15

Classifying TIM Barrel Protein Domain Structure by an Alignment Approach Using Best Hit Strategy and PSI-BLAST

Chia-Han Chu[1], Chun Yuan Lin[2], Cheng-Wen Chang[1],
Chihan Lee[3] and Chuan Yi Tang[4]
[1]*National Tsing Hua University,*
[2]*Chang Gung University,*
[3]*Chipboud Technology Corporation,*
[4]*Providence University*
Taiwan

1. Introduction

High-tech large-scale sequencing projects have identified a massive number of amino acid sequences for both known and putative proteins, but information on the three-dimensional (3D) structures of these proteins is limited. Several structure databases, such as the Structural Classification of Proteins (SCOP (Andreeva et al., 2008), release version 1.73) and the Class, Architecture, Topology, and Homologous superfamily (CATH (Cuff et al., 2009), release version 3.2.0), contain fewer than 60,000 entries in the Protein Data Bank (PDB (Berman et al., 2000), released on 12 May, 2009). This number of entries constitutes only about 15% of entries in Swiss-Prot (Bairoch et al., 2004), release version 57.2, with more than 400,000 entries). Either X-ray diffraction or NMR can be used to determine the 3D structure of a protein, but each method has its limitation (Dubchak et al., 1995). As such, extracting structural information from sequence databases is an important and complementary alternative to these experimental methods, especially when swiftly determining protein functions or discovering new compounds for medical or therapeutic purposes.

From ASTRAL SCOP 1.73, it has been estimated that ~10% of known enzymes have triosephosphate isomerase (TIM) barrel domains. Moreover, TIM barrel proteins have been identified in five of six enzyme classes, oxidoreductases, transferases, hydrolases, lyases and isomerases, in the Enzyme nomenclature (ENZYME (Bairoch, 2000), released on 5 May, 2009) database; the ligases class does not contain TIM barrel protein. TIM barrel proteins are diverse in sequence and functionality and thus represent attractive targets for protein engineering and evolutionary studies. It is therefore important to examine TIM barrel protein domain structure classification in SCOP and ENZYME.

In SCOP, there are six levels of hierarchy: class, fold, superfamily, family, protein domain and species. The classification of protein structures has, more recently, been facilitated by computer-aided algorithms. Previous research (Chou & Zhang, 1995; Dubchak et al., 1995; Lin et al., 2005, 2007) has shown that an overall prediction accuracy rate of 70-90% can be

easily achieved by using only amino acid sequence information to classify most of proteins into four major classes in SCOP (all-alpha (α), all-beta (β), alpha/beta (α/β) and alpha+beta (α+β)) (Murzin, 1995). For the α/β class (constituting TIM barrel proteins), the overall prediction accuracy rate achieved 97.9% (Lin et al., 2005, 2007). However, less optimal results were obtained if a more complicated category was used, such as protein folding patterns. The overall prediction accuracy rate for classifying 27 fold categories in SCOP only achieved only 50-70% using amino acid sequence information (Ding & Dubchak, 2001; Huang et al., 2003; Lin et al., 2005, 2007; Shen & Chou, 2006; Vapnik, 1995; Yu et al., 2003). Although the classification for the SCOP fold category is still a challenge, the overall prediction accuracy rate for the TIM barrel fold is 93.8% (Yu et al., 2003). Based on the above results, it is possible to further classify TIM barrel proteins into the SCOP superfamily and family categories. Four projection methods, PRIDE (Carugo & Pongor, 2002; Gáspári et al., 2005), SGM (Rogen & Fain, 2003), LFF (Choi et al., 2004) and SSEF (Zotenko et al., 2006, 2007), have been proposed for protein structure comparisons. Zotenko *et al*. (Zotenko et al., 2006) compared these four methods for classifying proteins into the SCOP fold, superfamily and family categories and showed that the SSEF method had the best overall prediction accuracy rate. The SSEF method utilizes 3D structure information to generate the triplet of secondary structure elements as the footprints in the comparisons.

Hence, in this chapter, an alignment approach using the pure best hit strategy, denoted PBH, is proposed to classify the TIM barrel protein domain structures in terms of the superfamily and family categories in SCOP. This approach requires only amino acid sequence information to generate alignment information, but secondary and 3D structure information is also applied in this approach, respectively, to compare the performances with each other. This work is also used to perform the classification for the class category in ENZYME. Two testing data sets, TIM40D and TIM95D from ASTRAL SCOP 1.71 (Chandonia et al., 2004), were tested to evaluate this alignment approach. First, for any two proteins, we adopt the tools CLUSTALW (Thompson et al., 1994), SSEA (Fontana et al., 2005) and CE (Shindyalov & Bourne, 1998) to align the amino acid sequences, secondary and 3D structures, respectively, to obtain the scores of sequence identity, secondary structure identity and RMSD. These scores are then used to build an alignment-based protein-protein identity score network. Finally, a PBH strategy is used to determine the prediction result of a target protein by selecting the protein having the best score for the target protein according to this network. This score can be calculated by a single parameter, such as sequence identity, or mixed parameters by combing two or three single parameters, such as combining sequence identity and secondary structure identity. In this chapter, we only consider the single parameter. To verify the stability of the proposed alignment approach, we also use the novel TIM barrel proteins in TIM40D and TIM95D from ASTRAL SCOP 1.73 that do not existed in ASTRAL SCOP 1.71. For this test, the alignment-based protein-protein identity score network constructed by the TIM barrel proteins from ASTRAL SCOP 1.71 and the PBH strategy are used to predict the classification result for each novel TIM barrel protein. In addition, we further adopt the PSI-BLAST method as a filter for the PBH strategy, denoted the BHPB strategy, to reduce the number of false positives. The experimental results demonstrated that the alignment approach with the PBH strategy or BHPB strategy is a simple and stable method for TIM barrel protein domain structure classification, even when only the amino acid sequence information is available.

2. Materials

2.1 TIM barrel proteins from ASTRAL SCOP 1.71

Two data sets, TIM40D and TIM95D, were used to evaluate the proposed PBH and BHPB alignment strategies. TIM40D contains 272 TIM barrel protein domain sequences (abbreviated to TIM sequences) extracted from the 40D set in ASTRAL SCOP 1.71, in which any two proteins must have ≤40% sequence identity based on PDB SEQRES records. TIM95D contains 439 TIM sequences extracted from the 95D set in ASTRAL SCOP 1.71, in which any two proteins must have ≤95% sequence identity based on PDB SEQRES records. For TIM40D and TIM95D, we directly retrieved amino acid sequences and 3D structures from ASTRAL SCOP 1.71 but excluded redundant and possible mutant data. Secondary structure information for each TIM barrel protein with eight states (H, I, G, E, B, S, T and _) was first derived from the digital shape sampling and processing (DSSP (Kabsch & Sander, 1983)) program. Then the eight states for each TIM barrel protein were then reduced to three states (H, E and C)

Superfamily categories	Index	N_{40D}*	N_{95D}*
Triosephosphate isomerise (TIM)	1	3	16
Ribulose-phosphate binding barrel	2	19	30
Thiamin phosphate synthase	3	2	2
FMN-linked oxidoreductases	4	15	22
Inosine monophosphate dehydrogenase (IMPDH)	5	3	5
PLP-binding barrel	6	8	10
NAD(P)-linked oxidoreductase	7	8	21
(Trans)glycosidases	8	82	134
Metallo-dependent hydrolases	9	18	22
Aldolase	10	31	48
Enolase C-terminal domain-like	11	12	24
Phosphoenolpyruvate/pyruvate domain	12	12	22
Malate synthase G	13	1	2
RuBisCo, C-terminal domain	14	4	10
Xylose isomerise-like	15	7	15
Bacterial luciferase-like	16	7	9
Nicotinate/Quinolinate PRTase C-terminal domain-like	17	4	5
PLC-like phosphodiesterases	18	5	5
Cobalamin (vitamin B12)-dependent enzymes	19	5	6
tRNA-guanine transglycosylase	20	2	2
Dihydropteroate synthetase-like	21	4	6
UROD/MetE-like	22	4	4
FAD-linked oxidoreductase	23	3	3
Pyridoxine 5′-phosphate synthase	24	1	1
Monomethylamine methyltransferase MtmB	25	1	1
Homocysteine S-methyltransferase	26	2	3
(2r)-phospho-3-sulfolactate synthase ComA	27	1	2
Radical SAM enzymes	28	3	3
GlpP-like	29	1	1
CutC-like	30	1	1
ThiG-like	31	1	2
TM1631-like	32	2	2

*N_{40D}: the number of TIM sequences in TIM40D
*N_{95D}: the number of TIM sequences in TIM95D

Table 1. Non-redundant data sets, TIM40D and TIM95D, of superfamily categories in SCOP

Index	N_{40D}*	N_{95D}*	Index	N_{40D}*	N_{95D}*	Index	N_{40D}*	N_{95D}*
1.1	3	16	9.6	4	5	16.3	2	3
2.1	2	5	9.7	1	1	16.4	2	2
2.2	2	4	9.8	1	1	17.1	2	3
2.3	4	7	9.9	1	1	17.2	2	2
2.4	10	13	9.11	1	1	18.1	1	1
2.5	1	1	9.12	1	1	18.2	2	2
3.1	2	2	9.13	1	1	18.3	2	2
4.1	15	22	10.1	18	29	19.1	2	2
5.1	3	5	10.2	2	3	19.2	1	1
6.1	7	9	10.3	3	5	19.3	1	2
6.2	1	1	10.4	3	6	19.4	1	1
7.1	8	21	10.5	3	3	20.1	2	2
8.1	25	48	10.6	2	2	21.1	2	4
8.3	26	41	11.1	1	6	21.2	2	2
8.4	4	12	11.2	11	18	22.1	2	2
8.5	13	18	12.1	1	5	22.2	2	2
8.6	3	3	12.2	1	2	23.1	2	2
8.7	2	2	12.3	1	2	23.2	1	1
8.8	3	3	12.5	4	4	24.1	1	1
8.9	1	1	12.7	4	6	25.1	1	1
8.10	1	2	12.8	1	3	26.1	2	3
8.11	1	1	13.1	1	2	27.1	1	2
8.12	1	1	14.1	4	10	28.1	1	1
8.13	1	1	15.1	1	1	28.2	1	1
8.14	1	1	15.2	1	1	28.3	1	1
9.1	1	2	15.3	2	10	29.1	1	1
9.1	2	2	15.4	1	1	30.1	1	1
9.2	1	3	15.5	1	1	31.1	1	2
9.3	2	2	15.6	1	1	32.1	2	2
9.4	1	1	16.1	2	2	-	-	-
9.5	1	1	16.2	1	2	-	-	-

