BIOINFORMATICS ORIGINAL PAPER Vol. 26 no. 8 2010, pages 1022–1028
doi:10.1093/bioinformatics/btq084

Sequence analysis

Advance Access publication February 26, 2010

Feature-incorporated alignment based ligand-binding residue prediction for carbohydrate-binding modules

Wei-Yao Chou[1,2,3], Wei-I Chou[2,3,4], Tun-Wen Pai[5,6], Shu-Chuan Lin[2,3], Ting-Ying Jiang[2,3], Chuan-Yi Tang[1] and Margaret Dah-Tsyr Chang[2,3,*]

[1]Department of Computer Science, [2]Institute of Molecular and Cellular Biology, [3]Department of Life Science, National Tsing Hua University, Hsinchu, Taiwan 300, [4]Simpson Biotech Co., Ltd, Taoyuan County, Taiwan 333, [5]Department of Computer Science and Engineering and [6]Center for Marine Bioscience and Biotechnology, National Taiwan Ocean University, Keelung, Taiwan 202, Republic of China

Associate Editor: Martin Bishop

ABSTRACT

Motivation: Carbohydrate-binding modules (CBMs) share similar secondary and tertiary topology, but their primary sequence identity is low. Computational identification of ligand-binding residues allows biologists to better understand the protein–carbohydrate binding mechanism. In general, functional characterization can be alternatively solved by alignment-based manners. As alignment accuracy based on conventional methods is often sensitive to sequence identity, low sequence identity among query sequences makes it difficult to precisely locate small portions of relevant features. Therefore, we propose a feature-incorporated alignment (FIA) to flexibly align conserved signatures in CBMs. Then, an FIA-based target-template prediction model was further implemented to identify functional ligand-binding residues.

Results: *Arabidopsis thaliana* CBM45 and CBM53 were used to validate the FIA-based prediction model. The predicted ligand-binding residues residing on the surface in the hypothetical structures were verified to be ligand-binding residues. In the absence of 3D structural information, FIA demonstrated significant improvement in the estimation of sequence similarity and identity for a total of 808 sequences from 11 different CBM families as compared with six leading tools by Friedman rank test.

Contact: dtchang@life.nthu.edu.tw

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on December 12, 2009; revised on February 19, 2010; accepted on February 22, 2010

1 INTRODUCTION

Carbohydrate-binding modules (CBMs) are defined as a set of protein domains capable of recognizing and binding polysaccharide ligands. With the exception of some lectins and sugar transport proteins, most CBMs mediate the interaction between the substrate and the enzyme, which in turn increases local substrate concentration at the active site of the catalytic domain (Southall *et al.*, 1999). CBMs are protein domains sharing low sequence identities that possess conserved structure topology. Up to now, CBMs have been classified into 59 families (still growing) by CAZy database (http://www.cazy.org/). Among these families, CBM20, CBM21, CBM25, CBM26, CBM34, CBM41, CBM45, CBM48 and CBM53 are reported to possess starch-binding activity. The most recent review comprehensively introduces the importance of functions and structures of CBM20 (Christiansen *et al.*, 2009). The primary sequence identities of CBM families are low; however, bioinformatics analysis suggests that some CBMs constitute a CBM clan such as CBM20s, CBM21s, CBM48s and CBM53s. A better understanding of ligand-binding residues in CBMs can boost protein engineering for industrial development in food processing and biofuel production (Guzman-Maldonado and Paredes-Lopez, 1995; Schmidt and Dauenhauer, 2007).

