Bioinformation
www.bioinformation.net

open access
Hypothesis

A decision tree model for the prediction of homodimer folding mechanism

Abishek Suresh [1, 2], Velmurugan Karthikraja [1], Sajitha Lulu [1], Uma Kangueane [1], Pandjassarame Kangueane [1, 2,*]

[1] Biomedical Informatics, Pondicherry 607402, [2]AIMST University, Semeling 08100, Malaysia; Pandjassarame Kangueane - E-mail: kangueane@bioinformation.net; Phone: +91 413 2633 589; Fax: +91 413 2633 722; *Corresponding author

Received October 20, 2009; Accepted November 09, 2009; Published November 17, 2009

Abstract:
The formation of protein homodimer complexes for molecular catalysis and regulation is fascinating. The homodimer formation through 2S (2 state), 3SMI (3 state with monomer intermediate) and 3SDI (3 state with dimer intermediate) folding mechanism is known for 47 homodimer structures. Our dataset of forty-seven homodimers consists of twenty-eight 2S, twelve 3SMI and seven 3SDI. The dataset is characterized using monomer length, interface area and interface/total (I/T) residue ratio. It is found that 2S are often small in size with large I/T ratio and 3SDI are frequently large in size with small I/T ratio. Nonetheless, 3SMI have a mixture of these features. Hence, we used these parameters to develop a decision tree model. The decision tree model produced positive predictive values (PPV) of 72% for 2S, 58% for 3SMI and 57% for 3SDI in cross validation. Thus, the method finds application in assigning homodimers with folding mechanism.

Keywords: folding, homodimer, decision tree, prediction, mechanism

Background:
Homodimers play an important role in catalysis and regulation. The formation of homodimer interface is structurally intriguing **[1]**. The mechanism of formation of such homodimer interfaces is further appealing. Structures for 47 homodimers with known folding information are now available as given in **Table 1 (supplementary material) [2-46]**. These homodimers are formed through 2-sate (2S) **[2-28]**, or 3-state with monomer intermediate (3SMI) **[36-46]** or 3-state with dimer intermediate (3SDI) **[29-35]**. A couple of homodimers have been described as cancer targets **[47, 48, 49]**. Hence, the future definition of homodimers as drug targets is evident. Therefore, it is important to understand both homodimer association and its folding mechanism of formation. A number of attempts have been made to relate homodimer structures to folding mechanism to decipher folding specific structural features **[50-54]**. We recently documented the relationship between structural features describing homodimer folding mechanism **[55]**. Nevertheless, folding information on homodimers is far less than the known number of homodimer structures stored in databases **[1]**. Therefore, it is of interest to predict folding mechanism to known homodimer structures. We created an improved dataset of 47 homodimer structures from PDB with known folding mechanism to glean parameters and to develop models for homodimer folding mechanism prediction given their structures. We then use these parameters to design a decision tree model to classify homodimer structures with unknown folding mechanism.

Methodology:
Dataset:
We created a dataset of 47 homodimer structures from PDB with known folding information taken from respective literature **(Table 1 in supplementary material)**. The dataset consists of twenty eight 2S, twelve 3SMI and seven 3SDI structures. **Table 1 (see supplementary material)** also provides information on structural parameters such as monomer length (ML), interface area (B/2) and interface to total residue (I/T) ratio for each structure. The structural features in the dataset are summarized in **Table 2 (see supplementary material)**.

Monomer length (ML):
Monomer length (ML) refers to the protein length of monomers forming the homodimer complex. The distribution of 2S, 3SMI and 3SDI with ML is shown in **Figure 1a**. The figure illustrates the minimum and maximum limits of ML for 2S, 3SMI and 3SDI homodimers in the dataset. The length of 2S proteins are found in the range of 45 to 271, 3SMI in the range of 72 and 381, while 3SDI between 90 and 835. There is some degree of ML overlap between the three categories of homodimers.

Interface area (B/2):
Interface area (B/2) is defined as the change in accessible surface area (delta ASA) when going from monomer state to dimer state during complex formation. Accessible surface area (ASA) is calculated using the software SURFACE RACER 5.0 **[56]** using the algorithm described by Lee and Richard **[56]**. The distribution of 2S, 3SMI and 3SDI with B/2 is shown in **Figure 1b**. The figure shows the graphical representation of homodimers according to their interface area. 2S proteins have B/2 range between 156 -2507 $Å^2$ and 3SMI proteins range within 309 and 2317 $Å^2$. However, 3SDI dimers lie between 1351 and 2317 $Å^2$.

Interface to total residue (I/T) ratio:
It is the ratio between the numbers of interface residues per monomer (residues involved in homodimer interactions at the interface) to the total number of residues in monomer protein. Interface residues are identified using ASA calculation described in previous section. The distribution of 2S, 3SMI and 3SDI with I/T ratio is shown in **Figure 1c.** The figure shows the graphical representation of homodimers to I/T ratio. Here, the 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%.

Decision tree model:
A decision model is a clear logical model that can be easily understood by persons who are not mathematically inclined. The decision tree model is a classification tree to classify the target variable (folding mechanism in this case) based on the predictor variables (ML, B/2 and I/T) described in previous sections. The cumulative frequencies of the three predictors (ML, B/2 and I/T) were used to decide the values in the logical conditions of the decision tree. A flowchart describing the decision tree model is illustrated in **Figure 3**. The model checks for ML, I/T and B/2 for each known homodimer structures to assign their folding mechanism using human expert cut-off values as shown in **Figure 3**.

