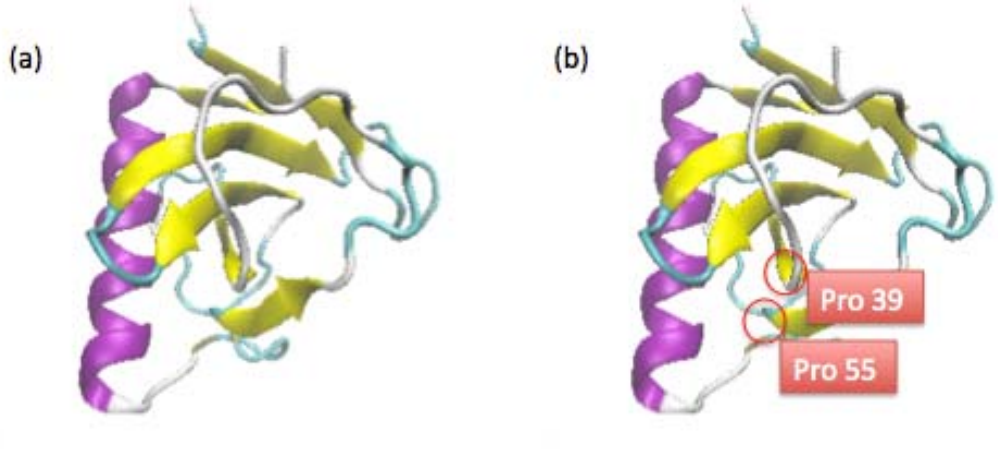


Figure 12: (a) PDB structure of 1I0V, (b) PDB structure of 1I0V with the critical Pro residues labelled

6 CONCLUSIONS

Understanding the relationship between sequence, structure, and folding rate of a protein is critical to gain insight into the protein folding mechanism. In our current work, we used a model which takes into account the 'sequence order' in a protein in predicting the folding rate of a protein. Our results show that 'sequence order' is indeed a very important determinant of a protein's folding rate along with the length. We used this model to comute the folding rate distributions of a proteome from which we look at the extreme limits of folding rate in a cell. We find that many proteins in a cell require the help of the chaperone machinery to be able to fold in a biologically relevant time. We have also shown that halophiles have a difrent distribution of folding rates when compared to mesophiles. We further find the fastest and slowest folding proteins among those representing all known protein families. We analyzed certain structures among these fast/slow folders to gain valuable insight into the structure-folding rate relationship.

7 FUTURE WORK

Understanding how simple mutations affect folding rate of a protein is one of the most important challenges in current bioinformatics. Our current model is not sensitive enough to predict folding rate changes due to mutations. We need a larger and more diverse experimental dataset to be able to create a more sensitive prediction model.

Is ribosomal translation speed connected to protein folding rates? A recent study [37], reported that slowing bacterial translation rates enchanced eukaryotic protein folding effeciency in a prokaryotic system. It is important to know if the folding pathways of proteins evolved in the context of translation rates.

Cited Literature

1. Anfinsen, C.B., et al., *The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain*. PNAS, 1961. 47(9): p.1309–1314.
2. Zeeb, M. and J. Balbach, *Protein folding studied by real-time NMR spectroscopy*. J. Methods, 2004. 34(1): p.65–74.
3. Fabian, H. and D. Naumann, *Methods to study protein folding by stopped-flow FT-IR*. J.Methods, 2004. 34(1): p.28–40.
4. Zarrine-Afsara, A. and A. R. Davidson, *The analysis of protein folding kinetic data produced in protein engineering experiments*. J. Methods, 2004. 34(1): p.41–50.
5. Maity, H.; et al., *Protein folding: The stepwise assembly of foldon units.* Proc Nat Acad Sci USA, 2005. 102(13): p.4741–4746.
6. Xiao, H. et al., *Mapping protein energy landscapes with amide hydrogen exchange and mass spectrometry. I. A generalized model for a two-state protein and comparison with experiment.* Protein Sci, 2005. 14: p.543–557.
7. Maxwell, K. L. et al., *Protein folding: Defining a "standard" set of experimental conditions and a preliminary kinetic data set of two-state proteins.* Protein Sci, 2005. 14: p.602–616.
8. Plaxco, K. W. et al., *Contact order, transition state placement, and the refolding rates of single domain proteins*. J Mol Biol, 1998. 277: p.985–994.
9. Sousa, S.F. et al., *Protein-ligand docking: current status and future challenges.* Proteins, 2006. 65(1): p. 15-26.