To predict ligand-binding residues in a protein domain, protein-ligand docking programs are typically applied to simulate the binding site of a target protein relative to a ligand conformation and orientation. In general, docking programs require precise atom coordinate information to calculate the optimal geometric pose under stereo-chemical restraints and are usually based on certain criteria such as binding affinity and minimal free energy (Thomsen and Christensen, 2006; Yang and Chen, 2004). Nevertheless, the time complexity of the optimized procedure is extremely high such that docking programs are commonly solved by heuristic algorithms like the genetic algorithm. Besides, real protein–ligand interactions always involve conformational change. When the structural flexibility is also taken into account, the computational complexity becomes prohibitively complicated. Furthermore, although more than 10.1 million protein sequences are currently available in UniProtKB/TrEMBL database (Boutet *et al.*, 2007), as of January 2010, only 62 926 protein structures have been deposited in Protein Data Bank (Berman *et al.*, 2007). Thus, the number of resolved protein structures is far behind that of protein sequences. In reality, it is expensive and time consuming to resolve a protein structure and non-productive to determine all protein structures. Nevertheless, the lack of a resolved protein structure does not hinder researchers from identifying crucial functional residues governing reaction mechanisms. When tertiary structural information is not available, it is still possible to predict ligand-binding residues from sequence alone. Recently, some pure sequence-based prediction methodologies applied sequence alignment to generate a column of conserved residues and calculated the residue conservation score based on training with known functional residues (Capra and Singh, 2007; Chen and Jeong, 2009; Fischer *et al.*, 2008). In terms of

*To whom correspondence should be addressed.

alignment implementation, the global optimal pairwise alignment can be solved in $O(n^2)$, where n stands for the length of the longer sequence (Needleman and Wunsch, 1970). In the last decades, the most widely used alignment programs were heuristic or approximate based on different improvement techniques. Among them, ClustalW is designed mainly based on gap penalty adjustments (Larkin *et al.*, 2007). T-COFFEE constructs an alignment library by weighting the consistency of ClustalW (global alignment) and Lalign (local alignment) (Notredame *et al.*, 2000). MUSCLE first constructs a draft alignment, and then refines it using a variant of tree-dependent restricted partitioning (Edgar, 2004). DIALIGN-TX identifies local alignments with ungapped segment comparisons to greedily evolve a multiple alignment (Subramanian *et al.*, 2008). ProbCons utilizes an assortment of probabilistic modeling and consistency-based alignment techniques to increase alignment accuracy (Do *et al.*, 2005). MAFFT applies fast Fourier transform to rapidly locate homologous regions and simplifies the scoring function to reduce computational time (Katoh *et al.*, 2002). Generally speaking, among these six mentioned tools, ProbCons achieves the highest accuracy and MUSCLE is the most efficient on BAliBASE, a standard alignment benchmark for various sequence properties (Bahr *et al.*, 2001).

However, one common deficiency in sequence alignment is that the alignment accuracy relies heavily on sequence identity (Yang and Honig, 1999). Low sequence identity implies that only few positions are conserved and thus difficult to be recognized. Moreover, a simplified mathematic model may not reflect the real complicated biological model. Therefore, instead of designing a generalized model, biologists are concerned about the functionally meaningful features. Previous studies show that hydrophobic stacking interactions of aromatic residues and hydrogen bonding of polar amino acids in CBMs confer essential roles in ligand-binding recognition (Boraston *et al.*, 2004; Pell *et al.*, 2003; Ponyi *et al.*, 2000; Xie *et al.*, 2001), and secondary structure is the core topology conserved in CBMs. Therefore, based on the observations of the conserved features in CBM families, we first designed a feature-incorporated alignment (FIA) to flexibly anchor aromatic residues with adjacent polar residues with β-stranded structures as the most conserved features. Based on FIA, we further developed a ligand-binding residue prediction system employing the target-template comparison model. Our contributions are 2-fold: superior alignment accuracy and a pure sequence-based prediction model for *in silico* identification of ligand-binding residues in CBMs.

2 SYSTEM AND METHODS

2.1 System overview

The central idea of the proposed ligand-binding residue prediction model was to locate key conserved secondary structure elements and aromatic residues in CBM families. FIA was designed to flexibly anchor aromatic residues with adjacent polar residues and β-stranded structures. In addition, annotated sequences with reported ligand-binding residues were adopted as templates to detect the conserved ligand-binding residue in the target sequence. The proposed ligand-binding residue prediction system for CBM families can be divided into four modules as depicted in Figure 1. The first two steps are preprocesses to predict secondary structures and to annotate the aromatic amino acids with adjacent polar residues. The last two detecting stages are applied for pairwise alignment of the target sequence against the annotated sequences and to evaluate the aligned aromatic residues in the target sequence.