*N_{40D}: the number of TIM sequences in TIM40D
*N_{95D}: the number of TIM sequences in TIM95D

Table 2. Non-redundant data sets, TIM40D and TIM95D, of family categories in SCOP

Class categories	Index	N_{40D}*	N_{95D}*
Oxidoreductases	1	27	46
Transferases	2	31	53
Hydrolases	3	68	106
Lyases	4	58	97
Isomerases	5	23	49
undefined	-	67	91

*N_{40D}: the number of TIM sequences in TIM40D
*N_{95D}: the number of TIM sequences in TIM95D
The sum of N_{40D} and N_{95D} are 274 and 442, respectively. TIM sequences:
"d1pii_2" and "d1pii_1" in TIM40D and TIM95D have multiple EC numbers for class categories;
"d1b9ba_" and "d1jvna1" in TIM95D have multiple EC numbers for class categories

Table 3. Non-redundant data sets, TIM40D and TIM95D, of class categories in ENZYME

according to the scheme outlined by Jones (Jones, 1999). The TIM sequence "d1cwn_" (SCOP id) in TIM95D was excluded because of lack of secondary structure information (only 438 TIM sequences in TIM95D were tested). The TIM barrel proteins (from ASTRAL SCOP 1.71 and the Universal Protein Resource (UniProt (Bairoch et al., 2005))) for each of TIM40D and TIM95D were classified into 32 superfamily categories, 91 family categories and 5 class categories (Tables 1, 2 and 3; supplemental Table S1(Chu, 2011)).

2.2 Novel TIM barrel proteins from ASTRAL SCOP 1.73

Novel TIM barrel proteins from ASTRAL SCOP 1.73 that do not exist in ASTRAL SCOP 1.71 were also tested. The intersection among the TIM barrel proteins from ASTRAL SCOP 1.71 and 1.73 for TIM40D (Figure 1(A)) and TIM95D (Figure 1(B)) are shown. The number of TIM sequences are represented in green (ASTRAL SCOP 1.71), light blue (ASTRAL SCOP 1.73) and orange (ASTRAL SCOP 1.73 that are not presented in 1.71). In TIM40D (Figure 1(A)), we identified 258 TIM sequences (ASTRAL SCOP 1.71 and 1.73), 14 TIM sequences (exclusively ASTRAL SCOP 1.71) and 64 novel TIM sequences (exclusively ASTRAL SCOP 1.73: 12 of 64 were categorized as new). In TIM95D (Figure 1(B)), we identified 439 TIM sequences (ASTRAL SCOP 1.71 and 1.73) and 79 novel TIM sequences (exclusively ASTRAL SCOP 1.73: 12 of 79 were categorized as new). These 12 novel TIM sequences within the new categories were identical and thus were excluded in the alignment approach. Hence, 52 (TIM40D) and 67 (TIM95D) novel TIM sequences from ASTRAL SCOP 1.73 were used to evaluate the stability of the proposed PBH alignment strategy, respectively. (see supplemental Table S2 (Chu, 2011)).

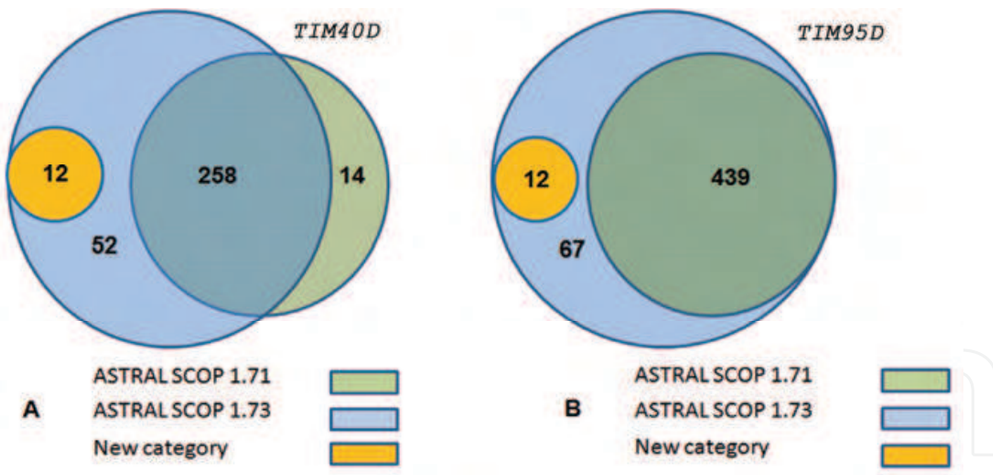


Fig. 1. Intersection among TIM sequences for TIM40D and TIM95D between ASTRAL SCOP 1.71 and 1.73. (A) In TIM40D, there are 272 (ASTRAL SCOP 1.71) and 322 (ASTRAL SCOP 1.73) TIM sequences. (B) In TIM95D, there are 439 (ASTRAL SCOP 1.71) and 518 (ASTRAL SCOP 1.73) TIM sequences.

3. Results and discussion

3.1 Performance analysis

The standard percentage prediction accuracy rate Q_i (Rost & Sander, 1993) was used to evaluate the proposed alignment approach and Q_i is defined as

$$Q_i = \frac{p_i}{n_i} \times 100, \quad (1)$$

where n_i is the number of test proteins in the ith superfamily/family/class category and p_i is the number of test proteins being correctly predicted in the ith superfamily/family/class. The overall prediction accuracy rate Q is given by

$$Q = \sum_{i=1}^{k} q_i Q_i , \quad (2)$$

where $q_i = n_i/K$, where K is the total number of test proteins. Q_i is equivalent to Recall (Gardy et al., 2003), which is defined as

$$\text{Recall}_i = \frac{TP_i}{(TP_i + FN_i)}, \quad (3)$$

where TP_i (true positives) is the number of correctly predicted proteins in the ith superfamily/family/class category, and FN_i (false negatives) is the number of missed proteins in the ith superfamily/family/class category. Precision (Gardy et al., 2003) was also used to evaluate the proposed alignment approach. Precision is defined as

$$Precision_i = \frac{TP_i}{(TP_i + FP_i)}, \quad (4)$$

where FP_i (false positives) is the number of pseudo proteins predicted in the ith superfamily/family/class category. In addition, the Matthews Correlation Coefficient (*MCC* for short) (Matthews, 1975) was used to measure the prediction quality of classifications by utilizing the proposed PBH and BHPB alignment strategies. *MCC* accounts for TP_i, FP_i, TN_i and FN_i as a balanced measure, which can be used for categories with varying sizes. *MCC* returns a value +1 for the perfect prediction quality, 0 for the average random prediction quality, or -1 for an inverse prediction quality. The formula of *MCC* is defined as

$$MCC = \frac{TP_i \times TN_i - FP_i \times FN_i}{\sqrt{(TP_i + FP_i) \times (TP_i + FN_i) \times (TN_i + FP_i) \times (TN_i + FN_i)}} \quad (5)$$

3.2 Alignment approach with the PBH strategy

Zotenko *et al*. (Zotenko et al., 2006) compared four projection methods, PRIDE, SGM, LFF and SSEF, to classify the 40D data set (ASTRAL SCOP 1.69) into superfamily and family categories in SCOP. There are 246 TIM barrel proteins classified into 24 superfamily categories and 210 TIM barrel proteins classified into 42 family categories. Based on the overall Q values, SSEF outperformed LFF, SGM and PRIDE for TIM barrel protein structure classification (Table 4).

For the proposed PBH alignment strategy, the overall Q values for TIM40D and TIM95D from ASTRAL SCOP 1.71 are shown in Table 5. In Table 5, the threshold (see Methods) is determined without decreasing the Q value, which is achieved without a threshold. The Q

Method	SSEF Q (%)	LFF Q (%)	SGM Q (%)	PRIDE Q (%)
Superfamily*	78.0	73.6	63.0	41.5
Family*	74.3	72.9	57.1	45.2

*: the performances of SSEF, LFF, SGM and PRIDE are extracted from the additional file (Zotenko et al., 2006)

Table 4. Overall *Q* values for SSEF, LFF, SGM and PRIDE in TIM40D (ASTRAL SCOP 1.69)

value will decrease when a score, which is higher or lower than the threshold given in Table 5, is assigned as the threshold. For TIM40D, the best *Q* value (84.2%) for the superfamily classification is derived according to secondary structure identity, 76.1% for the family classification is derived according to sequence identity and 48.2% for the class classification is derived according to sequence identity (Table 5). The *Q* value of 48.2% for the class classification is not valid. In TIM40D, 67 of 274 TIM sequences with undefined class categories (derived from UniProt) were initially assumed to be false negatives before the test (see Discussion). Using amino acid sequence or secondary structure information, the PBH alignment strategy yields results as good as SSEF (footprint information). This alignment approach will be useful for TIM barrel proteins lacking 3D structure information. Moreover, for the class classification, the *Q* value of 48.2% is better than the *Q* value of 35% (under Rank 1 condition) by a non-alignment method proposed by Dobson and Doig (Dobson & Doig, 2005). For TIM95D, the best *Q* value (93.2%) for the superfamily classification is derived according to secondary structure identity, 90.0% for the family classification is derived according to sequence identity and 65.2% for the class classification is derived according to secondary structure identity. Similarly, for the class classification, 91 of 442 TIM sequences with undefined class categories in TIM95D were initially assumed to be false negatives before the test.