10. Weikl, T. R. and K. A. Dill, *Folding kinetics of two-state proteins: Effect of circularization, permutation, and crosslinks.* J Mol Biol, 2003. 332: p.953–963.
11. Zhou, H. and Y. Zhou, *Folding rate prediction using total contact distance.* Biophys J, 2002. 82(1 Pt 1): p. 458-463.
12. Nolting, B., et al., *Structural determinants of the rate of protein folding.* J Theor Biol, 2003. 223(3): p. 299-307.
13. Ivankov, D.N. and A.V. Finkelstein, *Prediction of protein folding rates from the amino acid sequence-predicted secondary structure.* Proc Natl Acad Sci U S A, 2004. 101(24): p. 8942-8944.
14. OuYang, Z. and J. Liang, *Predicting protein folding rates from geometric contact and amino acid sequence.* Protein Sci, 2008. 17: p.1256–1263.
15. Huang, L. T. and M. M. Gromiha, *Analysis and prediction of protein folding rates using quadratic response surface models.* J Comput Chem, 2008. 29: p.1675–1683.
16. Jiang, Y.; Iglinski, P. and L. Kurgan, *Prediction of protein folding rates from primary sequences using hybrid sequence representation*. J Comput Chem, 2009. 30: p.772–783.
17. Gao, J. et al., *Accurate prediction of protein folding rates from sequence and sequence-derived residue flexibility and solvent accessibility.* Proteins, 2010. 78: p.2114–2130.
18. Shen, H. B. et al., *Prediction of protein folding rates from primary sequence by fusing multiple sequential features.* J Biomed Sci Eng, 2009. 2: p.136–143.
19. Chou, K. C. and H. B. Shen, *FoldRate: A web-server for predicting protein folding rates from primary sequence.* Open Bioinformatics J, 2009. 3: p.31–50.

20. Gou J., et al., *Predicting protein folding rates using the concept of Chou's pseudo amino acid composition.* J Comput Chem, 2011. 32(8): p.1612-7.

21. Ouyang Z., *Prediction of Protein Binding and Folding From Molecular Geometry and Empirical Scoring Functions.* University of Illinois at Chicago, 2010.

22. Plaxco, K.W., et al., *Topology, stability, sequence, and length: defining the determinants of two-state protein folding kinetics.* Biochemistry, 2000. 39(37): p. 11177-11183.

23. Gromiha, M.M. and S. Selvaraj, *Comparison between long-range interactions and contact order in determining the folding rate of two-state proteins: application of long- range order to folding rate prediction.* J Mol Biol, 2001. 310(1): p. 27-32.

24. Gong, H., et al., *Local secondary structure content predicts folding rates for simple, two- state proteins.* J Mol Biol, 2003. 327(5): p. 1149-1154.

25. Gromiha, M.M., *A statistical model for predicting protein folding rates from amino acid sequence with structural class information.* J Chem Inf Model, 2005. 45(2): p. 494-501.

26. Thirumalai D., *From minimal models to real proteins: Time scales for protein folding kinetics*. J Phys I France, 1995. 5: p.1457–1467.

27. Finkelstein, A.V. and A.Y.A Badretdinov, *Rate of protein folding near the point of thermodynamics equilibrium between the coil and the most stable chain fold.* Fold Des, 1997. 2: p.115–121.

28. Koga, N. and S. Takada, *Role of native topology and chain-length scaling in protein folding: A simulation study with a Go-like model.* J Mol Biol, 2001. 313: p.171–180.

29. Galzitskaya OV., et al., *Chain length is the main determinant of the folding rate of proteins with three-state folding kinetics.* Proteins, 2003. 51: p.162–166.

30. Ivankov, D. N., et al., *Contact order revisited: Influence of protein size on the folding rate.* Protein Sci, 2003. 12: p.2057–2062.

