Journal of Proteomics & Bioinformatics - Open Access

Research Article
OPEN ACCESS Freely available online
doi:10.4172/jpb.1000152
JPB/Vol.3 Issue 10

CART Assignment of Folding Mechanisms to Homodimers with Known Structures

Abishek Suresh[1]*, Pattabhiraman Lalitha[1] and Pandjassarame Kangueane[1,2]

[1]*Faculty of Applied Sciences, Department of Biotechnology, AIMST University, Semeling, Kedah, Malaysia*
[2]*Biomedical Informatics, Irulan Sandy Annex, Pondicherry 607 402, India*

Abstract

Protein homodimers play a critical role in catalysis and regulation and their mechanism of folding is intriguing. The mechanisms of homodimer folding (2-state [2S] without intermediates and 3-state [3S] with either monomer [3SMI] or dimer [3SDI] intermediates) have been observed and documented for about 46 homodimers (27 2S; 12 3SMI; 7 3SDI) with known 3D structures. Determination of folding mechanisms through classical denaturation experiments is both time consuming, tedious, and expensive. Therefore, it is of interest to predict their folding mechanism. Furthermore, a large number of homodimers structures with unknown folding mechanism are available in the PDB. Hence, it is compelling to predict their folding mechanism using structural features intrinsic of each complex structure. Thus, we developed a classification and regression tree (CART) model using predictive parameters ((a) monomer protein size (ML); (b) interface area (B/2); (c) interface to total residues (I/T) ratio) derived from a dataset (46 homodimers with both known structures and folding mechanism) for folding mechanisms prediction. The dataset was subjectively divided into training (13 2S; 6 3SMI; 3 3SDI) and testing (14 2S; 6 3SMI; 4 3SDI) sets for validation. The model performed fairly well for predicting 2S and 3SMI in both during training and testing using ML and I/T as predictive variables. However, it should be noted that the performance of model in classifying 3SDI is poor. Nonetheless, the model was not stable with the inclusion of the predictive variable B/2 and hence, was not considered during training and testing. The CART model produced accuracies of 85% (2S), 83% (3SMI) and 100% (3SDI) with positive predictive values (PPV) of 100% (2S), 83% (3SMI) and 75% (3SDI) during training. It then produced accuracies of 100% (2S) and 50% (3SMI) with positive predictive values (PPV) of 74% (2S), 60% (3SMI) during testing. Thus, we then used the model to assign folding mechanisms to protein homodimers with known structures and unknown folding mechanisms. This exercise provides a framework for predicted homodimer structures with unknown folding mechanism for further verification through folding experiments. The CART model was able to assign folding mechanisms to all (169) the homodimer structures (with unknown folding data) due its automatically robust learning capabilities unlike the manually developed decision model which left some structures unassigned.

Abbreviations: 2S: 2 State; 3S: 3 State; 3SDI: 3 State Dimer Intermediate; 3SMI: 3 State Monomer Intermediate; B/2: Interface Area; CART: Classification And Regression Tree; I/T: Interface to Total residue ratio; ML: monomer length; PPV: positive predictive value

Background

Homodimers play an important role in catalysis and cellular regulation. Moreover, a couple of homodimers have been described as cancer-targets (U.S. Patent office), (Tanaka et al., 2007; Schulke et al., 2003). Thus, the importance of homodimer structures is recognized. The formation of homodimers in cellular biology is interesting and the mechanism (2-state (2S), 3-state (3S)) of folding is more fascinating (Zhanhua et al., 2005). Two-state (2S) homodimers (Zhanhua et al., 2005; Wales et al, 2004; Bowie et al., 1989; Milla and Sauer, 1994; Steif et al., 1993; Jana et al., 1997; Topping et al., 2004; Stone et al., 2002; Grant et al., 1992; Bajaj et al., 2004; Kretschmar and Jaenicke, 1999; Johnson et al, 1992; Tamura et al., 1995; Gloss et al., 2001; Timm et al., 1994; Li et al., 1997; Kim et al., 2001; D'Alfonso et al., 2002; Dirr and Reinemer, 1991; Wallace et al., 1998; Kaplan et al., 1997; Ahmad et al.,1998; Mainfroid et al., 1996; Yang et al., 1994) fold without the formation of a stable intermediate. Three state (3S) homodimers fold with the formation of a stable dimeric (3SDI), (Ramstein et al., 2003; Zhu et al., 2003; Grimsley et al., 1997; Clark et al., 1993; Motono et al.,1999; Mei et al., 1997; Doyle et al., 2000) or monomeric (3SMI) intermediate (Mateu, 2002; Ruller et al., 2003; Apiyo et al., 2001; Malvezzi-Campeggi et al., 1999; Stroppolo et al., 2000; Malecki et al., 1997; Aceto et al.,1992; Gokhale et al., 1996; Park and Bedouelle, 1998; Wójciak et al., 2003; Liang et al., 2003). The folding data is obtained by denaturation techniques involving thermal and chemical agents. The denatured fraction is studied by CD, NMR and absorption. But these experiments are very time consuming and tedious. Thus, folding data is known only for a few homodimers, although large number of structures is known and available at the protein databank (PDB). Denaturation experiments (using temperature and chemical agents), although tedious to perform, have played a vital role in understanding the structural architecture and folding pattern of homodimers.