	Method	Sequence identity		Secondary structure identity		RMSD	
		Q (%)	Threshold	*Q* (%)	Threshold	*Q* (%)	Threshold
TIM40D	Superfamily	83.1	<14	84.2	<67	40.4	>1.9
	Family	76.1	<14	75.0	<67	37.1	>1.9
	Class (ENZYME)	48.2	<13	47.4	<67	21.2	>1.8
TIM95D	Superfamily	92.5	<14	93.2	<68	68.0	>2.0
	Family	90.0	<14	89.3	<68	66.4	>2.0
	Class (ENZYME)	64.0	<16	65.2	<72	48.0	>1.8

Table 5. Overall *Q* values for the PBH alignment strategy in TIM40D and TIM95D (ASTRAL SCOP 1.71)

Overall, the *Q* values of the PBH alignment strategy using secondary structure information are similar to those using amino acid sequence information in TIM40D and TIM95D. For practical purposes, however, it may be best to use only amino acid sequence information. In addition, the *Q* values of the PBH alignment strategy using the RMSD do not yield valid results. The RMSD (global alignment result in this chapter) may not be a valid feature for

the alignment approach to perform TIM barrel protein domain structure classification. In Table 5, the threshold is too low for sequence identity, suggesting that the sequence identity of the target and its selected proteins (within the same category) is low. When the threshold is set higher than the above sequence identity, the target protein becomes a false negative and then the *Q* value under the "no threshold" condition will decrease. The threshold is high for secondary structure identity, suggesting that the secondary structure identity of the target and its selected proteins (within the same category) is high. These results imply that although TIM barrel proteins have diverse sequences they have very similar secondary structures. This inference matches the recent observation for the TIM barrel proteins.

Tables 6 and 7 show overall Q and Precision values for various categories in TIM40D and TIM95D from ASTRAL SCOP 1.71, respectively. Only the categories with more than ten TIM sequences are listed; tests for RMSD were omitted because the results were invalid. In Tables 6 and 7, the threshold is determined with the best Precision value without decreasing the Q value, which is achieved without a threshold. Precision values with the threshold outperform or are equals to those without the threshold. However, it is very difficult to determine the appropriate threshold to obtain the best Precision value for routine alignment practices. This problem may be omitted by the BHPB strategy to reduce the number of false positives.

	Method	**Sequence identity**			**Secondary structure identity**		
	Index	***Q* (%)**	**Precision[1] (%)**	**Precision[2] (%)**	***Q* (%)**	**Precision[1] (%)**	**Precision[2] (%)**
Superfamily	2	94.7	64.3	78.3(17)	84.2	64.0	84.2(78)
	4	73.3	78.6	100.0(18-22)	73.3	100.0	100.0(<77)
	8	87.8	92.3	93.5(13)	86.6	95.9	95.9(<67)
	9	83.3	78.9	78.9(<15)	72.2	86.7	92.9(68-70)
	10	83.9	78.8	81.3(16)	96.8	78.9	81.0(74)
	11	83.3	76.9	100.0(17-18)	91.7	91.7	100.0(73-75)
	12	75.0	90.0	100.0(15)	83.3	100.0	100.0(<77)
Family	2.4	100.0	66.7	100.0(19-29)	100.0	71.4	100.0(80-86)
	4.1	73.3	78.6	100.0(18-22)	73.3	100.0	100.0(<77)
	8.1	88.0	84.6	88.0(13)	92.0	95.8	95.8(<67)
	8.3	84.6	95.7	95.7(<17)	84.6	88.0	88.0(<73)
	8.5	100.0	86.7	100.0(17)	92.3	100.0	100.0(<75)
	10.1	83.3	78.9	100.0(18-19)	88.9	72.7	84.2(76)
	11.2	90.9	76.9	100.0(17-18)	90.9	83.3	100.0(76-79)
Class (ENZYME)	1	66.7	64.3	85.7(21-22)	74.1	71.4	71.4(<72)
	2	45.2	58.3	58.3(<13)	45.2	66.7	70.0(64-66)
	3	60.3	60.3	62.1(14)	54.4	61.7	63.8(68-70)
	4	77.6	65.2	71.4(16)	75.9	62.0	69.8(74)
	5	61.0	53.8	61.0(15-16)	65.2	45.5	48.4(70-71)

Table 6. Overall *Q* and Precision values for the PBH alignment strategy in TIM40D (ASTRAL SCOP 1.71)

	Method	Sequence identity			Secondary structure identity		
	Index	*Q* (%)	Precision[1] (%)	Precision[2] (%)	*Q* (%)	Precision[1] (%)	Precision[2] (%)
Superfamily	1	100.0	94.1	100.0(17-44)	100.0	88.9	94.1(80-85)
	2	96.7	78.4	82.9(17)	90.0	84.4	96.4(79)
	4	90.9	87.0	90.9(15-16)	86.4	95.0	100.0(70-81)
	6	90.0	100.0	100.0(<22)	100.0	100.0	100.0(<84)
	7	100.0	95.2	100.0(16-23)	100.0	100.0	100.0(<82)
	8	91.8	98.4	98.4(<14)	95.5	98.5	98.5(<68)
	9	81.8	94.7	94.7(<16)	77.3	89.5	94.4(68-74)
	10	95.8	86.8	88.5(16)	97.9	87.0	94.0(76-77)
	11	100.0	96.0	100.0(17-21)	100.0	96.0	100.0(74-80)
	12	100.0	100.0	100.0(<26)	100.0	100.0	100.0(<79)
	14	100.0	100.0	100.0(<31)	100.0	83.3	100.0(74-85)
	15	93.3	82.4	82.4(<16)	93.3	93.3	93.3(<76)
Family	1.1	100.0	94.1	100.0(17-44)	100.0	88.9	94.1(80-85)
	2.4	100.0	76.5	100.0(20-29)	100.0	86.7	100.0(77-86)
	4.1	90.9	87.0	90.9(15-16)	86.4	95.0	100.0(70-81)
	7.1	100.0	95.2	100.0(16-23)	100.0	100.0	100.0(<82)
	8.1	95.8	97.9	97.9(<14)	97.9	97.9	97.9(<68)
	8.3	92.7	100.0	100.0(<17)	92.7	92.7	95.0(73-74)
	8.4	100.0	92.3	100.0(16-35)	100.0	100.0	100.0(<85)
	8.5	100.0	90.0	100.0(17)	94.4	94.4	100.0(72-74)
	10.1	96.6	87.5	100.0(18-19)	96.6	84.8	93.3(76-77)
	11.2	100.0	94.7	100.0(17-21)	100.0	94.7	100.0(74-80)
	14.1	100.0	100.0	100.0(<31)	100.0	83.3	100.0(74-85)
	15.3	100.0	66.7	100.0(18-67)	100.0	90.9	100.0(76-95)
Class (ENZYME)	1	89.1	80.4	91.1(21-23)	89.1	80.4	82.0(70-71)
	2	77.4	75.9	75.9(<16)	79.2	87.5	87.5(<73)
	3	67.0	71.7	72.4(14-15)	69.8	71.2	73.3(72-73)
	4	91.8	80.2	84.0(17)	90.7	79.3	80.7(73)
	5	83.7	78.8	82.0(15)	87.8	79.6	79.6(<72)

Table 7. Overall *Q* and Precision values for the PBH alignment strategy in TIM95D (ASTRAL SCOP 1.71)

The *Q* values for Ribulose-phosphate binding barrel, (Trans)glycosidases, Aldolase and Enolase C-terminal domain-like in the superfamily categories were above 84.2%, whereas FMN-linked oxidoreductases, Metallo-dependent hydrolases, and Phosphoenolpyruvate/pyruvate domain had lower *Q* values (Table 6). All of family categories except for FMN-linked oxidoreductases had *Q* values above 76.1%. Only one of the class categories, Transferases, had a *Q* value below 48.2%. For the superfamily categories, Ribulose-phosphate binding barrel, (Trans)glycosidases and Metallo-dependent hydrolases, the *Q* values derived according to sequence identity were better than those derived according to secondary structure identity. In contrast, the *Q* values of Aldolase, Enolase C-terminal domain-like and Phosphoenolpyruvate/pyruvate domain derived

according to secondary structure identity were better than those derived according to sequence identity. FMN-linked oxidoreductases yielded the same *Q* values based on sequence identity and secondary structure identity. For the family category, Type II chitinase had a better *Q* value derived according to sequence identity than derived according to secondary structure identity. Amylase, catalytic domain and Class I aldolase produced better *Q* values derived according to secondary structure identity than derived according to sequence identity. Tryptophan biosynthesis enzymes, FMN-linked oxidoreductases, beta-glycanases and D-glucarate dehydratase-like had the same *Q* values derived according to sequence identity and secondary structure identity. For the class categories, the *Q* values of Hydrolases and Lyases derived according to sequence identity were better than those derived according to secondary structure. Oxidoreductases and Isomerases yielded better *Q* values derived according to secondary structure identity than derived according to sequence identity. The remaining category, Transferases, had the same *Q* values derived according to sequence identity and secondary structure identity. These results demonstrated that the proposed PBH alignment strategy is more useful for certain TIM barrel proteins than others.

In TIM95D superfamily categories, FMN-linked oxidoreductases and Metallo-dependent hydrolases yielded Q values less than 93.2% (Table 7); others all yielded Q values above 93.2%. All of family categories obtained Q values above 90.0% and no class category obtained Q values below 65.2%. For the superfamily categories, Ribulose-phosphate binding barrel, FMN-linked oxidoreductases and Metallo-dependent hydrolases, and for the family categories, FMN-linked oxidoreductases and Type II chitinase, the Q values derived according to sequence identity were better than those derived according to secondary structure identity. For other superfamily and family categories, the same Q values were obtained using sequence identity and secondary structure identity. For the class categories, Transferases, Hydrolases and Isomerases had better Q values derived according to secondary structure identity than those derived according to sequence identity; Lyases had a better Q value derived according to sequence identity than that derived according to secondary structure identity. The last category, Oxidoreductases, produced the same Q values derived according to sequence identity and secondary structure identity.