Validation:
An internal cross validation is performed for 47 homodimers in Table 1 using the decision tree model described above. The results of the validation using true positive (TP), false positive (FP) and positive predictive value (PPV) is given in Table 5. PPV (%) is defined as TP/(TP+FP)*100.

Assignment dataset:
We created a dataset of 149 homodimers with unknown folding information for prediction and assignment of folding mechanism using structural parameters **(Table 3 in supplementary material)**. The structural features in the dataset are summarized in **Table 4 (see supplementary material)**. A classification of 149 homodimers into three target categories using the decision tree model is given in **Table 6 (see supplementary material)**.

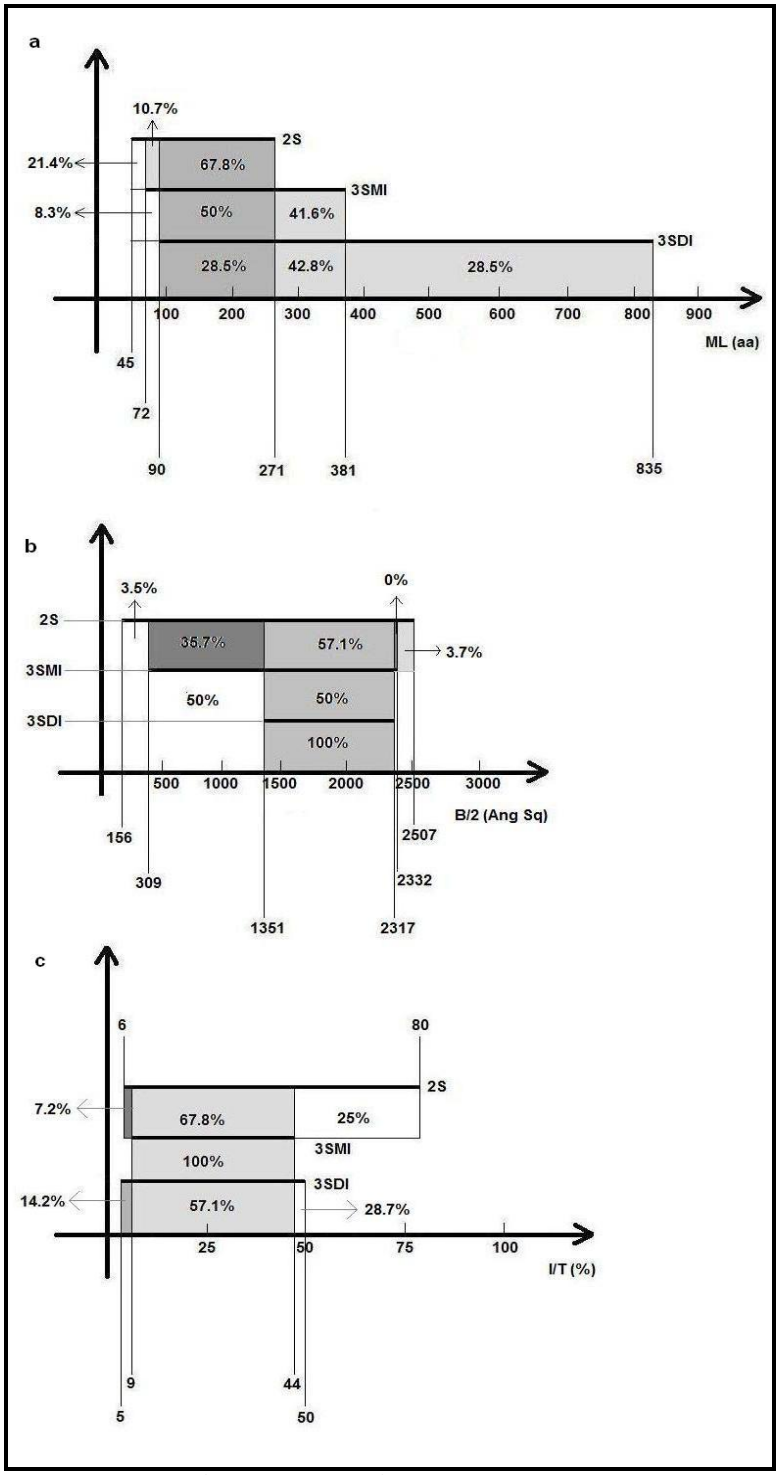

Figure 1: Distribution of 2S, 3SMI and 3SDI for ML, B/2 and I/T is shown. (a) An illustration of the minimum and maximum limits of ML for 2S, 3SMI and 3SDI homodimers in the dataset is presented. The X – axis represents monomer length. The overlap regions are shown horizontally. 2S proteins range from 45 to 271, 3SMI range from 72 to 381 and 3SDI range from 90 to 835. (b) An illustration of the minimum and maximum limits of ML for 2S, 3SMI and 3SDI homodimers in the dataset is presented. The X – axis represents interface area. The overlap regions are shown horizontally. 2S proteins range from 156 to 2507, 3SMI range from 309 to 2332 and 3SDI range from 1351 to 2317. (c) Distribution of 2S, 3SMI and 3SDI for I/T ratio. An illustration of the minimum and maximum limits of I/T for 2S, 3SMI and 3SDI homodimers in the dataset is presented. The X – axis represents I/T ratio. The overlap regions are shown horizontally. 2S proteins range from 6 to 80, 3SMI range from 9 to 44 and 3SDI range from 5 to 50. It should be noted that there is no Y-axis variable defined in this case. However, a Y-axis is shown for convenience of visualization.