A review of unfolding data of homodimers by Neet and Timm (1994) showed that some homodimers denature by a two-state equilibrium transition (2S) while others have stable intermediates in the process (3S). In addition, the conformational stability of these homodimers was found to be related to the size of the polypeptide and the nature of subunit interface. Li et al. (2005) identified structural parameters (based on 41 homodimer structures) for the classification of homodimers into 2S and 3S. The cluster of small-sized proteins with large interface area and high interface hydrophobicity were found to be 2S homodimers, while the cluster of large proteins with small interface area and low interface hydrophobicity were found

***Corresponding author:** Abishek suresh, Faculty of Applied Sciences, Department of Biotechnology, AIMST University, Semeling, Kedah, Malaysia; E-mail: abisheksuresh@gmail.com

Accepted October 19, 2010; **Published** October 21, 2010

Citation: Suresh A, Lalitha P, Kangueane P (2010) CART Assignment of Folding Mechanisms to Homodimers with Known Structures. J Proteomics Bioinform 3: 279-285. doi:10.4172/jpb.1000152

to belong to the 3S category. This study shows the importance of structural features such as monomer protein size (ML) and interface area (B/2) for distinguishing 2S from 3S. Tsai et al. (1997) studied 187 stable and 57 symmetrically related oligomeric interfaces. The architecture of 2S interfaces was found to be similar to protein cores, where the monomer chains fold cooperatively. On the other hand, 3S interfaces were found to resemble binding of already folded proteins and mirrored the monomer architecture only in general outline.

Levy et al. (2004) suggested that the native protein 3D structure is the major factor governing the choice of homodimer folding and binding mechanism in 11 homodimers with known unfolding data. The study showed that as far as protein folding is concerned, the protein topology is the main factor governing the protein binding mechanism and the degree of topological frustration of a monomer determines whether the binding will occur between two unfolded or folded chains. Mei et al. (2005), defined Interface amino acid residue (IAR - distance between the first and last amino acid that take part in the inter-subunit interaction) and squared loop length (SLL - sum of the squared distances between two successive residues of the monomer) in 32 homodimer structures to find a possible correlation between protein size, sequence and quaternary structure. They propose that medium-sized proteins of classes A (2S) and C (3SDI) models are highly stable due to their large IAR and SLL .

Lulu et al. (2009) used a dataset of 42 homodimers and showed that interface to total (I/T) residues ratio is large for 2S than 3S (3SMI and 3SDI). I/T values of 3S structures clustered together despite varying monomer protein size. Thus, I/T ratio was considered as an important parameter for distinguishing 2S from 3S. Karthikraja et al. (2009) created a dataset of 47 homodimers (twenty-eight 2S, twelve 3SMI and seven 3SDI) to examine the types of interfaces. 2S proteins were observed to be small sized, 3SMI were medium sized, while 3SDI proteins often existed as large-sized proteins. I/T measure was also used to group 2S, 3SMI and 3SDI homodimers into categories with large I/T (>50%), moderate I/T (50-25%) and small I/T (<25%) interfaces. The study provided a 2-dimensional insight into the interaction of the interface residues, while considering 2S, 3SMI and 3SDI homodimers. Suresh et al. (2009) described a decision tree model to classify 47 homodimers whose folding data was already known (by denaturation experiments) (Wójciak et al., 2003). The model worked based on the structural parameters protein-size (ML), number of interface to total residue ratio (I/T) and interface Area (B/2) and yielded positive predictive values of 71.4%, 58.4% and 57.1% in classifying 2S, 3SMI and 3SDI respectively. The model was further drawn to predict the folding information to a set of homodimers structures whose folding information was not known. The manually set up decision model was able to establish relationship between structural features and folding mechanism despite its inability to assign some structures with folding mechanism. Thus, assignment of folding mechanism for homodimer structures is of both interest and need using simple yet robust classification models. Here, we describe the development, performance and application of a Classification and Regression Tree (CART) model based on structure derived predictive variables such as ML, I/T and B/2 for the prediction of homodimer folding mechanisms given structures.

Methodology

Homodimer structure dataset with known folding mechanism: A dataset of 46 homodimers with known folding data (Table 1) was used in this analysis to develop the CART model. The dataset consists of 2S (27), 3SMI (12), and 3SDI (7) homodimers. The predictive parameters such (ML), interface area (B/2), and ratio of interface to total residues (I/T) were calculated for each entry in the dataset (Table 2).

Monomer length: Monomer length (ML) refers to the protein length of monomers forming the homodimer complex (Table 1). The ML range for 2S (45 – 271 residues), 3SMI (72 – 381 residues) and 3SDI (90 – 835 residues) is given in (Table 2).

Interface area (B/2): Interface area was calculated using change in solvent accessible surface area (ASA) from a monomer state to a dimer state. ASA was calculated using algorithm described and implemented in the software Surface Racer 5.0 (Tsodikov et al.,

Folding Data		#	PDB ID
2S		27	2CPG; 1ARQ;1ARR; 1ROP; 5CRO; 1BFM; 1A7G; 1VQB; 1B8Z; 1ETY; 1Y7Q; 1A8G; 1SIV; 1VUB; 1HDF; 1CMB; 3SSI; 1WRP; 1BET; 1BUO; 1OH0; 1BEB; 2GSR; 1GTA; 2BQP; 1HTI; 1EE1;
3S	3SMI	12	1A43; 1QLL; 1DFX; 1YAI; 1SPD; 1RUN; 11GS; 2TDM; 1TYA; 1CVI; 1ND5; 2CRK;
	3SDI	7	1MUL; 1HQO; 1PSC; 1LUC; 1CM7; 1AOZ; 1NL3;
Total		46	

2S- 2 State folding; 3S- 3 State folding; 3SMI- 3 State folding with Monomer Intermediate; 3SDI- 3 State folding with Dimer Intermediate.

Table 1: List of PDB structures with known folding information used to develop the CART prediction model.

		ML				B/2				I/T			
Fold	#	Min	Max	Mean	SD	Min	Max	Mean	SD	Min	Max	Mean	SD
2S	27	45	271	119.4	62.4	156	2507	1442.2	566.3	6	80	38.7	18.8
3SMI	12	72	835	290.6	107.6	309	2332	1204.2	563.1	9	44	18.2	10.5
3SDI	7	90	835	397.4	236.9	1351	2317	1757.8	354.7	5	50	17.8	15.3
Total	46												

ML = monomer subunit length; Min = Minimum; Max = Maximum; SD = Standard deviation

Table 2: Characteristics features of the homodimer dataset with known folding mechanism and structural data. These parameters were used in the development of a CART model for the prediction of folding mechanism for homodimers with known structural data and unknown folding data.