3.3 Estimating stability using the PBH alignment strategy

Novel TIM sequences in TIM40D (n=52) and TIM95D (n=67) from ASTRAL SCOP 1.73 were used to estimate the stability of the proposed PBH alignment strategy. Table 8 presents the overall *Q* values for novel TIM sequences. The definition and observation of the threshold in Table 8 is the same as that in Table 5. In Table 8, the best *Q* values for the superfamily, family and class classifications in TIM40D and TIM95D from ASTRAL SCOP 1.73 were derived according to sequence identity. For TIM40D, the best *Q* value was 94.2% for superfamily, 90.4% for family and 40.4% for class; for TIM95D, the best *Q* value was 91.0% for superfamily, 88.1% for family and 47.8% for class. Similarly, for the class classification, 20 of 52 (TIM40D) and 25 of 67 (TIM95D) novel TIM sequences with undefined class categories were initially assumed to be false negatives before the test (see supplemental Table S3 (Chu, 2011)). These results suggest that the proposed PBH alignment strategy is stable and suitable for TIM barrel protein domain structure classification.

	Method	Sequence identity		Secondary structure identity		RMSD	
		Q (%)	Threshold	*Q* (%)	Threshold	*Q* (%)	Threshold
TIM40D	Superfamily	94.2	<16	90.4	<72	57.7	>1.7
	Family	90.4	<16	84.6	<74	55.8	>1.7
	Class (ENZYME)	40.4	<17	40.4	<72	25.0	>1.7
TIM95D	Superfamily	91.0	<16	86.6	<72	59.7	>1.8
	Family	88.1	<16	80.6	<78	58.2	>1.8
	Class (ENZYME)	47.8	<14	44.8	<72	29.9	>1.7

Table 8. Overall *Q* values for novel TIM sequences in TIM40D and TIM95D (ASTRAL SCOP 1.73)

3.4 Alignment strategy with the BHPB strategy

The high *Q* value derived according to sequence identity using the PBH alignment strategy can decrease the false positives via the homologous finding method. PSI-BLAST is an established method that detects subtle relationships between proteins that are structurally distant or functionally homologous owing to a position-specific scoring matrix generated from multiple alignments of the top-scoring BLAST responses to a given query sequence. The PSI-BLAST package was integrated into the NCBI standalone BLAST package (Altschul et al., 1997). All of our tests were implemented using Perl combined with the CPAN bioperl package (http://www.cpan.org/).

Table 9 presents the overall *Q* values for TIM sequences in TIM40D and TIM95D from ASTRAL SCOP 1.71 using PSI-BLAST as a filter. The definition and observation of the threshold in Table 9 is the same as that in Table 5. For TIM40D (Table 9), the best *Q* values acquired according to sequence identity were 76.1% for superfamily, 73.9% for family, and 41.6% for class. For TIM95D, the secondary structure identity was used obtain the best *Q* value of 62.2% for class, whereas sequence identity was used to obtain the best *Q* values for superfamily (88.8%) and family (88.4%). Based on Tables 5 and 9, the *Q* values obtained using the BHPB alignment strategy were slightly lower than those obtained using the PBH alignment strategy. The lower *Q* values may be a consequence of proteins for which no homolog was found using PSI-BLAST method; such proteins were thus false negatives. Although the overall *Q* values using the PBH alignment strategy were higher than those using the BHPB alignment strategy, Precision values obtained using the BHPB alignment strategy were higher than those using the PBH alignment strategy. Tables 10 (TIM40D from ASTRAL SCOP 1.71) and 11 (TIM95D) show the overall *Q* and Precision values for TIM sequences within various categories. The definitions of the threshold in Tables 10 and 11 are the same as that in Tables 6 and 7, respectively.

3.4.1 Q analysis

In Table 10, for the superfamily categories, the same categories as those observed in Table 6 obtained *Q* values above 76.1%; however, all of the categories obtained better or equal *Q* values derived according to sequence identity than derived according to secondary

	Method	Sequence identity		Secondary structure identity		RMSD	
		Q (%)	Threshold	*Q* (%)	Threshold	*Q* (%)	Threshold
TIM40D	Superfamily	76.1	<14	73.5	<67	39.0	>1.9
	Family	73.9	<14	70.6	<67	37.1	>1.9
	Class (ENZYME)	41.6	<17	40.9	<73	20.1	>1.8
TIM95D	Superfamily	88.8	<14	87.2	<68	67.4	>2.0
	Family	88.4	<14	86.3	<68	66.4	>2.0
	Class (ENZYME)	61.3	<18	62.2	<73	47.7	>1.8

Table 9. Overall *Q* values for the BHPB alignment strategy in TIM40D and TIM95D (ASTRAL SCOP 1.71)

	Method	Sequence identity			Secondary structure identity		
	Index	*Q* (%)	Precision[1] (%)	Precision[2] (%)	*Q* (%)	Precision[1] (%)	Precision[2] (%)
Superfamily	2	89.5	100.0	100.0(<20)	84.2	80.0	88.9(79)
	4	73.3	100.0	100.0(<23)	73.3	100.0	100.0(<77)
	8	78.0	100.0	100.0(<14)	74.4	98.4	98.4(<67)
	9	44.4	100.0	100.0(<26)	44.4	88.9	100.0(78-79)
	10	83.9	96.3	96.3(<17)	83.9	92.9	92.9(<75)
	11	83.3	90.9	100.0(17-18)	83.3	100.0	100.0(<80)
	12	75.0	90.0	100.0(15)	75.0	100.0	100.0(<77)
Family	2.4	100.0	100.0	100.0(<30)	100.0	90.9	100.0(77-86)
	4.1	73.3	100.0	100.0(<23)	73.3	100.0	100.0(<77)
	8.1	88.0	100.0	100.0(<14)	92.0	100.0	100.0(<67)
	8.3	84.6	100.0	100.0(<17)	76.9	95.2	95.2(<75)
	8.5	92.3	100.0	100.0(<18)	84.6	100.0	100.0(< 81)
	10.1	83.3	93.8	100.0(18-19)	83.3	88.2	93.8(75-76)
	11.2	90.9	90.9	100.0(17-18)	90.9	100.0	100.0(<80)
Class (ENZYME)	1	66.7	78.3	85.7(21-22)	70.4	73.1	76.0(75-76)
	2	38.7	60.0	63.2(17)	32.3	71.4	71.4(<73)
	3	42.6	61.7	69.0(18)	44.1	60.0	62.5(72-73)
	4	70.7	78.8	85.4(18-19)	69.0	74.1	76.9(75-76)
	5	60.9	63.6	63.6(<17)	56.5	56.5	59.1(78-79)

Table 10. Overall *Q* and Precision values for the BHPB alignment strategy in TIM40D (ASTRAL SCOP 1.71)

structure identity. For the family categories, only FMN-linked oxidoreductases had a *Q* value less than 73.9%. Amylase, catalytic domain was the only category that had a lower *Q* value when using sequence identity instead of secondary structure identity. For the class categories, only Transferases had a *Q* value less than 41.6%. For Transferases, Lyases and Isomerases, the *Q* values derived according to sequence identity were higher than those

derived according to secondary structure identity. In Table 11, for the superfamily categories, (Trans)glycosidases, Metallo-dependent hydrolases and Xylose isomerase-like, had *Q* values less than 88.8%. For the family categories, only beta-glycanases had a *Q* value less than 88.4%. For all superfamily and family categories, the *Q* values derived according to sequence identity were higher than or equal to those derived according to secondary structure identity. All of the class categories had *Q* values higher than 62.2%. For Hydrolases and Isomerases, the *Q* values derived according to secondary structure identity were higher than those derived according to sequence identity.

	Method	**Sequence identity**			**Secondary structure identity**		
	Index	***Q* (%)**	**Precision[1] (%)**	**Precision[2] (%)**	***Q* (%)**	**Precision[1] (%)**	**Precision[2] (%)**
Superfamily	1	100.0	100.0	100.0(<45)	100.0	94.1	94.1(<86)
	2	96.7	100.0	100.0(<18)	90.0	93.1	96.4(77-79)
	4	90.9	100.0	100.0(<17)	86.4	100.0	100.0(<82)
	6	90.0	100.0	100.0(<22)	90.0	100.0	100.0(<88)
	7	100.0	100.0	100.0(<24)	100.0	100.0	100.0(<82)
	8	86.6	100.0	100.0(<14)	85.1	99.1	99.1(<72)
	9	63.6	100.0	100.0(<26)	63.6	93.3	100.0(78-79)
	10	93.8	97.8	97.8(<17)	93.8	93.8	95.7(75-77)
	11	100.0	100.0	100.0(<22)	100.0	100.0	100.0(<81)
	12	100.0	100.0	100.0(<26)	100.0	100.0	100.0(<79)
	14	100.0	100.0	100.0(<31)	100.0	100.0	100.0(<86)
	15	80.0	100.0	100.0(<18)	80.0	100.0	100.0(<80)
Family	1.1	100.0	100.0	100.0(<45)	100.0	94.1	94.1(<86)
	2.4	100.0	100.0	100.0(<30)	100.0	92.9	100.0(77-86)
	4.1	90.0	100.0	100.0(<17)	86.4	100.0	100.0(<82)
	7.1	100.0	100.0	100.0(<24)	100.0	100.0	100.0(<82)
	8.1	95.8	100.0	100.0(<14)	95.8	100.0	100.0(<68)
	8.3	87.8	100.0	100.0(<17)	85.4	97.2	97.2(<75)
	8.4	100.0	100.0	100.0(<36)	100.0	100.0	100.0(<85)
	8.5	94.4	100.0	100.0(<18)	88.9	100.0	100.0(<83)
	10.1	93.1	96.4	100.0(18-25)	93.1	93.1	96.4(75-86)
	11.2	100.0	100.0	100.0(<22)	100.0	100.0	100.0(<81)
	14.1	100.0	100.0	100.0(<31)	100.0	100.0	100.0(<86)
	15.3	100.0	100.0	100.0(<68)	100.0	100.0	100.0(<96)
Class (ENZYME)	1	89.1	87.2	91.1(21-23)	87.0	83.3	85.1(78-79)
	2	75.4	81.6	83.3(17)	75.4	88.9	88.9(<73)
	3	59.4	72.4	74.1(18)	64.2	73.9	74.7(72-73)
	4	89.7	87.9	88.8(17)	88.7	86.9	88.7(77)
	5	81.6	93.0	93.0(<19)	83.7	85.4	85.4(<80)