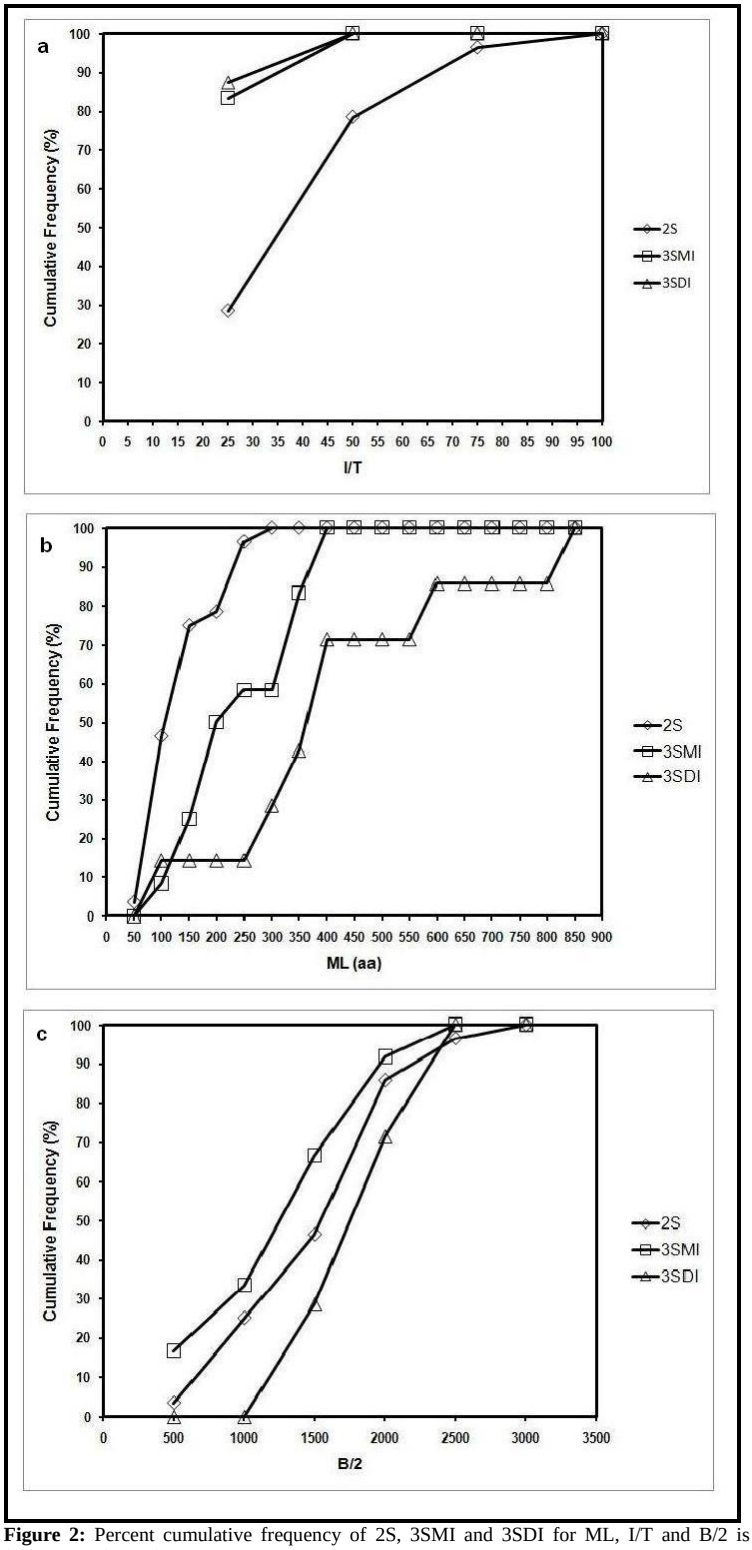

Figure 2: Percent cumulative frequency of 2S, 3SMI and 3SDI for ML, I/T and B/2 is given. (a) The distribution of the cumulative frequency of ML for 2S, 3SMI and 3SDI homodimers in the dataset is presented. About 90% of 2S, 60% of 3SMI and 15% of 3SDI are covered when ML <= 250. Hence, ML <=250 was selected as a decision condition in the development of the model. (b) The distribution of the cumulative frequency of I/T ratio for 2S, 3SMI and 3SDI homodimers in the dataset is presented. About 30% of 2S and 90% of 3SMI and 3SDI are covered when I/T <= 25%. Hence, I/T <=25% was selected as a decision condition in the development of the model. (c) The distribution of the cumulative frequency of interface area for 2S, 3SMI and 3SDI homodimers in the dataset is presented. About 50% of 2S, 70% of 3SMI and 30% of 3SDI are covered when B/2 <= 1500. Hence, B/2 <= 1500 was selected as a decision condition in the development of the model.