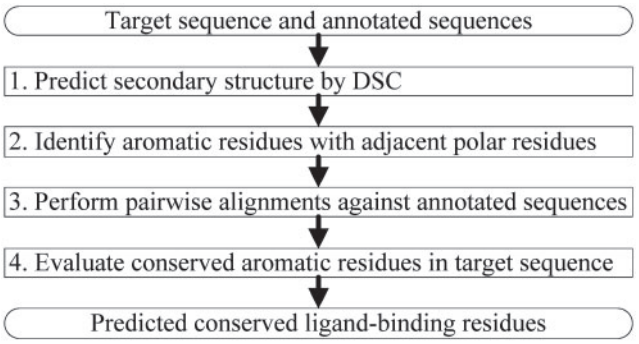

Fig. 1. Dataflow of the FIA-based prediction model.

2.2 Predicting secondary structure by discrimination of protein secondary structure class

Secondary structures are better conserved than loop regions in a tertiary structure, and the β-stranded element is a key conserved structure topology in CBM families (Hashimoto, 2006). These features serve as a good constraint core segment in sequence alignment. Therefore, we applied Discrimination of protein Secondary structure Class (DSC) (King and Sternberg, 1996). DSC achieved overall three-state accuracy of 70.1% and performed best in enzyme size of 90–170 kDa. DSC predicts the probabilities of an amino acid in a loop, α-helix and β-sheet for primary sequences. When the reported probability of an amino acid in a β-sheet was higher than 60%, the amino acid would be annotated as 'in β-strand' in this study. In addition, the predicted β-sheets with lengths less than two did not meet the annotation criteria.

2.3 Identifying aromatic residues with adjacent polar residues

A previous study indicated that aromatic ligand-binding residues with adjacent polar residues may form a cleft to bind a carbohydrate ligand (Tormo *et al.*, 1996). Polar residues usually exhibit higher hydrophilicity when exposed on the outer protein surface, and we have observed that the reported aromatic ligand-binding residues of CBMs always reside in motifs containing polar residues. Hence, this stage identified the aromatic residues with adjacent polar residues within two amino acids in neighboring, and such an aromatic residue was defined as a hydrophilic aromatic residue. These hydrophilic aromatic residues W, F and Y were denoted by the new terms W_h, F_h and Y_h, respectively. Then, the original 20 amino acids were extended to 23 amino acids for the following procedures in this study. These two conserved features in CBMs representing characteristic core ligand-binding regions are expected to be aligned in sequence alignment.

2.4 Performing FIA against template sequences

In this stage, the system executed pairwise alignments of the target sequence against relevant template sequences to accumulate conserved information for the last evaluation stage. The standard dynamic programming approach for global alignment (Needleman and Wunsch, 1970) with an affine gap penalty (Gotoh, 1982) was applied. Needleman's algorithm guarantees the generation of an optimal solution and the affine gap penalty is a better gap model reflecting real evolutionary mechanisms without increasing time complexity. Moreover, the goodness and fitness of the scoring function affects the alignment accuracy. Therefore, in this study, the scoring function was redesigned and focused on hydrophilic aromatic amino acids and secondary structure conservation as described in the previous sections. The pairwise version of FIA definition is declared as follows. Suppose that two

sequences of the target sequence and the annotated sequence over 23 amino acids including the three newly defined hydrophilic amino acids are denoted by S_t and S_a, respectively, where $|S_t|$ and $|S_a|$ stand for the lengths of S_t and S_a, respectively. $S_t[i]$ indicates the i-th amino acid in S_t. The goal of FIA is to optimize the alignment with properly inserted gaps within S_t and S_a. The affine penalty model requires three two-dimensional arrays of M, I and D to compute and store temporary sub-solutions. $M[i][j]$, $I[i][j]$ and $D[i][j]$ represent the optimal alignment arrangements for the first i amino acids of S_t and the first j amino acids of S_a ending with substitution, insertion and deletion, respectively. In other words, the dynamic approach recursively computed sub-solutions to optimize $M[|S_t|][|S_a|]$ from Equation (1) for the two complete query sequences.