Table 11. Overall *Q* and Precision values for the BHPB alignment strategy in TIM95D (ASTRAL SCOP 1.71)

3.4.2 Precision analysis

In Tables 10 and 11, Precision values with the threshold were higher or equal to those without the threshold, thus making it difficult to determine the feasible threshold to obtain the best Precision value for routine alignment practices. However, the differences between Precision values with and without the threshold were greatly reduced by using the BHPB alignment strategy with the exception of Hydrolases of TIM40D. Using the BHPB alignment strategy in TIM40D, the average Precision values without the threshold were 96.7% for superfamily, 97.8% for family, and 68.5% for class (Table 10). Using the BHPB alignment strategy in TIM95D, the average Precision values without the threshold were 99.8% for superfamily, 99.7% for family, and 84.4% for class (Table 11). The best average Precision values were derived according to sequence identity. The PSI-BLAST method in the BHPB alignment strategy can filter out some of the false positives. Figures 2 and 3 indicate the

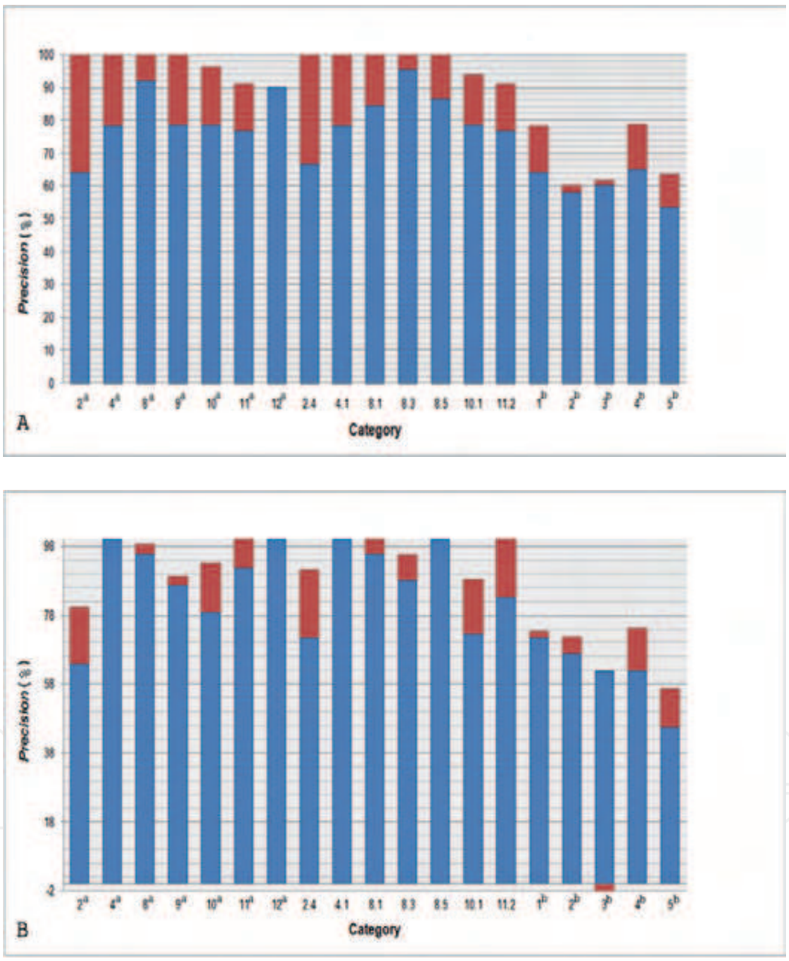


Fig. 2. The increase in Precision values for TIM40D (ASTRAL SCOP 1.71) using the BHPB alignment strategy. (A) Increase in Precision values for TIM40D derived according to sequence identity. (B) Increase in Precision values for TIM40D derived according to secondary structure identity. Superscript 'a' or 'b' indicates the superfamily categories or the class categories, respecively. Categories without a superscript indicated the family categories. The blue bar indicates Precision values using the PBH alignment strategy and the red bar indicates the increase in Precision values using the BHPB alignment strategy.

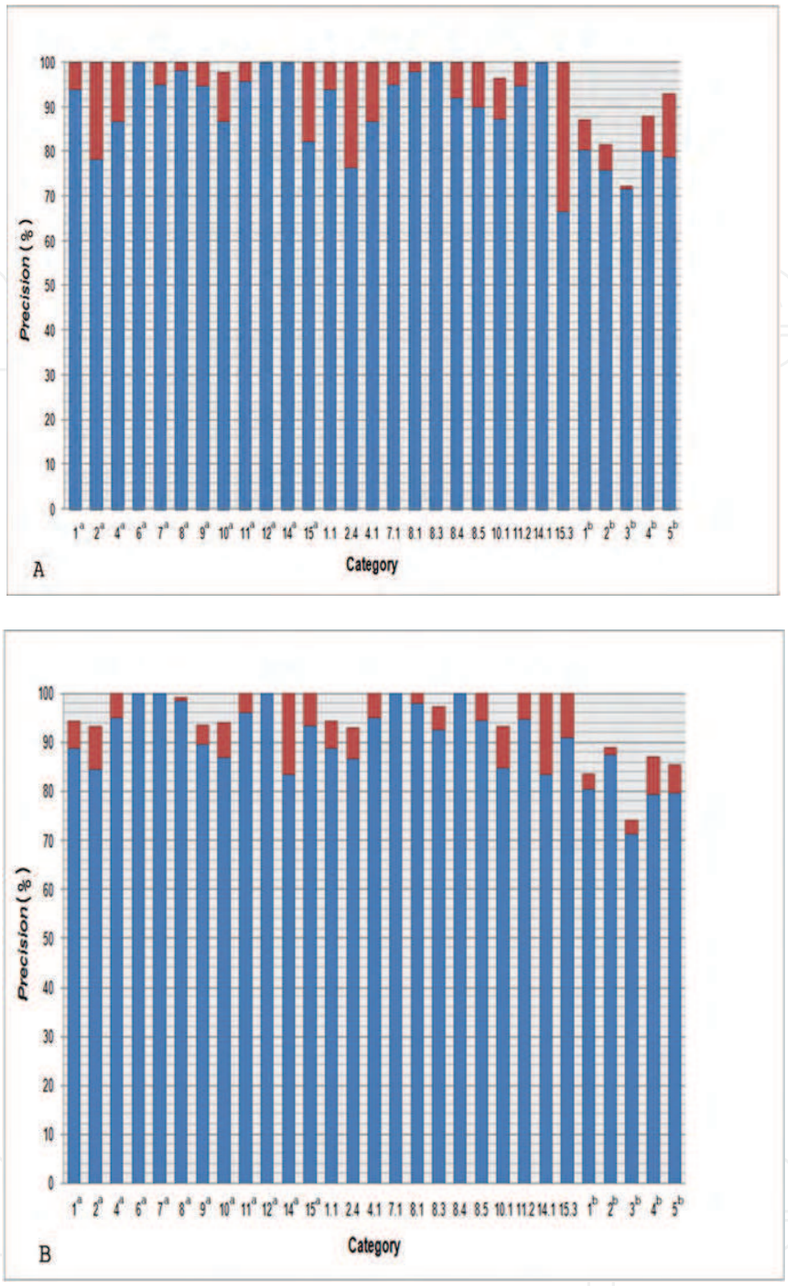


Fig. 3. The increase in Precision values for TIM95D (ASTRAL SCOP 1.71) using the BHPB alignment strategy. (A) Increase in Precision values for TIM95D derived according to sequence identity. (B) Increase in Precision values for TIM95D derived according to secondary structure identity. Superscript 'a' or 'b' indicates the superfamily categories or the class categories, respecively. Categories without a superscript indicated the family categories. The blue bar indicates Precision values using the PBH alignment strategy and the red bar indicates the increase in Precision values using the BHPB alignment strategy.

increase in Precision values for TIM40D and TIM95D from ASTRAL SCOP 1.71 using the BHPB alignment strategy as compared to the PBH alignment strategy, respectively. The

increase in Precision values was computed by comparing the results shown in Tables 6, 7, 10 and 11 (see supplemental Table S4 (Chu, 2011)). Based on Figures 2 and 3, Precision values for almost all categories improved when using the BHPB alignment strategy. The average increases in Precision values for TIM40D using sequence identity were 16.8% for superfamily, 16.7% for family and 8.1% for class. The average increases in Precision values using secondary structure identity were 7.1% for superfamily, 9.5% for family and 7.0% for class. The average increases in Precision values derived according to sequence identity were higher than those derived according to secondary structure identity for TIM40D and TIM95D. Thus, the BHPB alignment strategy yields higher Precision values than the PBH alignment strategy.