Discussion:

Protein homodimer molecules have been defined as drug targets in cancer **[48-49]**. Thus, homodimers have commercial importance in drug discovery. The different folding mechanisms associated with homodimers are interesting and their study is often attractive. Homodimer denaturation experiments using fluorescence **[3, 4, 8, 13 -15, 19, 21-27, 30-43, 45, 46]**, circular dichroism **[2, 3, 5-12, 14, 20, 26, 27, 29, 31-40, 43, 44]**, NMR **[18]** and adsorption **[38]** have been used to establish folding mechanism (2S, 3SMI, 3SDI) for a list of homodimers given in **Table 1 (see supplementary material)**. This is time consuming, laborious and tedious. The number of homodimer structures with unknown folding mechanism is substantial **[1]**. Therefore, it is of interest to predict homodimer folding mechanism given their 3dimenisonal structures. A number of studies have been documented to relate folding and structural features **[50-54]**. We recently described the trends in parameters (monomer size, interface residues, interface area, hydrophobicity factor, hydrophilic residues and charged residues) for distinguishing 2S from 3S proteins **[55]**. However, no attempt has been made to predict their folding mechanism given their structures in complex state. Here, we describe a novel decision tree model using predictors ML, B/2 and I/T to predict folding mechanism (target variable) given their structures in complex state.

The decision tree model is developed based on the prevalence of weight associated with these predictors in a dataset of structures with known folding data **(Figure 1)**. The distribution of its percent cumulative frequency of predictor variables in the datasets are given in **Figures 2**. **Figure 2a** gives percent cumulative frequency of 2S, 3SMI and 3SDI for ML. More than 90% of 2S lie when ML <= 250. When ML = 250 only about 15% of 3SDI and 60% of 3SMI are covered. Hence, ML <=250 was selected as a decisive condition in the development of the model. **Figure 2b** gives percent cumulative frequency of 2S, 3SMI and 3SDI for I/T ratio. About 90% of 3SMI and 3SDI lie when I/T <= 25%. When I/T <= 25%, only about 30% of 2S is covered. Therefore, I/T <=25% was selected as a decision condition in the development of the model. **Figure 2c** gives percent cumulative frequency of 2S, 3SMI and 3SDI for B/2. When B/2 <= 1500, about 70% 3SMI, 50% 2S and 30% 3SDI are covered. So, B/2 <= 1500 was selected as a decision condition in the development of the model. Thus, percent cumulative frequency values for predictors are used in the design and development of the decision tree model **(Figure 3)**. The conditional values of the predictor variables are selected based on their biased cumulative frequency in the target categories (datasets). The decision tree model checks for predictor values within defined conditional values for multiple variables in a subsequent manner sequentially so as to reach the respective nodes to predict and assign target variables.

The decision tree model was applied to classify the dataset of 47 homodimers (with known folding data) in a cross validation experiment. The model produced the positive predictive values (PPV) 71.4%, 58.4% and 57.1% for 2S, 3SMI and 3SDI, respectively **(Table 5 in supplementary material)**. We then extended the application of the decision tree model to a dataset of 149 homodimers with no folding data known. The model was able to assign folding data to 132 (88.5%) of 149 structures to predicted target variables with only 17 structures unable to classify **(Table 6 in supplementary material)**. This predicted data serves a framework to understand their folding mechanism given their structures. It should be noted that these predicted mechanism should be verified using denaturation experiments.

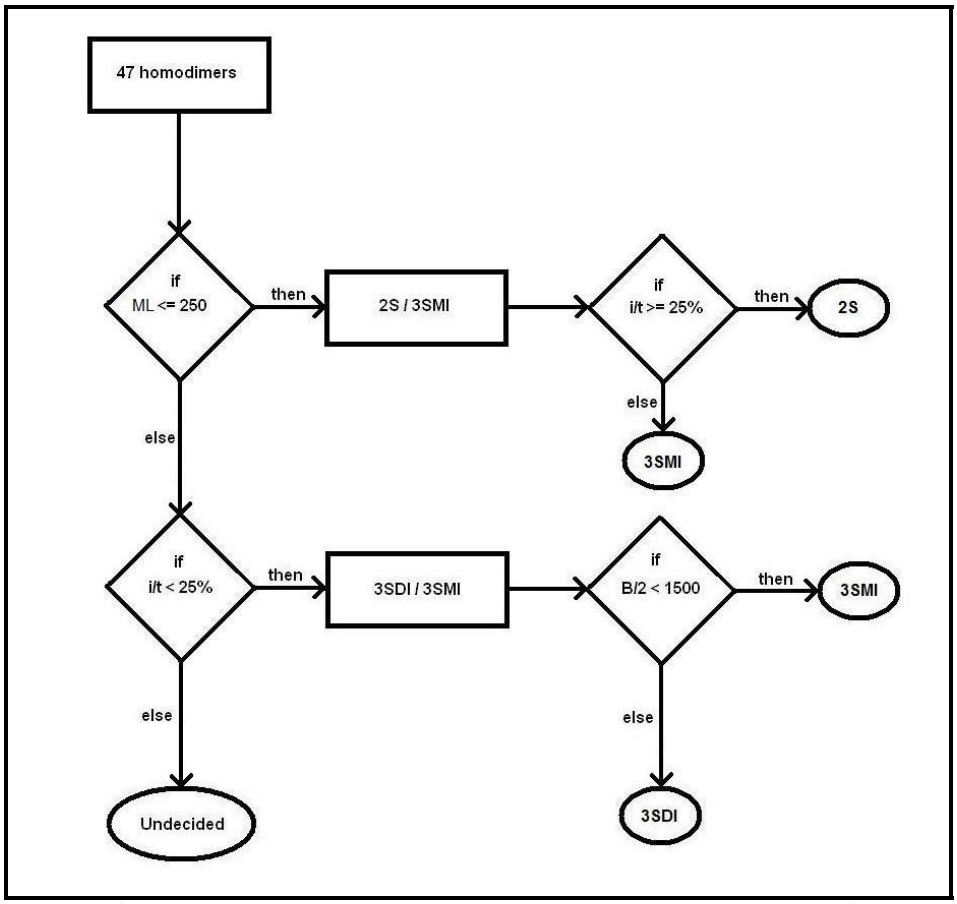

Figure 3: A flowchart describing the decision tree model is given. The decision tree model checks for predictor values within defined conditional values for multiple variables in a subsequent manner sequentially so as to reach the respective nodes to predict and assign target variables.