Research Article
OPEN ACCESS Freely available online
doi:10.4172/jpb.1000152
JPB/Vol.3 Issue 10

2002). 2S proteins have B/2 range between 156 and 2507Å^2 and 3SMI proteins range within 309 and 2332Å^2. However 3SDI dimers are found between 1351 and 2317 Å^2.

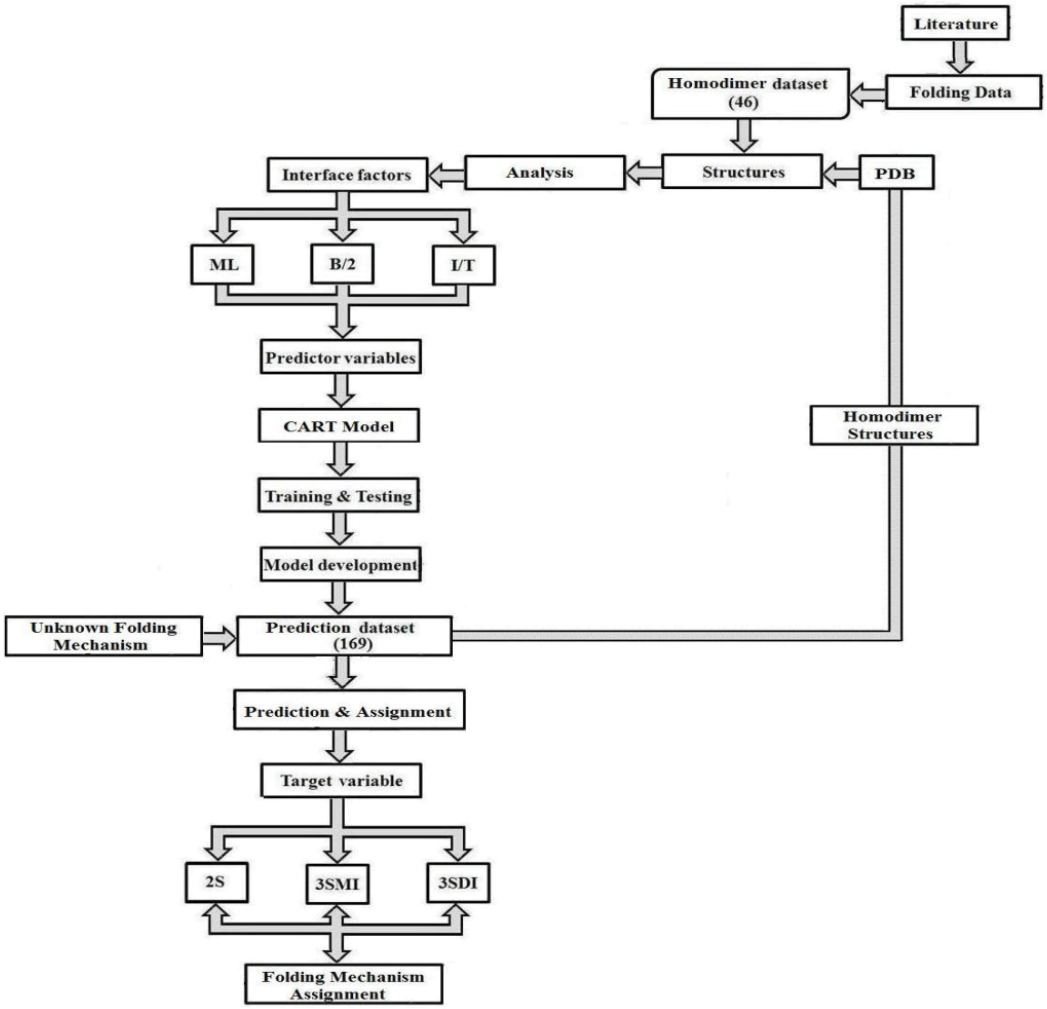
Interface to total residue (I/T) ratio: It is the ratio between interface residues (number of residues per monomer involved in homodimer interactions) to the total number of residues in the monomer protein. 3SDI proteins lie in the range of 5 to 50% and 3SMI in the range of 9 to 44%; while the 2S proteins lie in the range of 6 to 80%.

Predictor variables: The predictor variables used in model development are monomer protein size (ML), interface-area (B/2), and interface residues to total residues number ratio (I/T). The minimum, maximum, average and standard-deviation values are shown for each predictor (Table 2).

Target variables: The three target variables defined in the model are 2S, 3SMI and 3SDI.

Training & testing dataset: The 46 homodimers dataset with known folding mechanisms is arbitrarily divided into 2 subsets of 22 homodimers (thirteen 2S; six 3SMI; three 3SDI) and 24 homodimers (fourteen 2S; six 3SMI; four 3SDI) representing the training and testing set, respectively.

CART model: Classification and Regression Tree (CART) is a robust decision-tree system used for data-mining and predictive modeling (Breiman et al., 1984; Steinberg and Colla, 1997). CART is also known as Binary recursive partitioning; binary because the data (parent node) is split into two child nodes based on certain splitting conditions; it is recursive because the process is repeated with the child node acting as the parent in the subsequent iteration. The CART looks for all possible splits in the group of predictor variables. The Geni impurity criterion (probability measure of misclassification in a group of known data) is used as the splitting rule by default. During the splitting process, intermediate nodes that can no longer be split from terminal nodes. When all the active nodes become terminal nodes, the tree growing process terminates. Each node of the decision tree (including the terminal node) presents some information about a target variable or a group of target variables.