3.4.3 MCC analysis

Figure 4 presents the *MCC* measures of (1) the PBH alignment strategy derived according to sequence identity (PBH(1D) for short), (2) the PBH alignment strategy derived according to secondary structure identity (PBH(2D) for short), (3) the BHPB alignment strategy derived according to sequence identity (BHPB(1D) for short) and (4) the BHPB alignment strategy derived according to secondary structure identity (BHPB(2D) for short) for TIM40D and TIM95D from ASTRAL SCOP 1.71, respectively. (see supplemental Table S5 (Chu, 2011))

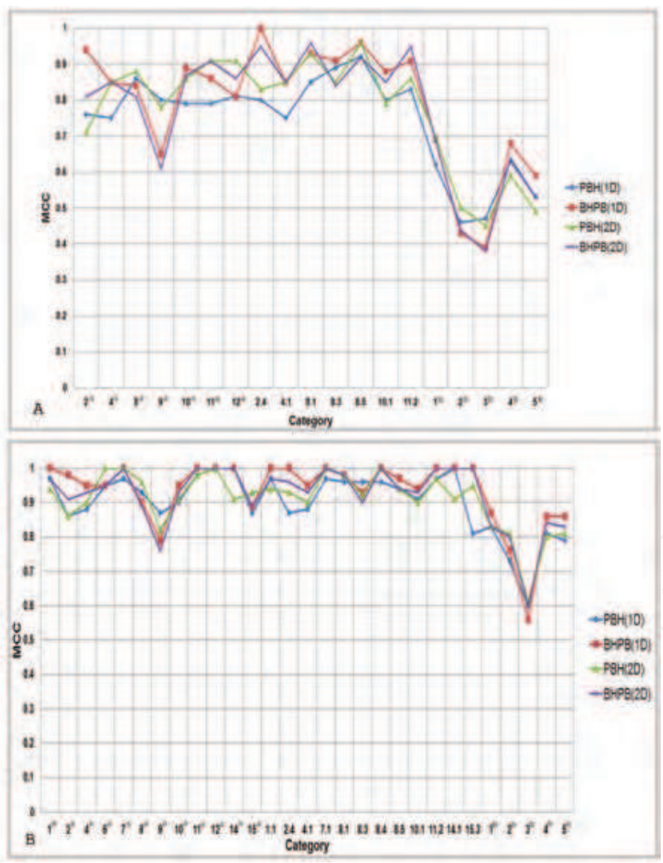


Fig. 4. *MCC* scores of PBH(1D), PBH(2D), BHPB(1D) and BHPB(2D) for TIM40D and TIM95D (ASTRAL SCOP 1.71). (A) *MCC* scores for TIM40D. (B) *MCC* scores for TIM95D. Superscript 'a' or 'b' indicates the superfamily categories or the class categories. Categories without a superscript indicate the family categories.

Using the PBH and BHPB alignment strategies, all of the superfamily categories had *MCC* scores greater than 0.7 except Metallo-dependent hydrolases when using the BHPB alignment strategy (Figure 4(A)); Ribulose-phosphate binding barrel, Enolase C-terminal domain-like, and Phosphoenolpyruvate/pyruvate domain had *MCC* scores greater than 0.9. All of the family categories had *MCC* scores greater than 0.7; Tryptophan biosynthesis enzymes, Amylase, catalytic domain, beta-glycanases, Type II chitinase, and D-glucarate dehydratase-like had *MCC* scores greater than 0.9. All of the class categories had *MCC* scores between 0.3~0.7, which is not an optimal score. From Figure 4(B), all of the superfamily and family categories had *MCC* scores greater than 0.7; 13 categories had the optimal *MCC* score (+1), indicating perfect prediction quality. All of the class categories had *MCC* scores between 0.5~0.9. The above results demonstrate that the proposed PBH or BHPB alignment strategy yielded high prediction quality for TIM barrel protein domain structure classification.

3.5 Discussion

Here we further investigate why the alignment approach with the PBH or BHPB strategy is not sufficient to classify the class category. For the above experiments, all of the EC annotations for TIM sequences in TIM40D and TIM95D from ASTRAL SCOP 1.71 and 1.73 were derived from UniProt. There are 24.5% TIM40D (67 of 274) and 20.6% TIM95D (91 of 442) TIM sequences listed as undefined in class from ASTRAL SCOP 1.71; there are 38.5% TIM40D (20 of 52) and 37.3% TIM95D (25 of 67) novel TIM sequences listed as undefined in class from ASTRAL SCOP 1.73. These TIM sequences with undefined class categories were initially assumed to be false negatives before the test. Therefore, the *Q* values for class obtained by the PBH or the BHPB alignment strategy derived according to sequence identity or secondary structure identity is poor. However, the ENZYME functions of some of these TIM sequences with undefined class categories derived from UniProt have been described in PDB. Thus, the EC annotations derived from PDB were integrated into TIM40D and TIM95D from ASTRAL SCOP 1.71 and 1.73 (see supplemental Table S3 (Chu, 2011)), and the above experiments for the class classification were repeated. After the PDB integrations, 13.6% TIM40D (38 of 279) and 11.1% TIM95D (50 of 450) TIM sequences remained undefined from ASTRAL SCOP 1.71; further, 11.5% TIM40D (6 of 52) and 9.0% TIM95D (6 of 67) novel TIM sequences remained undefined from ASTRAL SCOP 1.73. These six novel TIM sequences were identical in TIM40D and TIM95D from ASTRAL SCOP 1.73.

3.5.1 Improvement in Q

Figure 5 compares the *Q* values for TIM40D and TIM95D from ASTRAL SCOP 1.71 with UniProt and PDB EC annotations using the PBH and BHPB alignment strategies. (see supplemental Table S6 (Chu, 2011)) The *Q* values for the class classification using the PBH and BHPB alignment strategies improved after integrating the PDB EC annotations. By integrating the PDB EC annotations, some of the false negatives from UniProt were eliminated. The alignment approach using either the PBH or BHPB strategy was useful for the class classification. For TIM40D, the best *Q* value of 62.0% (an increase from 48.2%) for class was derived according to sequence identity or secondary structure identity using the PBH alignment strategy; the best *Q* value of 53.4% (an increase from 41.6%) for class was derived according to sequence identity using the BHPB alignment strategy. For TIM95D, the best *Q* value of 78.2% (an increase from 65.2%) for class was derived according to secondary

structure identity using the PBH alignment strategy; the best *Q* value of 72.9% (an increase from 62.2%) for class was derived according to sequence identity or secondary structure identity using the BHPB alignment strategy. For the novel TIM sequences in TIM40D and TIM95D from ASTRAL SCOP 1.73, the best *Q* values were 73.1% (TIM40D) and 79.1% (TIM95D) using the PBH alignment strategy (see supplemental Table S6 (Chu, 2011)).

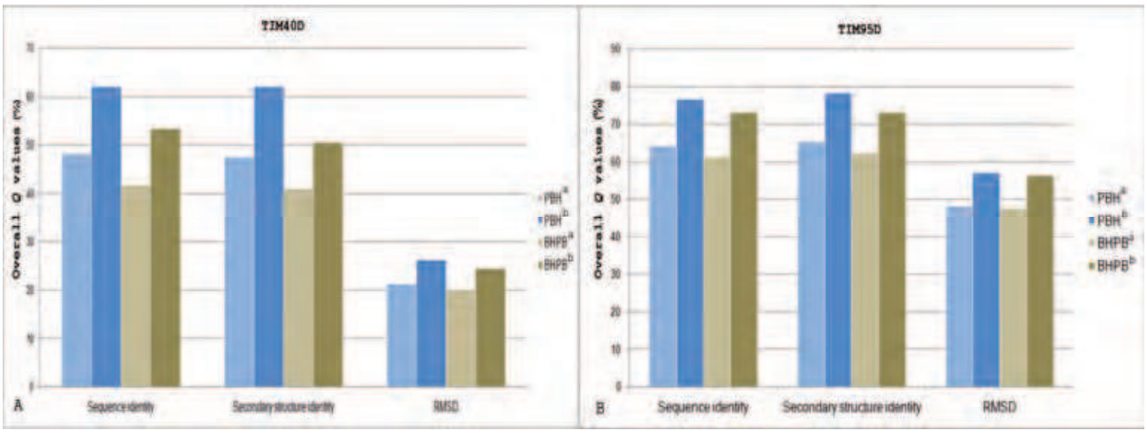


Fig. 5. Comparisons for TIM40D and TIM95D (ASTRAL SCOP 1.71) with UniProt and PDB EC annotations. (A) The *Q* values for TIM40D. (B) The *Q* values for TIM95D. Superscript 'a' or 'b' indicates TIM sequences available using only UniProt EC annotations or TIM sequences available using UniProt and PDB EC annotations.

3.5.2 Improvement in MCC

Table 12 presents *MCC* scores for TIM sequences in TIM40D and TIM95D from ASTRAL SCOP 1.71 with UniProt and PDB EC annotations using the PBH and BHPB alignment strategies. All of the class categories had *MCC* scores between 0.4~0.8 in TIM40D; greater than 0.7 in TIM95D; Oxidoreductases and Lyases also had *MCC* scores greater than 0.9 using the BHPB alignment strategy. Hence, the proposed PBH or BHPB alignment strategy also yielded high prediction quality for class.

	Category	Index	Sequence identity *MCC*		Secondary structure identity *MCC*	
			PBH	**BHPB**	**PBH**	**BHPB**
TIM40D	Class (ENZYME)	1	0.72	0.80	0.75	0.76
		2	0.49	0.47	0.50	0.42
		3	0.68	0.61	0.67	0.57
		4	0.67	0.69	0.67	0.65
		5	0.50	0.53	0.52	0.49
TIM95D	Class (ENZYME)	1	0.89	0.93	0.87	0.89
		2	0.76	0.78	0.81	0.80
		3	0.79	0.74	0.81	0.73
		4	0.87	0.91	0.86	0.90
		5	0.74	0.78	0.81	0.79

Table 12. *MCC* scores for TIM40D and TIM95D (ASTRAL SCOP 1.71) with UniProt and PDB EC annotations

3.5.3 Inferring ENZYME function for TIM barrel proteins with undefined class categories

After integrating the PDB EC annotations into the above tests, there remained 38 (TIM40D) and 50 (TIM95D) TIM sequences with undefined class categories from ASTRAL SCOP 1.71; 6 novel TIM sequences had undefined class categories from ASTRAL SCOP 1.73. Therefore, we used the proposed alignment approach to infer the ENZYME functions for TIM barrel proteins with undefined class.