Conclusion:
It was of interest to predict and classify the homodimer folding mechanism given their structures in complex state. A novel decision tree model is described using structural features (ML, B/2, I/T) derived from known structures to assign folding mechanism for homodimers given their structures. The decision tree model correctly classified with positive predictive values (PPV) 72% for 2S, 58% for 3SMI and 57% for 3SDI into their respective groups in cross validation. Thus, the method finds application in grouping protein homodimer structures with unknown folding data. A number of homodimer structures with unknown folding information are available in PDB. We applied the model to a set of 149 homodimers with unknown folding data. The model classified 132 (88.5% of 149) homodimers into 2S (39), 3SMI (61) and 3SDI (32). Consequently, a framework is established for these 132 known structures with predicted folding data for further experimental verification and confirmation.

Author's contribution:
PK conceived the idea and designed the experiment. VK and AS performed the analysis and summarized results. SL participated in the analysis and UK helped in manuscript preparation.

Acknowledgement:
VK, AS and SL wish to express their sincere thanks to all members of Biomedical Informatics for providing necessary support and material for the analysis. SL is a visitor to Biomedical Informatics.

References:
[1] C Zhanhua *et al., Bioinformation* **1**: 28 (2005) [PMID: 17597849]
[2] TE Wales *et al., Protein Sci.* **13**: 1918 (2004) [PMID: 15169951]
[3] JU Bowie, RT Sauer, *Biochemistry* **28**: 7139 (1989) [PMID: 2819054]
[4] ME Milla, RT Sauer, *Biochemistry* **33**: 1125 (1994) [PMID: 8110744]
[5] C Steif *et al., Biochemistry* **32**: 3867 (1993) [PMID: 8471599]
[6] R Jana *et al., J Mol Biol.* **273**: 402 (1997) [PMID: 9344748]
[7] TB Topping, LM Gloss, *J Mol Biol.* **342**: 247 (2004) [PMID: 15313621]
[8] YK Mok *et al., Protein Sci.* **5**: 310 (1996) [PMID: 8745409]
[9] H Liang, TC Terwilliger, *Biochemistry* **30**: 2772 (1991) [PMID: 2007116]
[10] J Ruiz-Sanz *et al., Eur J Biochem.* **271**: 1497 (2004) [PMID: 15066175]
[11] TB Topping *et al., J Mol Biol.* **335**: 1065 (2004) [PMID: 14698300]
[12] JR Stone *et al., J Biol Chem.* **277**: 5448 (2002) [PMID: 11741982]
[13] SK Grant *et al., Biochemistry* **31:** 9491 (1992) [PMID: 1390732]
[14] K Bajaj *et al., Biochem J.* **380**: 409 (2004) [PMID:14763902]
[15] M Kretschmar, R Jaenicke, *J Mol Biol.* **291**: 1147 (1999) [PMID: 10518950]
[16] CM Johnson *et al., Biochemistry* **31**: 9717 (1992) [PMID: 1390748]
[17] A Tamura *et al., J Mol Biol.* **249**: 636 (1995) [PMID: 7783216]
[18] LM Gloss *et al., J Mol Biol.* **312**: 1121 (2001) [PMID: 11580254]
[19] DE Timm *et al., Biochemistry* **33**: 4667 (1994) [PMID: 8161524]
[20] X Li *et al., J Biol Chem.* **272**: 27324 (1997) [PMID: 9341182]
[21] D Kim *et al., Protein Sci.* **10:** 741 (2001) [PMID: 11274465]
[22] L D'Alfonso *et al., Biochemistry* **41**: 326 (2002) [PMID: 11772032]
[23] HW Dirr, P Reinemer, *Biochem Biophys Res Commun.* **180**: 294 (1991) [PMID: 1930226]
[24] LA Wallace *et al., Biochemistry* **37**: 5320 (1998) [PMID: 9548764]
[25] W Kaplan *et al., Protein Sci.* **6**: 399 (1997) [PMID: 9041642]
[26] N Ahmad *et al., Biochemistry* **37:** 16765 (1998) [PMID: 9843447]
[27] V Mainfroid *et al., J Mol Biol.* **257**: 441 (1996) [PMID: 8609635]
[28] ZW Yang *et al., Protein Sci.* **13**: 830 (1994) [PMID: 14978314]
[29] J Ramstein *et al., J Mol Biol.* **331**: 101 (2003) [PMID: 12875839]
[30] L Zhu *et al., J Mol Biol.* **328**: 235 (2003) [PMID: 12684011]
[31] JK Grimsley *et al., Biochemistry* **36**: 14366 (1997) [PMID: 9398154]
[32] AC Clark *et al., J Biol Chem.* **268**: 10773 (1993) [PMID: 8496144]
[33] C Motono *et al., Biochemistry* **38**: 1332 (1999) [PMID: 9930995]
[34] G Mei *et al., Biochemistry* **36**: 10917 (1997)[PMID: 9283082]
[35] SM Doyle *et al., Biochemistry* **39**: 11667 (2000) [PMID: 10995234]
[36] MG Mateu *J Mol Biol.* **318**: 519 (2002) [PMID: 12051856]
[37] R Ruller *et al., Arch Biochem Biophys.* **411**: 112 (2003) [PMID: 12590929]
[38] D Apiyo *et al., Biochemistry* **40**: 4940 (2001) [PMID: 11305909]
[39] F Malvezzi-Campeggi *et al., Arch Biochem Biophys.* **370**: 201 (1999) [PMID: 10510278]
[40] ME Stroppolo *et al., Arch Biochem Biophys.* **377**: 215 (2000) [PMID: 10845696]
[41] J Malecki, Z Wasylewski, *Eur J Biochem.* **243**: 660 (1997) [PMID: 9057829]
[42] A Aceto *et al., Biochem J.* **285**: 241 (1992) [PMID:1637306]
[43] RS Gokhale *et al., Biochemstry* **35**: 7150 (1996) [PMID: 8679542]
[44] YC Park, H Bedouelle, *J Biol Chem.* **273:** 18052 (1998) [PMID: 9660761]
[45] P Wójciak *et al., Int J Biol Macromol.* **32**: 43 (2003) [PMID: 12719131]
[46] Y Liang *et al., J Biol Chem.* **278**: 30098 (2003) [PMID: 12771138]
[47] The United States Patent and Trademark Office database, USA
[48] T Tanaka *et al., J Biol Che.* **282**: 29987 (2007) [PMID: 17656367]
[49] N Schülke *et al., Proc Natl Acad Sci.* **100**: 12590 (2003) [PMID: 14583590]
[50] KE Neet, DE Timm, *Protein Sci.* **3**: 2167 (1994) [PMID: 7756976]
[51] CJ Tsai, *et al., Protein Sci.* **6**: 1793 (1997) [PMID: 9300480]
[52] Y Levy *et al., Proc Natl Acad Sci.* **101**: 511 (2004) [PMID: 14694192]
[53] G Mei *et al., Febs J.* **272**: 16 (2005) [PMID: 15634328]
[54] L Li *et al., Bioinformation* **1**: 42 (2005) [PMID: 17597851]
[55] S Lulu *et al., J Mol Graph Model.* **28:** 88 (2009) [PMID: 19442545]
[56] OV Tsodikov *et al., J Comput Chem.* **23**: 600 (2002) [PMID: 11939594]