CART- Classification and Regression Tree; ML- Monomer-Length; B/2- Interface Area; I/T- ratio of number of Interface residues to total number residues in the monomer chain. The unknown dataset is a dataset of homodimer structures from PDB with unknown folding mechanism information.

Figure 1: A flowchart showing the methodology employed in the development and utility of the CART model.

The CART version 6.0 (Salford Systems, http://salford-systems.com/cart.php) is used in this study. The Target (2S, 3SMI, 3SDI) and predictor variables (ML, I/T and B/2) are defined in the CART system. Figure 1 shows a flow-chart describing the methodology employed in the development of the CART model and its implementation. CART classification and model development consists of the training and the testing phase. In the training phase, CART reads the predictor variables from a training subset of 22 homodimers and learns to classify the dataset building an overly large tree (Table 3, Table 4). CART then classifies the testing set using the overly large tree and calculates the error rate. The largest tree or sub-tree with the minimum error rate is chosen as the CART model. The results from the testing phase are given in (Table 5, Table 6).

CART prediction: The CART model is then applied to predict the folding information to a dataset of 169 homodimers whose folding information was not known (Table 7).

Results

A dataset of 46 known homodimer structures with known folding data collected from literature is given in (Table 1). The dataset of 46 homodimers is split indiscriminately into two subsets of 22 and 24 homodimers, each subset representing a uniform distribution of the protein folding types. The predictor variables (ML, B/2 and I/T) used in the model was calculated for each structure in training and testing test. (Table 2) shows the respective minimum, maximum, and average and standard deviation values of the predictors in the dataset The model uses predictor variables ML and I/T in the classification process during training and testing. Node 1 (root node) consists of the 22 homodimers. The data is split based on various conditions of the predictor variables at each level. At each intermediate node, a case went to the left child node if and only if a condition statement regarding the predictor variable was satisfied. The classification model consisted of 4 terminal nodes. Information regarding the distribution of folding cases (2S, 3SMI or 3SDI) is available at each node. Table 3 and Table 4 show the training results of the CART model. The CART classification model produced positive predictive values (PPV, ratio of true positives to the total number of true and false positives) 100%, 83.3% and 75% (with an accuracy of 84.6%, 83.4% and 100%) in classifying 2S, 3SMI and 3SDI homodimers respectively during the training phase. During the learning phase, CART trains on the training subset of 22 known homodimers, and develops a maximal tree. The model developed from training is further tested on the second subset of 24 homodimers (Table 5 and Table 6). The CART model displayed positive predictive values of 73.7%, and 60% (with an accuracy of 100%, and 50%) in testing the subset for 2S, and 3SMI, respectively. However, it was not able to perform well in predicting 3SDI during testing. The CART model is then applied to a dataset of 169 homodimers whose folding data is not known. A folding mechanism was assigned to all structures (Table 7) unlike the manually set up decision tree model as described elsewhere (Wójciak et al., 2003). The CART model classified 78 homodimers as 2S, 45 as 3SMI and 46 homodimers under 3SDI respectively. Table 7 also gives a detailed comparison of the results obtained from CART and the decision-tree studied earlier.

Fold	Training Set	CART classification		
		2S	3SMI	3SDI
2S	13	11	2	0
3SMI	6	0	5	1
3SDI	3	0	0	3

Table 3: CART classification of the training set. The CART model uses the training set for learning features of the target variable using known predictor variables and hence, uses the same for classification in internal cross validation.

Folding	Dataset	TP	TN	FP	FN	AC (%)	PPV (%)
2S	13	11	0	0	2	85	100
3SMI	6	5	0	2	1	83	83
3SDI	3	3	0	1	0	100	75

TP = true positive; TN = true negative; FP = false positive; FN = false negative; AC = accuracy; PPV = positive predictive value; AC = (TP + TN)/(TN+TP+FP+FN); PPV = TP / (TP+FP)

Table 4: Summary of results from a training experiment is given (see Table 3).

Fold	Testing Set	Prediction by CART		
		2S	3SMI	3SDI
2S	14	14	0	0
3SMI	6	3	3	0
3SDI	4	2	2	0

Table 5: CART classification of an independent testing set.

	Dataset	TP	TN	FP	FN	Accuracy (%)	PPV (%)
2S	14	14	0	3+2	0	100	74
3SMI	6	3	0	2	3	50	60
3SDI	4	0	0	0	4	0	0

TP = true positive; TN = true negative; FP = false positive; FN = false negative; AC = accuracy; PPV = positive predictive value; AC = (TP + TN)/(TN+TP+FP+FN); PPV = TP / (TP+FP)

Table 6: Summary of results from a testing experiment is given (see Table 5).

Discussion

The mechanism of homodimer folding and binding was studied based on thermal denaturation experiments using fluorescence (Bowie et al., 1989; Milla and Sauer,1994; Mok et al., 1996; Grant et al., 1992; Bajaj et al., 2004; Kretschmar and Jaenicke, 1999; Timm et al., 2001; Kim et al., 2001; D'Alfonso et al., 2002; Dirr and Reinemer, 1991; Wallace et al., 1998; Kaplan et al., 1997; Ahmad et al.,1998; Mainfroid et al., 1996; Zhu et al., 2003; Grimsley et al., 1997; Clark et al., 1993; Motono et al.,1999; Mei et al., 1997; Doyle et al., 2000; Mateu, 2002; Ruller et al., 2003; Apiyo et al., 2001; Malvezzi-Compeggi et al., 1999; Stroppolo et al., 2000; Malecki et al., 1997; Aceto et al., 1992; Gokhale et al., 1996; Wójciak et al., 2003; Liang et al., 2003), NMR (Tamura et al., 1995), Circular Dichroism (Wales et al., 2004; Bowie et al., 1989; Steif et al., 1993; Jana et al., 1997; Topping and Gloss 2004; Mok et al., 1996; Liang et al., 1991; Ruiz-Sanz et al., 2004; Topping et al., 2004; Stone et al., 2002; Bajaj et al., 2004; Li et al., 1997; Ahmad et al., 1998; Mainfroid et al., 1996; Ramstein et al., 2003; Grimsley et al., 1997; Clark et al., 1993; Motono et al., 1999; Mei et al., 1997; Doyle et al., 2000; Mateu, 2002; Ruller et al., 2003; Apiyo et al., 2001; Malvezzi-Campeggi et al., 1999; Stropopolo et al., 2000; Gokhale et al., 1996; Park et al., 1998) and absorption (Apiyo et al., 2001). Homodimers fold by three folding mechanisms (2S, 3SMI and 3SDI). Three dimensional structures for these homodimers with known folding mechanism are already available in the protein data bank (PDB). The consideration of homodimers as drug targets in cancer has been realized (Tanaka et al., 2007) (Shulke et al., 2007).Therefore, it is of importance to study homodimers binding and folding. The documentation of folding mechanisms for homodimers through denaturation experiments is tedious. Thus, folding mechanism is known only for a handful of