31. Naganathan, A. N. and V. Munoz, *Scaling of folding times with protein size.* J Am Chem Soc, 2005. 127: p.480–481.

32. Wu C. H., et al., *The universal protein resource (uniprot): an expanding universe of protein information.* Nucleic Acids Res, 2006. 34(1): p.187.

33. Hubbard, T. J., et al., *SCOP: A Structural Classification of Proteins database.* Nucleic Acids Research, 1999. **27** (1): p.254–256.

34. Brenner, S. E., et al., *The ASTRAL compendium for protein structure and sequence analysis.* Nucleic Acids Research, 2000. 28: p.254-256.

35. Chung, Y. J., et al., *Size Comparisons among Integral Membrane Transport Protein Homologues in Bacteria, Archaea, and Eucarya.* J. Bacteriol, 2001**.** 183 (3): p.1012-1021.

36. Andrade, M. A., et al., *Protein Repeats: Structures, Functions, and Evolution.* Journal of Structural Biology, 2001. 134: p.117–131.

37. Siller, E., et al., *Slowing Bacterial Translation Speed Enhances Eukaryotic Protein Folding Efficiency.* J. Mol. Biol, 2010. 396: p.1310–1318.

38. Pedersen, S., *Escherichia coli ribosomes translate in vivo with variable rate.* EMBO J., 1984. 3: p.2895–2898.

39. Liang, S. T., et al., *mRNA composition and control of bacterial gene expression.* J. Bacteriol, 2000. 182: p.3037–3044.

40. Mathews, M. B., et al., *Origins and principles of translational control. In Translational Control of Gene Expression.* Sonenberg, N., Hershey, J. W. B. & Mathews, M. B., eds, 2000. p. 1–31.

41. Levinthal, Cyrus. *How to Fold Graciously*. Mossbauer Spectroscopy in Biological Systems: Proceedings of a meeting held at Allerton House, Monticello, Illinois, 1969. p.22–24.

42. Dill, K.A., Ghosh, K., and J.D. Schmit, *Physical limits of cells and proteomes*. Proc Natl Acad Sci, 2011. 108: p.17876–17882.

43. Kubelka, J., Hofrichter, J., and W.A. Eaton, *The protein folding 'speed limit'*. Curr Opin Struct Biol, 2004. 14: p.76–88.

44. Zaccai, G., and H. Eisenberg, *Halophilic Proteins and the Influence of Solvent on Protein Stabilization*. Trends Biochem Sci, 1990. 15: p.333-337.

45. Kramers, H. A., *Brownian motion in a field of force and the diffusion model of chemical reactions*. Physica, 1940. 7(4): p.284-304.

46. Eaton, W. A., et al., *Fast kinetics and mechanisms in protein folding.* Annu Rev Biophys Biomol Struct, 2000. 29: p.327-359

47. Mamathambika, B.S., and J.C. Bardwell, *Disulfide-linked protein folding pathways*. Annu. Rev. Cell. Dev. Biol, 2008. 24 : p.211–235.

48. Kiefhaber, T., et al., *Structure of a rapidly formed intermediate in ribonuclease T1 folding*. Protein Sci, 1992. 1: p.1162-1172.

49. Ma, B., Guo, J., and H. Zhang, *Direct correlation between proteins' folding rates and their amino acid compositions: an ab initio folding rate prediction.* Proteins, 2006. 65: p.362–372.

50. Gromiha, M. M., Thangakani, A. M., and S. Selvaraj, *FOLD-RATE: prediction of protein folding rates from amino acid sequence*. Nucleic Acid Res, 2006. 34: p.70-74.

51. L-T. Huang and M. M. Gromiha, *Analysis and prediction of protein folding rates using quadratic response surface models.* J. Comp. Chem, 2008. 29: p.1675-16

VITA

NAME: Sandeep C. Gorla

EDUCATION: B.S. Equivalent, Biotechnology, National Institute of Technology Durgapur, Durgapur, India, 2010.

M.S. Bioinformatics, University of Illinois at Chicago, Illinois, 2012, under the advising of Professor Jie Liang.