Edited by V. S. Mathura
Citation: Suresh *et al.,* Bioinformation 4(5): 197-205 (2009)

Supplementary material:

Table 1: Dataset of 47 homodimer structures from PDB with known folding information

PDB ID	Folding	ML (aa)	B/2 (Å^2)	IR	I/T	Folding Reference #
2cpg	2S	45	1632	24	71	[2]
1arq	2s	53	2007	42	80	[3]
1arr	2S	53	1962	30	75	[4]
1rop	2S	63	1345	34	54	[5]
5cro	2S	66	648	16	29	[6]
1bfm	2S	69	1593	40	60	[7]
1a7g	2S	82	918	44	32	[8]
1vqb	2S	87	850	47	26	[9]
1b8z	2S	90	1894	19	53	[10]
1ety	2S	98	2079	36	49	[11]
1y7q	2S	98	1508	40	43	[12]
1a8g	2S	99	1785	31	44	[12]
1siv	2S	99	1684	28	42	[13]
1vub	2S	101	1074	18	29	[14]
1hdf	2s	102	156	5	6	[15]
1cmb	2S	104	1813	42	38	[16]
3ssi	2S	108	866	38	29	[17]
1wrp	2S	108	2243	39	48	[18]
1bet	2S	107	1366	41	42	[19]
1buo	2S	121	1972	50	41	[20]
1oh0	2S	131	1036	53	24	[21]
1beb	2s	162	527	15	10	[22]
2gsr	2S	207	1331	49	18	[23]
1gsd	2S	208	1477	52	18	[24]
1gta	2S	218	1505	51	21	[25]
2bqp	2S	234	955	47	41	[26]
1hti	2S	248	1685	46	18	[27]
1ee1	2S	271	2507	48	23	[28]
1mul	3SDI	90	1739	25	50	[29]
1hqo	3SDI	258	1656	31	20	[30]
1psc	3SDI	329	1353	25	12	[31]
1luc	3SDI	355	2072	52	17	[32]
1cm7	3SDI	363	2317	43	16	[33]
1aoz	3SDI	552	1817	9	5	[34]
1nl3	3SDI	835	1351	20	5	[35]
1a43	3SMI	72	921	22	44	[36]
1qll	3SMI	121	432	6	12	[37]
1dfx	3SMI	125	1472	17	34	[38]
1yai	3SMI	151	309	6	9	[39]
1spd	3SMI	154	658	13	13	[40]
1run	3SMI	197	1542	22	21	[41]
11gs	3SMI	209	1197	19	17	[42]
2tdm	3SMI	316	2332	63	20	[43]
1tya	3SMI	319	1513	23	13	[44]
1cvi	3SMI	342	1444	37	13	[44]
1nd5	3SMI	354	1512	31	12	[45]
2crk	3SMI	381	1119	30	11	[46]

Table 2: The minimum, maximum, mean and standard deviation value of the predictor variables is given for 47 homodimers.

Parameters	Min	Max	mean	S.D
Length	45	835	190.5	148.8
B/2	156	2507	1429.2	550.7
I/T (%)	5	80	30	19
IR	6	96	40	15

Table 3: An assignment dataset of 149 homodimers with unknown folding data.