Research Article
OPEN ACCESS Freely available online
doi:10.4172/jpb.1000152
JPB/Vol.3 Issue 10

Folding	Count	PDB ID common to CART & Decision Tree	Count	Decision Tree Only	Count	CART Only	Assignment by Decision model	Assignment by CART
2S	41	1BH5; 1C6X; 1CDC; 1CQS; 1D1G; 1EN7; 1EXQ; 1FJH; 1F4Q; 1G0S; 1GD7; 1GGQ; 1G64; 1H8X; 1HSS; 1IOR; 1IPI; 1J30; 1JOG; 1JR8; 1K3S; 1K6Z; 1KSO; 1L5B; 1LHZ; 1LMW; 1LQ9; 1M7H; 1MKB; 1NA8; 1NWW; 1OR4; 1PP2; 1QFH; 1QR2; 1QXR; 1R5P; 1R9C; 1SMT; 1TLU; 1LR5;	1	1P60;	37	1A4U; 1AA7; 1AQ6; 1BBH; 1CBK; 1COZ; 1DQP; 1DQT; 1EAJ; 1EYV; 1FL1; 1FVD; 1G1M; 1I4S; 1JMV; 1JP3; 1L5X; 1M4I; 1M6P; 1MJH; 1N2A; 1O4U; 1ON2; 1ORO; 1OTV; 1OXB; 1PEO; 1R8J; 1RVE; 1RYA; 1S44; 1SCF; 1UC8; 1DAB; 2HHM; 3LYN; 3SDH;	42	78
3SMI	32	1A4I; 1AUO; 1BXG; 1DVJ; 1EN5; 1EOG; 1EZ2; 1F89; 1FUX; 1G1A; 1G8T; 1HJR; 1JDO; 1JFL; 1JYS; 1M13; 1MNA; 1M98; 1NFZ; 1PN2; 1Q8R; 1QMJ; 1QYA; 1REG; 8PRK; 9WGA; 1EV7; 1EWZ; 1I2W; 1N2O; 1RQL; 1Y6H;	38	1AA7; 1AOR; 1AQ6; 1BBH; 1CBK; 1COZ; 1DQP 1DQT; 1EAJ; 1EYV; 1FL1; 1FP3; 1HJ3; 1HSJ; 1I4S; 1I8T; 1JMV; 1JP3; 1JV3; 1M4I; 1M6P; 1MJH; 1N1B; 1N2A; 1NW1; 1ON2; 1ORO; 1OXB; 1PE0; 1R7A; 1RVE; 1RYA; 1S02; 1S44; 1SCF; 2SQC; 3LYN; 3SDH;	13	1AD1; 1BMD; 1D0R; 1EBL; 1F17; 1FWL; 1KGN; 1KIY; 1LBQ; 1LHP; 1N80; 1P60; 1QHI;	70	45
3SDI	22	1ADE; 1AFW; 1BJW; 1CNZ; 1DPG; 1EHI; 1EKP; 1F13; 1FCS; 1M3E; 1M9K; 1NU6; 1P3W; 1SOX; 1TRK; 7AAT; 1BDO; 1F6D; 1HDY; 1M0W; 1S2Q; 1V26;	17	1ADI; 1BMD; 1D0R; 1F17; 1EBL; 1FWL; 1G1M; 1KIY; 1LBQ; 1LHP; 1N80; 1O4U; 1OTV; 1QHI; 1UC8; 2DAB; 2HHM;	24	1ALK; 1AOR; 1CHM; 1FP3; 1HJ3; 1HSJ; 1I8T; 1IRI; 1JV3; 1K75; 1LK9; 1N1B; 1NW1; 1NY5; 1OAC; 1P43; 1PJQ; 1PN0; 1PT5; 1R7A; 1S02; 2GSA; 2NAC; 2SQC;	39	46
Unassigned							**18**	**0**
Total	**95**		**56**		**74**		**169**	**169**

Table 7: Assignment of folding mechanism to homodimers with known structures & unknown folding. The classification is compared with those of the Decision model discussed earlier (Suresh *et. al*).

such homodimers. A comprehensive literature survey identified 46 homodimer structures with known folding mechanism (Table 1). However, hundreds of structures are available in PDB with unknown folding mechanism (Table 7). Therefore, it is of interest to develop a CART based prediction model to assign folding mechanism to homodimers structures with unknown folding mechanisms. Suresh et al. (2009) developed a decision-model based on predictive structural parameters (ML, B/2, I/T). This model yielded positive predictive values of 71.4%, 58.4% and 57.1% in classifying 2S, 3SMI and 3SDI, respectively in a known dataset Application of this model to a dataset (169 structures) with unknown folding data resulted in 18 unassigned structures. This is due to the inability of the manually set-up decision model to assign function to the 18 structures. This then created interest to develop an automatically robust method for this purpose. Classification and Regression Tree (CART) was found as an alternative method for this application due to its robust learning capabilities from learning (training) set.