We first assessed the classification results of the class categories by the PBH alignment strategy for TIM sequences in TIM40D and TIM95D from ASTRAL SCOP 1.71 with UniProt and PDB EC annotations. We found that the target protein and its selected protein belong to the same superfamily category for most of the true positives identified in the alignment. Table 13 presents statistics for true positives and false negatives for class using the PBH alignment strategy. For true positives, 94% (162 of 173) and 99% (342 of 344) of the target and its selected proteins belonged to the same superfamily category derived according to PBH(1D) in TIM40D and TIM95D, respectively. For false negatives, however, 38% (40 of 106) and 31% (33 of 106) of the target and its selected proteins belonged to the same superfamily category derived according to PBH(1D) in TIM40D and TIM95D, respectively.

Statistic	TIM40D				TIM95D			
	PBH(1D)		PBH(2D)		PBH(1D)		PBH(2D)	
	TP_i	FP_i	TP_i	FP_i	TP_i	FP_i	TP_i	FP_i
s, f	154.0	32.0	146.0	37.0	335.0	31.0	332.0	31.0
$s, \overline{f}$	8.0	8.0	13.0	9.0	7.0	2.0	11.0	2.0
$\overline{s}$	11.0	28.0	14.0	22.0	2.0	23.0	9.0	15.0
sum	173.0	106.0	173.0	106.0	344.0	106.0	352.0	98.0

s : Target and its selected proteins belong to the same superfamily category

$\overline{s}$: Target and its selected proteins belong to the different superfamily categories

f : Target and its selected proteins belong to the same family category

$\overline{f}$: Target and its selected proteins belong to the different family categories

Table 13. Statistical results for true positives and false negatives for class using the PBH alignment strategy

Overall, 58% (of 279) and 76% (of 450) of the target and its selected proteins belonged to the same superfamily and class categories derived according to PBH(1D) in TIM40D and TIM95D, respectively. Similar observations were made based on PBH(2D) in TIM40D and TIM95D. We observed 19 (PBH(1D)) and 23 (PBH(2D)) TIM sequences with undefined class categories in TIM40D with the same superfamily category, respectively. We observed 19 (PBH(1D)) and 26 (PBH(2D)) TIM sequences with undefined class categories in TIM95D with the same superfamily category. Therefore, it may be possible to infer the ENZYME functions for TIM barrel proteins with undefined class categories, especially for TIM95D, according to the classification results predicted by the proposed alignment approach. Table 13 also shows that 14% of 279 and 7% of 450 target and selected proteins belong to the same

superfamily category, but they belong to different class categories derived according to PBH(1D) in TIM40D and TIM95D, respectively. Hence, all of TIM sequences of undefined class may not be correctly inferred by the proposed alignment approach with the PBH or the BHPB strategy. In the future, information regarding the active sites will be used in the proposed alignment approach to remedy discrepancies in undefined class. In the following test cases, all of the alignment results were displayed by DS Visualizer (Accelrys). The split structure superposition was displayed utilizing PyMol Molecular Viewer (DeLano, 2002).

4. Methods

4.1 The alignment approach with the PBH strategy

An alignment approach with the PBH strategy was proposed to perform TIM barrel protein domain structure classification (Figure 6). TIM40D and TIM95D can be used as the input for this alignment approach. In the alignment methods block, three alignment tools, CLUSTALW, SSEA and CE, were adopted to align any two of proteins by the amino acid sequences, secondary structures and 3D structures, respectively, to obtain the scores of sequence identity, secondary structure identity and RMSD. CLUSTALW is an established multiple sequence alignment tool (global alignment) for DNA/RNA or protein sequences based on a progressive pair-wise alignment method by considering sequence weighting, variations in amino acid substitution matrices and residue-specific gap penalty scores. It is widely used by biologists to investigate evolutional relationships among multiple protein sequences. CLUSTALW may not be the best choice for the sequence alignment because of recent advancements in programming, but it is still suitable for this alignment approach for two reasons. First, we simply want to obtain the score of sequence identity for any two proteins rather than the actual alignment information. Hence, the sequence identity score obtained by CLUSTALW is not significantly different from that obtained by other tools. Second, the design of most of other tools is focused on revising the multiple sequence alignment results, not improving the pair-wise alignment results, even using the pair-wise alignment results by CLUSTALW. SSEA is a multiple protein secondary structure alignment tool (either global or local alignment) that aligns entire elements (rather than residue-based elements [20]) of multiple proteins based on the H, C, and E states of SSEs. CE is a popular and accurate pair-wise protein 3D structural alignment tool that aligns residues in sequential order in space. If a protein domain sequence is not continuous, however, each continuous fragment in the domain will be aligned against the other protein using the CE alignment tool. Two criteria were adopted to resolve this problem. First, the sequence length of the continuous fragment must be at least 30 residues, and second the minimal RMSD of any two aligned fragments must be chosen. The default parameters of CLUSTALW (accurate, but slow mode in setting your pairwise alignment options) and SSEA (global alignment version) were used to align any two proteins in TIM40D and TIM95D to obtain scores for sequence and secondary structure identities with normalized values ranging from 0-100. The default parameters of CE were used to align any two proteins in TIM40D and TIM95D to obtain RMSD scores. After using CLUSTALW, SSEA and CE, these scores were used to build an alignment-based protein-protein identity score network.

In the best hit strategy block, each protein in the network was first considered as a target protein. Each target protein was then used to map the remaining proteins in the network. Finally, the prediction result of each target protein was determined by selecting the remaining proteins in the network according to certain parameters, which are critical for

classification of the target protein. In our method, a PBH strategy is used to determine the prediction result of a target protein by selecting the protein that has the best score for the target protein according to this network. This score is calculated by a single parameter (sequence identity, secondary structure identity, or RMSD). For the sequence or secondary structure identity, the remaining protein with the highest score for the target protein is selected; for the RMSD, the remaining protein with the lowest score for the target protein is selected. For n proteins in the network, the time complexity is $O(n^2)$ for n target proteins to find all selected proteins in this network using the PBH strategy owing to the bidirectional aspect of the network. We used Perl to implement the PBH finding program because it supports powerful data structures.

The single parameter threshold was applied in this classification model. When a threshold is given for this approach, a target protein is assigned to a null situation as a false negative if the highest score of sequence identity or secondary structure identity (or lowest score of RMSD) for the target protein among all remaining proteins is less than (or larger than, for RMSD) this threshold. Although the overall prediction accuracy cannot be improved by the threshold concept, it may be decreased when an unfavorable threshold is given; however, the number of false positives may be reduced when an appropriate threshold is used. In other words, Precision values may be improved by the threshold concept. Nevertheless, an appropriate threshold is very difficult to attain for the classification problem in a practical setting. Therefore, in the experimental tests, an appropriate threshold was chosen after processing was complete. Using the threshold concept, we observed the best possible Precision values by this alignment approach and the properties of TIM barrel proteins.

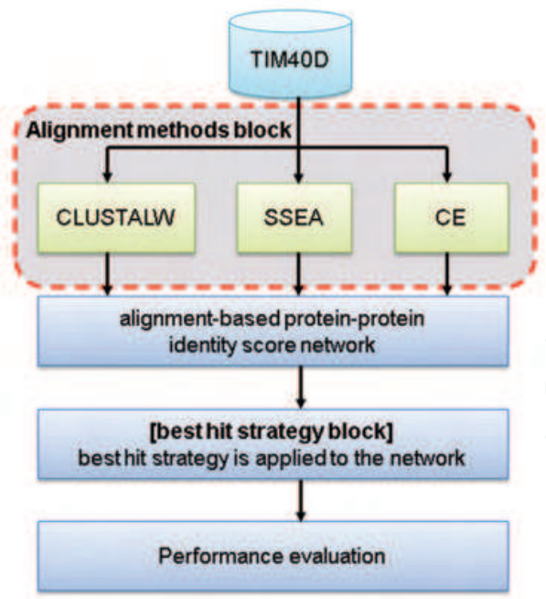


Fig. 6. Flow chart of the alignment approach with the PBH strategy

To experimentally test the novel TIM sequences from ASTRAL SCOP 1.73, the flow chart of the alignment approach with the PBH strategy is slightly different than that shown in Figure 6. For this test, the input is the novel TIM sequences from ASTRAL SCOP 1.73; however, the alignment-based protein-protein identity score network is built by TIM sequences from

ASTRAL SCOP 1.71. Therefore, the target protein is a novel TIM sequence from ASTRAL SCOP 1.73, and the remaining proteins are obtained from the TIM sequences (ASTRAL SCOP 1.71). All tools and materials used for this research are accessible from (Chu, 2011).

4.2 The alignment approach with the BHPB strategy

PSI-BLAST is a position-specific iterative BLAST that results from refinement of the position-specific scoring matrix (PSSM) and the next iterative PSSM. The position-specific scoring matrix is automatically constructed from a multiple alignment with the highest scoring hits in the BLAST search. The next iterative PSSM is generated by calculating position-specific scores for each position in the previous iteration. PSI-BLAST is typically used instead of BLAST to detect subtle relationships between proteins that are structurally distant or functionally homologous. Therefore, it is possible to utilize PSI-BLAST as a filter prior to the PBH strategy, denoted the BHPB strategy. The BHPB strategy can filter out potential false positives, which may improve Precision values. The flow chart of the alignment approach with the BHPB strategy is also slightly different than that shown in Figure 6. In the best hit strategy block, each target protein in the network is used to map a subset, but not all, of the remaining proteins in the network. This subset of remaining proteins is grouped from the network using PSI-BLAST method for the target protein. Hence, the selected protein with the best score for any target protein by the BHPB strategy may not be the same as that by the PBH strategy.

5. Conclusion

At the amino acid sequence level, TIM barrel proteins are very diverse; however, these proteins contain very similar secondary structures. Our results demonstrate that the alignment approach with the PBH strategy or BHPB strategy is a simple and stable method for TIM barrel protein domain structure classification, even when only amino acid sequence information is available.

6. Acknowledgment

Part of this work was supported by National Science Council (NSC) under contract NSC95-2627-B-007-002. The authors would like to thank Shu Hao Chang to help us to collect the TIM barrel proteins from the SCOP 1.71 version.

7. References

Altschul, S.F.; Madden, T.L.; Schäffer, A.A.; Zhang, J.; Zhang, Z.; Miller, W. & Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research,* 12.06.2008, Available from ftp://ncbi.nlm.nih.gov/blast/.