Folding	#	Result			PPV
		TP	FP	UD	
2S	28	20	8	0	71.4%
3SMI	12	7	5	0	58.4%
3SDI	7	4	3	0	57.1%

Table 4: The minimum, maximum, mean and standard deviation value of the predictor variables is given for 149 homodimers of the assignment dataset.

PDB	Assigned Folding	ML	B/2	IR	I/T
1A4I	3SMI	285	1435.8	39	0.14
1A4U	UD	254	2621.6	67	0.26
1AA7	3SMI	158	1170.4	28	0.18
1AD1	3SDI	264	1531.2	38	0.14
1ADE	3SDI	431	3206.6	98	0.23
1AFW	3SDI	390	2545.3	69	0.18
1ALK	UD	449	4042.7	112	0.25
1AOR	3SMI	605	1293.9	36	0.06
1AQ6	3SMI	245	2241.7	55	0.22
1AUO	3SMI	218	694.76	22	0.1
1BBH	3SMI	131	794.11	23	0.18
1BH5	2S	177	3969.4	105	0.59
1BJW	3SDI	381	2864.5	79	0.21
1BMD	3SDI	327	1659.5	43	0.13
1BXG	3SMI	349	1154.8	30	0.08
1C6X	2S	99	1852.1	46	0.46
1CBK	3SMI	160	972.67	30	0.18
1CDC	2S	96	3980.4	86	0.89
1CHM	UD	401	3789.2	105	0.26
1CNZ	3SDI	363	2549	64	0.18
1COZ	3SMI	126	1100.3	29	0.23
1CQS	2S	124	1067	31	0.25
1D1G	2S	164	1647.9	44	0.27
1DOR	3SDI	311	2314.6	60	0.19
1DPG	3SDI	485	2369.9	65	0.13
1DQP	3SMI	230	1827	53	0.23
1DQT	3SMI	117	902.69	27	0.23
1DVJ	3SMI	239	315.72	11	0.05
1EAJ	3SMI	124	760.89	26	0.21
1EBL	3SDI	309	2364.2	67	0.22
1EHI	3SDI	360	2714.4	74	0.2
1EKP	3SDI	365	2461.4	69	0.19
1EN5	3SMI	205	880.59	24	0.12
1EN7	2S	157	3444.1	75	0.48
1EOG	3SMI	208	1214.3	33	0.16
1EXQ	2S	147	1650.4	47	0.32
1EYV	3SMI	131	1165.5	28	0.21
1EZ2	3SMI	328	1412	34	0.1
1F13	2S	161	2050.4	48	0.3
1F17	3SDI	722	2802.6	92	0.13
1F4Q	3SDI	293	1704.9	43	0.15
1F89	3SMI	271	1475.3	36	0.13
1FC5	3SDI	397	2928	85	0.21
1FJH	2S	236	2093	58	0.25
1FL1	3SMI	192	1322	42	0.22
1FP3	3SMI	402	1240.1	33	0.08
1FUX	3SMI	164	877.71	25	0.15
1FWL	3SDI	296	1504.3	43	0.14
1FYD	UD	271	2692.4	69	0.25
1G0S	2S	201	3947.1	96	0.48
1G1A	3SMI	352	1388.5	45	0.13
1G1M	3SDI	287	1866.7	54	0.19
1G64	3SMI	241	936.99	26	0.11
1G8T	2S	169	2509.2	62	0.37
1GD7	2S	109	1681.2	43	0.39
1GGQ	2S	162	2193.2	58	0.35
1H8X	2S	107	1781.7	49	0.46
1HJ3	3SMI	91	70.29	4	0.04
1HJR	3SMI	158	503.45	16	0.1
1HSJ	3SMI	487	2167.2	56	0.11
1HSS	2S	111	1161.9	32	0.28
1I0R	2S	162	2277.3	65	0.4
1I4S	3SMI	147	1130.8	30	0.2
1I8T	3SMI	367	1267.9	42	0.11
1IPI	2S	114	1035.6	30	0.26
1IRI	UD	557	6766.2	180	0.32
1J30	2S	141	3351.4	84	0.59
1JD0	3SMI	260	1229.1	38	0.15
1JFL	3SMI	228	1363.9	40	0.17