We describe the performance of the CART model in assigning folding mechanisms to homodimer structures. The dataset of 46 homodimers (whose folding data is known) is indiscriminately split into two subsets of twenty-two (22) and twenty-four (24) homodimers. One subset acts as the training set (13-2S; 6-3SMI; 3-3SDI) while the other acts as a testing set (14-2S; 6-3SMI; 4-3SDI). The CART algorithm develops an overly large classification-tree based on structural parameters I/T ratio and Interface Area (B/2) from the training set (Table 3 and Table 4) shows the results of the training phase, wherein 2/13 of 2S homodimers are misclassified as 3SMI, 1/6 of 3SMI homodimers have been misclassified as 3SDI, while no misclassification is seen in 3SDI. The model thus created was found to yield positive predictive values 100%, 83% and 75% (with an accuracy of 85%, 83% and 100%) in learning to classify 2S, 3SMI and 3SDI homodimers respectively. Thus, an overly-large tree is grown

by the end of the training phase. CART classifies the test sample to determine the misclassification rate of the largest tree or every sub-tree developed in the training phase. The tree or sub-tree with the lowest misclassification rate is chosen as the CART model. The testing-phase CART showed positive predictive values of 74%, and 60% (with an accuracy of 100%, and 50%) in classifying 2S, and 3SMI homodimers respectively (Tables 5 and Table 6). However, the CART model was not able to classify the 3SDI appropriately. The CART was found to perform better than the decision model in classifying 2S and 3SMI.

The CART model was then applied to a larger dataset of 169 homodimers whose folding data was unknown. CART was able to assign folding information (target-variable) to all the unknown homodimers when the structural data from the dataset was passed through the classification model (Table 7). The decision-model was unable to fit 18 structures unlike the CART model. This was because the parameter values of these unassigned homodimers did not fall into any of the condition values set by the decision-tree. CART categorized 78, 45 and 46 homodimers of the dataset were into 2S, 3SMI and 3SDI homodimers respectively, while the decision model grouped 42, 70 and 39 homodimers into 2S, 3SMI and 3SDI, respectively. These results not only emphasize the importance of the structural features in the prediction of homodimer folding, but also prove CART to be a robust and efficient method for classification and prediction of features in multi-parameter datasets, compared to its predecessor. Further, the predicted data serves as a framework for better understanding of the folding mechanism given their structures. It should be kept in mind that these predicted results should be further verified using denaturation experiments.

Conclusion

Homodimer proteins fold through 3 different mechanisms 2S, 3SMI and 3SDI and are grouped accordingly. The presence of homodimers with unknown folding data emphasizes the need for a more efficient fold classification and prediction method. It was of interest to use the CART system in development of a classification model and further predict the folding information to an unknown dataset of 169 homodimers. The model performed fairly well for predicting 2S and 3SMI in both during training and testing using ML and I/T as predictive variables. However, it should be noted that the performance of model in classifying 3SDI is not discriminative in nature. Nonetheless, the model was not stable with the inclusion of the predictive variable B/2 and hence, was not considered further for prediction during training and testing. The CART model produced accuracies of 85% (2S), 83% (3SMI) and 100% (3SDI) with positive predictive values (PPV) of 100% (2S), 83% (3SMI) and 75% (3SDI) during training. It then produced accuracies of 100% (2S) and 50% (3SMI) with positive predictive values (PPV) of 74% (2S), 60% (3SMI) during testing. Thus, we then used the model to assign folding mechanisms to protein homodimers with known structures and unknown folding mechanisms. This exercise provides a framework for predicted homodimer structures with unknown folding mechanism for further verification through folding experiments. The CART model was able to assign folding mechanisms to all (169) the homodimer structures (with unknown folding data) due to its automatically robust learning capabilities unlike the manually developed decision model which left some structures unassigned.

Author's Contribution

PK conceived the idea. AS designed the experiment and performed the analysis with summarized results. PL participated in the analysis and helped in manuscript preparation.