Andreeva, A.; Howorth, D.; Chandonia, J.-M.; Brenner, S.E.; Hubbard, T.J.P.; Chothia, C. & Murzin, A.G. (2008). Data growth and its impact on the SCOP database: new developments. *Nucleic Acids Research*, Vol.36, pp. D419-D425.

Bairoch, A. (2000). The ENZYME database in 2000. *Nucleic Acids Research*, Vol.28, pp. 304-305.

Bairoch, A.I.; Apweiler, R.; Wu, C.H.; Barker, W.C.; Boeckmann, B.; Ferro, S.; Gasteiger, E.; Huang, H.; Lopez, R.; Magrane, M.; Martin, M.J.; Natale, D.A.; O'Donovan, C.; Redaschi, N. & Yeh, L.S. (2005). The Universal Protein Resource (UniProt). *Nucleic Acids Research*, Vol.33, pp. D154-D159.

Bairoch, A.; Boeckmann, B.; Ferro, S. & Gasteiger, E. (2004). Swiss-Prot: Juggling between evolution and stability. *Briefings in Bioinformatics*, Vol.5, pp. 39-55.

Berman, H.M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T.N.; Weissig, H.; Shindyalov, I.N. & Bourne, P.E. (2000). The Protein Data Bank. *Nucleic Acids Research*, Vol.28, pp. 235-242.

Carugo, O. & Pongor, S. (2002). Protein fold similarity estimated by a probabilistic approach based on C_α-C_α distance comparison. *Journal of Molecular Biology*, Vol.315, pp. 887-898.

Chandonia, J.M.; Hon, G.; Walker, N.S.; Lo, C.L.; Koehl, P.; Levitt, M. & Brenner, S.E. (2004). The ASTRAL compendium in 2004. *Nucleic Acids Research*, Vol.32, pp. D189-D192.

Choi, I.; Kwon, J. & Kim, S. (2004). Local feature frequency profile: a method to measure structural similarity in proteins. *Proceeding of the National Academy of Sciences of the United States of America*, Vol.101, pp. 3797-3802.

Chou, K.C. & Zhang, C.T. (1995). Prediction of protein structural classes. *Critical Reviews in Biochemistry and Molecular Biology*, Vol.30, pp. 275-349.

Chu, C.-H. (2011). TIM barrel supplemental data for the InTech bookchapter. *National Tsing Hua Uinversity, Computational Systems Biology & Bio-Medicine Laboratory*, 03.01.2011, Available from http://oz.nthu.edu.tw/~d938301/InTech/bookchapter/

Cuff, A.L.; Sillitoe, I.; Lewis, T.; Redfern, O.C.; Garratt, R.; Thornton, J. & Orengo, C.A. (2009). The CATH classification revisited—architectures reviewed and new ways to characterize structural divergence in superfamilies. *Nucleic Acids Research*, Vol.37, pp. D310-D314.

DeLano, W.L. (2002).The PyMOL Molecular Graphics System, DeLano Scientific, Palo Alto, CA, USA. http://www.pymol.org.

Ding, C.H.Q. & Dubchak, I. (2001). Multi-class protein fold recognition using support vector machines and neural networks. *Bioinformatics*, Vol.17, pp. 349-358.

Dobson, P.D. & Doig, A.J. (2005). Predicting enzyme class from protein structure without alignments. *Journal of Molecular Biology*, Vol.345, pp. 187-199.

Dubchak, I.; Muchnik, I.; Holbrook, S.R. & Kim, S.H.(1995). Prediction of protein folding class using global description of amino acid sequence. *Proceeding of the National Academy of Sciences of the United States of America*, Vol.92, pp. 8700-8704.

Fontana, P.; Bindewald, E.; Toppo, S.; Velasco, R.; Valle, G. & Tosatto, S.C.E. (2005). The SSEA server for protein secondary structure alignment. *Bioinformatics*, Vol.21, pp. 393-395.

Gardy, J.L.; Spencer, C.; Wang, K.; Ester, M.; Tusnday, G.E.; Simon, I.; Hua, S.; deFays, K.; Lambert, C.; Nakai, K. & Brinkman, F.S.L. (2003). PSORT-B: improving protein subcellular localization prediction for gram-negative bacteria. *Nucleic Acids Research*, Vol.31, pp. 3613-3617.

Gáspári, Z.; Vlahovicek, K. & Pongor, S. (2005). Efficient recognition of folds in protein 3D structures by the improved PRIDE algorithm. *Bioinformatics*, Vol.21, pp. 3322-3323.

Gloster, T.M.; Roberts, S.; Ducros, V.M.-A.; Perugino, G.; Rossi, M.; Hoos, R.; Moracci, M.; Vasella, A. & Davies, G.J. (2004). Structural studies of the β-Glycosidase from

Sulfolobus solfataricus in complex with covalently and noncovalently bound inhibitors. *Biochemistry*, Vol.43, pp. 6101-6109.

Huang, C.D.; Lin, C.T. & Pal, N.R. (2003). Hierarchical learning architecture with automatic feature selection for multi-class protein fold classification. *IEEE Transactions on NanoBioscience*, Vol.2, pp. 503–517.

Jones, D.T. (1999). Protein secondary structure prediction based on position-specific scoring matrices. *Journal of Molecular Biology*, Vol.292, pp. 195-202.

Kabsch, W. & Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, Vol.22, pp. 2577-2637.

Lin, K.L.; Lin, C.Y.; Huang, C.D.; Chang, H.M.; Yang, C.Y.; Lin, C.T.; Tang, C.Y. & Hsu, D.F. (2005). Methods of improving protein structure prediction based on HLA neural network and combinatorial fusion analysis. *WSEAS Transations on Information Science and Applications*, Vol.2, pp. 2146-2153.

Lin, K.L.; Lin, C.Y.; Huang, C.D.; Chang, H.M.; Yang, C.Y.; Lin, C.T.; Tang, C.Y. & Hsu, D.F. (2007). Feature combination criteria for improving accuracy in protein structure prediction. *IEEE Transactions on NanoBioscience*, Vol.6, pp. 186-196.

Matthews, B.W. (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et Biophysica Acta*, Vol.405, pp. 442-451.

Murzin, A.G.; Brenner, S.E.; Hubbard, T. & Chothia, C. (1995). SCOP: A structural classification of proteins database for the investigation of sequence and structures. *Journal of Molecular Biology*, Vol.247, pp. 536-540.

Rogen, P. & Fain, B. (2003). Automatic classification of protein structure by using gauss integrals. *Proceeding of the National Academy of Sciences of the United States of America*, Vol.100, pp. 119-124.

Rost, B. & Sander, C. (1993). Prediction of protein secondary structure at better than 70% accuracy. *Journal of Molecular Biology*, Vol.232, pp. 584-599.

Shen, H.B. & Chou, K.C. (2006). Ensemble classifier for protein fold pattern recognition. *Bioinformatics*, Vol.22, pp. 1717-1722.

Shindyalov, I.N. & Bourne, P.E. (1998). Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Engineering*, Vol.11, pp. 739-747.

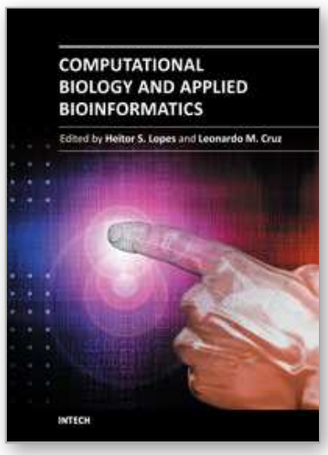
Thompson, J.D.; Higgins, D.G. & Gibson, T.J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. *Nucleic Acids Research*, Vol.22, pp. 4673-4680.

Vapnik, V.N. (1995). *The Nature of Statistical Learning Theory*, New York: Springer-Verlag.

Yu, C.S.; Wang, J.Y.; Yang, J.M.; Lyu, P.C.; Lin, C.J. & Hwang, J.K. (2003). Fine-grained protein fold assignment by support vector machines using generalized *n*Peptide coding schemes and jury voting from multiple-parameter sets. *Proteins*, Vol.50, pp. 531-536.

Zotenko, E.; Dogan, R.I.; Wilbur, W.J.; O'Leary, D.P. & Przytycka, T.M. (2007). Structural footprinting in protein structure comparison: The impact of structural fragments. *BMC Structural Biology*, Vol.7, pp. 53.

Zotenko, E.; O'Leary, D.P. & Przytycka, T.M. (2006). Secondary structure spatial conformation footprint: a novel method for fast protein structure comparison and classification. *BMC Structural Biology*, Vol.6, pp. 12.



Computational Biology and Applied Bioinformatics

Edited by Prof. Heitor Lopes

ISBN 978-953-307-629-4

Hard cover, 442 pages

Publisher InTech

Published online 02, September, 2011

Published in print edition September, 2011

Nowadays it is difficult to imagine an area of knowledge that can continue developing without the use of computers and informatics. It is not different with biology, that has seen an unpredictable growth in recent decades, with the rise of a new discipline, bioinformatics, bringing together molecular biology, biotechnology and information technology. More recently, the development of high throughput techniques, such as microarray, mass spectrometry and DNA sequencing, has increased the need of computational support to collect, store, retrieve, analyze, and correlate huge data sets of complex information. On the other hand, the growth of the computational power for processing and storage has also increased the necessity for deeper knowledge in the field. The development of bioinformatics has allowed now the emergence of systems biology, the study of the interactions between the components of a biological system, and how these interactions give rise to the function and behavior of a living being. This book presents some theoretical issues, reviews, and a variety of bioinformatics applications. For better understanding, the chapters were grouped in two parts. In Part I, the chapters are more oriented towards literature review and theoretical issues. Part II consists of application-oriented chapters that report case studies in which a specific biological problem is treated with bioinformatics tools.

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820

Fax: +385 (51) 686 166
www.intechopen.com

Fax: +86-21-62489821

IntechOpen