1JMV	3SMI	140	1233.1	32	0.23
1JOG	2S	129	1121.9	33	0.25
1JP3	3SMI	210	1793.9	44	0.21
1JR8	2S	105	1281.4	33	0.31
1JV3	3SMI	490	1498.2	83	0.17
1JYS	3SMI	226	1287.8	37	0.16
1K3S	2S	106	1148.7	31	0.29
1K6Z	2S	120	1402.7	36	0.3
1KGN	UD	296	2754	73	0.25
1KIY	3SDI	354	2888.7	73	0.2
1KSO	2S	93	1749.7	42	0.45
1L5B	2S	101	3252.3	80	0.79
1L5X	UD	270	3016.1	73	0.27
1LBQ	3SDI	356	1639	51	0.14
1LHP	3SDI	306	2158.8	56	0.18
1LHZ	2S	213	1759.3	58	0.27
1LK9	UD	425	4614	112	0.26
1LNW	2S	137	1247.2	52	0.38
1LQ9	2S	112	1650.1	46	0.41
1M3E	3SDI	459	2650.1	71	0.15
1M4I	3SMI	181	1327.7	40	0.22
1M6P	3SMI	146	1095.9	35	0.24
1M7H	2S	203	2020.7	50	0.25
1M98	3SDI	400	2891.8	74	0.19
1M9K	3SMI	316	1252.2	41	0.13
1MI3	3SMI	319	1301.1	38	0.12
1MJH	3SMI	143	1089.5	29	0.2
1MKB	2S	171	1809	54	0.31
1MNA	3SMI	276	831.41	23	0.08
1N80	3SDI	328	2606.7	74	0.22
1NA8	3SMI	151	60.86	17	0.11
1NFZ	3SMI	176	857.63	23	0.13
1NU6	3SDI	728	2342.6	65	0.09
1NW1	3SMI	365	1249.1	34	0.09
1NWW	2S	145	1605	42	0.29
1NY5	UD	384	3997.8	108	0.28
1OAC	UD	719	8022.8	221	0.31
1ON2	3SMI	135	1311.4	32	0.23
1OR4	2S	169	1933.5	44	0.26
1ORO	3SMI	213	1292.4	38	0.18
1OTV	3SDI	254	2298.4	60	0.23
1OX8	3SMI	105	748.72	20	0.19
1P3W	3SDI	385	2473.3	74	0.19
1P43	UD	436	1965.6	324	0.74
1PE0	3SMI	187	1369.1	35	0.19
1PJQ	UD	447	6479	162	0.36
1PN0	UD	652	13103	258	0.39
1PN2	3SMI	269	1158.7	31	0.11
1PP2	2S	122	1447.7	42	0.34
1PT5	UD	415	6455	167	0.4
1Q8R	3SMI	118	710.02	20	0.17
1QFH	2S	212	2441	64	0.3
1QHI	3SDI	304	1790.8	53	0.17
1QMJ	3SMI	132	609.43	17	0.13
1QR2	2S	230	2036.3	57	0.25
1QXR	2S	187	1874	48	0.26
1QYA	3SMI	293	1058.1	30	0.1
1R5P	2S	90	808.02	24	0.27
1R7A	3SMI	503	1035.7	34	0.07
1R8J	UD	272	3656	91	0.33
1R9C	2S	125	2022.8	56	0.45
1REG	3SMI	122	690.85	19	0.16
1RVE	3SMI	244	1605.3	46	0.19
1RYA	3SMI	160	1335.5	38	0.23
1S44	3SMI	180	1198.9	34	0.19
1SCF	3SMI	116	875.37	22	0.19
1SMT	2S	98	2030.2	52	0.53
1SOX	3SDI	463	1574.3	51	0.11
1TLU	2S	117	1503.8	44	0.37
1TRK	3SDI	678	4826.6	130	0.19
1UC8	3SDI	254	1946.4	52	0.2
2DAB	3SDI	280	2406.4	63	0.22
2GSA	UD	427	5178.7	146	0.34
2HHM	3SDI	266	1818.7	57	0.21

2NAC	UD	374	3896	103	0.27
3LYN	3SMI	122	1014.5	25	0.2
3SDH	3SMI	145	950.3	27	0.19
7AAT	3SDI	401	3426.8	97	0.24
8PRK	3SMI	282	1015.1	27	0.09
9WGA	3SMI	170	248.23	14	0.08

Table 5: Cross validation experiment positive predictive values (PPV) of the decision tree model when applied to the dataset of 47 homodimers.

Parameters	**Min**	**Max**	**Avg**	**S.D**
Length	90	728	259.8	142.5
B/2	60.8	13103.3	2049.8	1567.1
I/T (%)	4	89	24	13
IR	3	324	57	43.1

Table 6: Classification results of the assignment dataset.

2S	39	1BH5	1C6X	1CDC	1CQS	1D1G	1EN7	1EXQ	1F4Q	1FJH	1G0S
		1G64	1GD7	1GGQ	1H8X	1HSS	1IOR	1IPI	1J30	1JOG	1JR8
		1K35	1K6Z	1KSO	1L5B	1LHZ	1LNW	1LQ9	1M7H	1MKB	1NWW
		1OR4	1PP2	1QFH	1QR2	1QXR	1R5P	1R9C	1SMT	1TLU	
3SMI	61	1A41	1AA7	1AOR	1AQ6	1AUO	1BBH	1BXG	1CBK	1COZ	1DQP
		1DQT	1DVJ	1EAJ	1EN5	1EOG	1EYV	1EZ2	1F89	1FL1	1FP3
		1FUX	1G1A	1G8T	1HJ3	1HJR	1HSJ	1I4S	1I8T	1JDO	1JFL
		1JMV	1JP3	1JV3	1JYS	1M4I	1M6P	1M98	1MI3	1MJH	1MNA
		1NA8	1NFZ	1NW1	1ON2	1ORO	1OXB	1PEO	1PN2	1Q8R	1QMJ
		1QYA	1R7A	1REG	1RVE	1RYA	1S44	1SCF	3LYN	3SDH	8PRK
		9WGA									
3SDI	32	1AD1	1ADE	1AFW	1BJW	1BMD	1CNZ	1DOR	1DPG	1EBL	1EHI
		1EKP	1F13	1F17	1FCS	1FWL	1G1M	1KIY	1LBQ	1LHP	1M3E
		1M9K	1N80	1NU6	1OTV	1P3W	1QHI	1SOX	1TRK	1UC8	2DAB
		2HHM	7AAT								
UD	17	1A4U	1ALK	1CHM	1FYD	1IRI	1KGN	1L5X	1LK9	1NY5	1OAC
		1P43	1PJQ	1PN0	1PT5	1R8J	2GSA	2NAC			