References

1. Aceto A, Caccuri AM, Sacchetta P, Bucciarelli T, Dragani B, et al. (1992) Dissociation and unfolding of Pi-class glutathione transferase. Evidence for a monomeric inactive intermediate. Biochem J 285: 241-245.
2. Ahmad N, Srinivas VR, Reddy GB, Surolia A (1998)Thermodynamic characterization of the conformational stability of the homodimeric protein, pea lectin. Biochemistry 37: 16765-16772.
3. Apiyo D, Jones K, Guidry J, Wittung-Stafshede P (2001) Equilibrium unfolding of dimeric desulfoferrodoxin involves a monomeric intermediate: iron cofactors dissociate after polypeptide unfolding. Biochemistry 40: 4940-4948.
4. Bajaj K, Chakshusmathi G, Bachhawat-Sikder K, Surolia A, Varadarajan R (2004) Thermodynamic characterization of monomeric and dimeric forms of CcdB (controller of cell division or death B protein). Biochem J 380: 409-417.
5. Bowie JU, Sauer RT (1989) Equilibrium dissociation and unfolding of the Arc repressor dimer. Biochemistry 28: 7139-7143.
6. Breiman L, Friedman J, Olshen R, Stone C (1984) Classification and Regression Trees. Pacific Grove: Wadsworth.
7. Clark AC, Sinclair JF, Baldwin TO (1993) Folding of bacterial luciferase involves a non-native heterodimeric intermediate in equilibrium with the native enzyme and the unfolded subunits. J Biol Chem 268: 10773-9.
8. D'Alfonso L, Collini M, Baldini G (2002) Does beta-lactoglobulin denaturation occur via an intermediate state? Biochemistry 41: 326-333.
9. Dirr HW, Reinemer P (1991) Equilibrium unfolding of class pi glutathione S-transferase. Biochem Biophys Res Commun 180: 294-300.
10. Doyle SM, Braswell EH, Teschke CM (2000) SecA folds via a dimeric intermediate. Biochemistry 39: 11667-11676.
11. Gloss LM, Simler BR, Matthews CR (2001) Rough energy landscapes in protein folding: dimeric E. coli Trp repressor folds through three parallel channels. J Mol Biol 312: 1121-1134.
12. Gokhale RS, Agarwalla S, Santi DV, Balaram P (1996) Covalent reinforcement of a fragile region in the dimeric enzyme thymidylate synthase stabilizes the protein against chaotrope-induced unfolding. Biochemstry 35: 7150-7158.
13. Grant SK, Deckman IC, Culp JS, Minnich MD, Brooks IS, et al. (1992) Use of protein unfolding studies to determine the conformational and dimeric stabilities of HIV-1 and SIV proteases. Biochemistry 31: 9491-9501.
14. Grimsley JK, Scholtz JM, Pace CN, Wild JR (1997) Organophosphorus hydrolase is a remarkably stable enzyme that unfolds through a homodimeric intermediate. Biochemistry 36: 14366-74.
15. Jana R, Hazbun TR, Mollah AK, Mossing MC (1997) A folded monomeric intermediate in the formation of lambda Cro dimer-DNA complexes. J Mol Biol 273: 402-416.
16. Johnson CM, Cooper A, Stockley PG (1992) Differential scanning calorimetry of thermal unfolding of the methionine repressor protein (MetJ) from Escherichia coli. Biochemistry 31: 9717-9724.
17. Kaplan W, Hüsler P, Klump H, Erhardt J, Sluis-Cremer N, et al. (1997) Conformational stability of pGEX-expressed Schistosoma japonicum glutathione S-transferase: a detoxification enzyme and fusion-protein affinity tag. Protein Sci 6: 399-406.
18. Karthikraja V, Suresh A, Lulu S, Kangueane U, Kangueane P (2009) Types of interfaces for homodimer folding and binding. Bioinformation 4: 101-11.
19. Kim D, Nam GH, Jang DS, Yun S, Choi G, et al. (2001) Roles of dimerization in folding and stability of ketosteroid isomerase from Pseudomonas putida biotype B. Protein Sci 10: 741-752.
20. Kretschmar M, Jaenicke R (1999) Stability of a homo-dimeric Ca(2+)-binding member of the beta gamma-crystallin superfamily: DSC measurements on spherulin 3a from Physarum polycephalum. J Mol Biol 291: 1147-1153.
21. Levy Y, Wolynes PG, Onuchic JN (2004) Protein topology determines binding mechanism. Proc Natl Acad Sci 101: 511-516.
22. Li L, Gunasekaran K, Gan JG, Zhanhua C, Shapshak P, et al. (2005) Structural features differentiate the mechanisms between 2S (2 state) and 3S (3 state) folding homodimers. Bioinformation 1: 42-49.
23. Li X, Lopez-Guisa JM, Ninan N, Weiner EJ, Rauscher FJ 3rd, et al. (1997) Over-expression, purification, characterization, and crystallization of the BTB/POZ domain from the PLZF oncoprotein. J Biol Chem 272: 27324-27329.

Research Article
OPEN ACCESS Freely available online
doi:10.4172/jpb.1000152
JPB/Vol.3 Issue 10

24. Liang H, Terwilliger TC (1991) Reversible denaturation of the gene V protein of bacteriophage f1. Biochemistry 30: 2772-2782.

25. Liang Y, Du F, Sanglier S, Zhou BR, Xia Y, et al. (2003) Unfolding of rabbit muscle creatine kinase induced by acid. A study using electrospray ionization mass spectrometry, isothermal titration calorimetry, and fluorescence spectroscopy. J Biol Chem 278: 30098-30105.

26. Lulu S, Suresh A, Karthikraja V, Arumugam M, Kayathri R, et al. (2009) Structural features for homodimer folding mechanism. J Mol Graph Model 28: 88-94.

27. Mainfroid V, Terpstra P, Beauregard M, Frère JM, Mande SC, et al. (1996) Three hTIM mutants that provide new insights on why TIM is a dimer. J Mol Biol 257: 441-456.

28. Malecki J, Wasylewski Z (1997) Stability and kinetics of unfolding and refolding of cAMP receptor protein from Escherichia coli. Eur J Biochem 243: 660-669.

29. Malvezzi-Campeggi F, Stroppolo ME, Mei G, Rosato N, Desideri A (1999) Evidence of stable monomeric species in the unfolding of Cu,Zn superoxide dismutase from Photobacterium leiognathi. Arch Biochem Biophys 370: 201-207.

30. Mateu MG (2002) Conformational stability of dimeric and monomeric forms of the C-terminal domain of human immunodeficiency virus-1 capsid protein. J Mol Biol 318: 519-531.

31. Mei G, Di Venere A, Buganza M, Vecchini P, Rosato N, et al. (1997) Role of quaternary structure in the stability of dimeric proteins: the case of ascorbate oxidase. Biochemistry 36: 10917-10922.

32. Mei G, Di Venere A, Rosato N, Finazzi-Agrò A (2005) The importance of being dimeric. Febs J 272: 16-27.

33. Milla ME, Sauer RT (1994) P22 Arc repressor: folding kinetics of a single-domain, dimeric protein. Biochemistry 33: 1125-1133.

34. Mok YK, de Prat Gay G, Butler PJ, Bycroft M (1996) Equilibrium dissociation and unfolding of the dimeric human papillomavirus strain-16 E2 DNA-binding domain. Protein Sci 5: 310-319.