$$M[i][j] = \max\left\{\begin{array}{l} M[i-1][j-1]+\sigma(S_t[i],S_a[j]) \\ I[i-1][j]+\sigma(S_t[i],S_a[j]) \\ D[i][j-1]+\sigma(S_t[i],S_a[j]) \end{array}\right\}$$

$$I[i][j] = \max\left\{\begin{array}{l} M[i-1][j-1]-\sigma('-',S_a[j]) \\ I[i-1][j]-d \end{array}\right\} \quad (1)$$

$$D[i][j] = \max\left\{\begin{array}{l} M[i-1][j-1]-\sigma(S_t[i],'-') \\ D[i][j-1]-d \end{array}\right\}$$

$$\text{for } 1 \le i \le |S_t|, 1 \le j \le |S_a|$$

$$\sigma(a,b) = \left\{\begin{array}{l} \text{Case 1: EBM}(a,b)^*2, \\ \text{if } a \text{ and } b \text{ are both in } \beta\text{-sheet \&\& EBM}(a,b) > 0 \\ \text{Case 2: 0,} \\ \text{if } a \text{ and } b \text{ are in both } \beta\text{-sheet \&\& EBM}(a,b) < 0 \\ \text{Case 3: EBM}(a,b)^*2, \\ \text{if } a \text{ or } b \text{ is a gap in } \beta\text{-sheet} \\ \text{Case 4: EBM}(a,b), \\ \text{otherwise} \end{array}\right\} \quad (2)$$

In the original affine penalty model, each entry is calculated among the maximum value of the three conditions from M, I and D, respectively, where d represents the gap extension penalty. In our implementation, to avoid the abnormal cases of insertion concatenating deletion and deletion concatenating insertion, the calculations of the values in $I$and D ignore the conditions from D and I, respectively. In terms of the scoring function, the raw substitution value for two amino acids is obtained from the BLOSUM62 matrix, a substitution matrix containing over 20 amino acids for sequence alignment (Henikoff and Henikoff, 1992). By introducing three additional hydrophilic aromatic residues and the gap symbol, the original BLOSUM 62 matrix is extended to a 24×24 matrix named the Extended BLOSUM62 Matrix (EBM) as shown in Supplementary Table S1. When two hydrophilic aromatic residues align to each other, the substitution value is doubled from the ordinary aromatic residues. In addition, the substitution value for a residue and a gap is defined as the lowest substitution value, -4 in BLOSUM62. Subsequently, weighted and penalized substitution values vary based on different properties as described in Equation (2). The substitution value of two residues in β-strands is also increased. In addition, an insertion or a deletion in a β-strand is a serious mutation, so the substitution value for an amino acid and a gap in a β-strand element is double penalized. Through the above substitution value adjustments, even in a low sequence identity situation, the most important features of hydrophilic aromatic residues and β-strands are expected to be located better.

2.5 Evaluating aligned aromatic residues

After the target sequence is aligned pairwise to each annotated sequence, the final stage is to evaluate the alignment of the aromatic residues in the target sequence with the reported ligand-binding residues in annotated sequences. Based on homologous conservation, it is assumed that if an aromatic amino acid in the target sequence is aligned to a reported ligand-binding residue in an annotated sequence, this aromatic amino acid is likely to serve as a ligand-binding residue. The aligned aromatic residues in the target sequence with higher chances of aligning to reported ligand-binding residues in different templates are considered as the most possible candidates for ligand-binding residues. In implementation, the merging processes sum up the probabilities of adjacent aligned aromatic residues within a sliding window size of five into one representative aromatic residue. The final predicted ligand-binding residues are determined by the consistency from aligned aromatic residues in the target sequence, and the consistent rate for a putative ligand-binding residue is formulated in Equation (3) where R_p and R_c represent a putative ligand-binding residue and a conserved ligand-binding residue, respectively. Then, three confidence levels of consistency with template sequences (highly conserved, relatively conserved and unidentified) are defined. The highly conserved aromatic residues indicate the most promising ligand-binding residues possessing at least 50% consistency. The relatively conserved aromatic residues represent the median confidence level of putative aromatic residues with <50% consistency. The unidentified aromatic residues signify a lack of consistency between the aromatic residues in the target sequence and templates.

$$\text{Consistent rate}(R_p,R_c) = \frac{\text{frequency of } R_p \text{ aligning to } R_c}{\text{frequency of } R_c \text{ in templates}} * 100\% \quad (3)$$