35. Motono C, Yamagishi A, Oshima T (1999) Urea-induced unfolding and conformational stability of 3-isopropylmalate dehydrogenase from the Thermophile thermus thermophilus and its mesophilic counterpart from Escherichia coli. Biochemistry 38: 1332-1337.

36. Neet KE, Timm DE (1994) Conformational stability of dimeric proteins: quantitative studies by equilibrium denaturation. Protein Sci 3: 2167-2174.

37. Park YC, Bedouelle H (1998) Dimeric tyrosyl-tRNA synthetase from Bacillus stearothermophilus unfolds through a monomeric intermediate. A quantitative analysis under equilibrium conditions. J Biol Chem 273: 18052-18059.

38. Ramstein J, Hervouet N, Coste F, Zelwer C, Oberto J, et al. (2003) Evidence of a thermal unfolding dimeric intermediate for the Escherichia coli histone-like HU proteins: thermodynamics and structure. J Mol Biol 331: 101-121.

39. Ruiz-Sanz J, Filimonov VV, Christodoulou E, Vorgias CE, Mateo PL (2004) Thermodynamic analysis of the unfolding and stability of the dimeric DNA-binding protein HU from the hyperthermophilic eubacterium Thermotoga maritima and its E34D mutant. Eur J Biochem 271: 1497-1507.

40. Ruller R, Ferreira TL, de Oliveira AH, Ward RJ (2003) Chemical denaturation of a homodimeric lysine-49 phospholipase A2: a stable dimer interface and a native monomeric intermediate. Arch Biochem Biophys 411: 112-120.

41. Schülke N, Varlamova OA, Donovan GP, Ma D, Gardner JP, et al. (2003) The homodimer of prostate-specific membrane antigen is a functional target for cancer therapy. Proc Natl Acad Sci 100: 12590 -12595.

42. Steif C, Weber P, Hinz HJ, Flossdorf J, Cesareni G, et al. (1993) Subunit interactions provide a significant contribution to the stability of the dimeric four-alpha-helical-bundle protein ROP. Biochemistry 32: 3867-3876.

43. Steinberg D, Colla P (1997) CART-Classification and Regression Trees. San Diego, CA, Salford Systems.

44. Stone JR, Maki JL, Blacklow SC, Collins T (2002) The SCAN domain of ZNF174 is a dimer. J Biol Chem 277: 5448-5452.

45. Stroppolo ME, Malvezzi-Campeggi F, Mei G, Rosato N, Desideri A (2000) Role of the tertiary and quaternary structures in the stability of dimeric copper, zinc superoxide dismutases. Arch Biochem Biophys 377: 215-218.

46. Suresh A, Karthikraja V, Lulu S, Kangueane U, Kangueane P (2009) A decision tree model for the prediction of homodimer folding mechanism. Bioinformation 4: 197-205.

47. Tamura A, Kojima S, Miura K, Sturtevant JM (1995) A thermodynamic study of mutant forms of Streptomyces subtilisin inhibitor. II. Replacements at the interface of dimer formation, Val13. J Mol Biol 249: 636-645.

48. Tanaka T, Suh KS, Lo AM, De Luca LM (2007) p21WAF1/CIP1 is a common transcriptional target of retinoid receptors: pleiotropic regulatory mechanism through retinoic acid receptor (RAR)/retinoid X receptor (RXR) heterodimer and RXR/RXR homodimer. J Biol Che 282: 29987-29997.

49. The United States Patent and Trademark Office database, USA.

50. Timm DE, de Haseth PL, Neet KE (1994) Comparative equilibrium denaturation studies of the neurotrophins: nerve growth factor, brain-derived neurotrophic factor, neurotrophin 3, and neurotrophin 4/5. Biochemistry 33: 4667-4676.

51. Topping TB, Gloss LM (2004) Stability and folding mechanism of mesophilic, thermophilic and hyperthermophilic archael histones: the importance of folding intermediates. J Mol Biol 342: 247-260.

52. Topping TB, Hoch DA, Gloss LM (2004) Folding mechanism of FIS, the intertwined, dimeric factor for inversion stimulation. J Mol Biol 335: 1065-81.

53. Tsai CJ, Xu D, Nussinov R (1997) Structural motifs at protein-protein interfaces: protein cores versus two-state and three-state model complexes. Protein Sci 6: 1793-1805.

54. Tsodikov OV, Record MT, Sergeev YV (2002) Novel computer program for fast exact calculation of accessible and molecular surface areas and average surface curvature. J Comput Chem 23: 600-609.

55. Wales TE, Richardson JS, Fitzgerald MC (2004) Facile chemical synthesis and equilibrium unfolding properties of CopG. Protein Sci 13: 1918-1926.

56. Wallace LA, Sluis-Cremer N, Dirr HW (1998) Equilibrium and kinetic unfolding properties of dimeric human glutathione transferase A1-1. Biochemistry 37: 5320-5328.

57. Wójciak P, Mazurkiewicz A, Bakalova A, Kuciel R (2003) Equilibrium unfolding of dimeric human prostatic acid phosphatase involves an inactive monomeric intermediate. Int J Biol Macromol 32: 43-54.

58. Yang ZW, Tendian SW, Carson WM, Brouillette WJ, Delucas LJ, et al. (1994) Dimethyl sulfoxide at 2.5% (v/v) alters the structural cooperativity and unfolding mechanism of dimeric bacterial NAD+ synthetase. Protein Sci 13: 830-841.

59. Zhanhua C, Gan JG, Lei L, Sakharkar MK, Kangueane P (2005) Protein subunit interfaces: heterodimers versus homodimers. Bioinformation 1: 28-39.

60. Zhu L, Zhang XJ, Wang LY, Zhou JM, Perrett S (2003) Relationship between stability of folding intermediates and amyloid formation for the yeast prion Ure2p: a quantitative analysis of the effects of pH and buffer system. J Mol Biol 328: 235-254.