2.6 Materials

Employing CATH (Orengo *et al.*, 1997), a hierarchical protein structure classification tool, the proteins in CBM families classified into *2.60.40.10* (immunoglobulin topology) with structural information were employed as the templates. Among the template protein set, those structures with high sequence identity >30% were filtered out and only five CBM structures including the starch-binding domains derived from *Rhizopus oryzae* glucoamylase (*Ro*|2v8mA00) (Tung *et al.*, 2008), *Aspergillus niger* glucoamylase (*An*|1ac0A00) (Sorimachi *et al.*, 1997), *Thermoactinomyces vulgaris* α-amylase-I (*Tv*|1uh3A01) (Abe *et al.*, 2004), *Thermoactinomyces vulgaris* α-amylase-II (*Tv*|1bvzA01) (Kamitori *et al.*, 1999) and *Sulfolobus solfataricus* α-amylase (*Ss*|1eh9A03) (Feese *et al.*, 2000) remained as templates (2v8m is not classified in CATH yet, but it appears to share similar structural topology as well as functional residues in *2.60.40.10*). The remaining template structures are considered to possess representative and non-redundant characteristics and their corresponding ligand-binding residues were previously identified (Chou *et al.*, 2009; Liu *et al.*, 2007; Tung *et al.*, 2008) as shown in Supplementary Figure S1. The CBM domains of the template structures and the target proteins were assigned by CATH and GenBank (Benson *et al.*, 2009), respectively. The secondary structures were predicted by DSC. For CBM45 and CBM53 family members, not a single 3D structure has yet been resolved, hence two case studies of domain sequences from *Arabidopsis thaliana* water dikinase (*At*CBM45) and starch synthase (*At*CBM53) were applied to demonstrate the prediction model and to construct hypothetical structures. In addition, the performance of FIA-based prediction for both cases was compared with that of FRpred which combine information from the conservation at each site, its amino acid distribution, as well as its predicted secondary structure and relative solvent accessibility (Fischer *et al.*, 2008). The *At*CBM45 catalyzes the transfer of the β-phosphate of adenosine triphosphate (ATP) to either C-3 or C-6 of the glucosyl residue for starch phosphorylation (Mikkelsen *et al.*, 2006). The *At*CBM 53 contains a three repeated starch-binding domains at the N-terminal portion of starch synthase III (SSIII) involved in plant starch synthesis and plays a regulatory role in the synthesis of transient starch (Valdez *et al.*, 2008). In terms of broad alignment accuracy comparison, 808 sequences lacking of 3D structures were collected from 11 CBM families including CBM4, CBM6, CBM9, CBM20, CBM21, CBM25, CBM26, CBM34, CBM41, CBM48, CBM53, among which 38, 63, 36, 131, 66, 15, 5, 59, 65, 317 and 13 domain sequences, respectively, were thoroughly studied. Each sequence was randomly chosen as the representative for a species without redundancy. Note that CBM45 was omitted because of incomplete CBM domain assignment. In addition, the performance of FIA on these query sequences was compared with six leading alignment tools,

MUSCLE v3.6, ClustalW v2.0.11, DIALIGN-TX v1.0.2, T-COFFEE v5.31, ProbCons v1.12 and MAFFT v6.710b, using the default parameter settings. Finally, the Friedman rank test (Milton, 1937) was applied to verify the statistically significant differences of alignment accuracy between each pair of all alignment tools by SPSS v15. The resolved and hypothetical protein structures were rendered by WebLab ViewerPro v4.

3 RESULTS

3.1 FIA illustration

In general, the dynamic programming approach for sequence alignment optimizes sequential combination of insertion, deletion and substitution. The scoring function affects how this combination can be generated. Many algorithms aim to anchor similar short regions in alignment (Chou *et al.*, 2006; Edgar, 2004; Katoh *et al.*, 2002; Subramanian *et al.*, 2008). The goal of this novel approach is to flexibly anchor the key conserved regions according to their physiochemical properties. When two sequences share low sequence identity, such positions become more difficult to align. To increase the likelihood for alignment of conserved features, the substitution values of the conserved properties in CBMs are specifically weighted in FIA. As shown in Figure 2, the square and the bold arrows were extra weighted. In other words, if a path is through these two hydrophilic aromatic residues, a higher substitution value is earned. On the other hand, the scoring for an insertion in a β-strand shown in dotted arrow in Figure 2 paid an extra penalty whereas two residues in two β-strands gained extra scores [refer to Equation (2)].

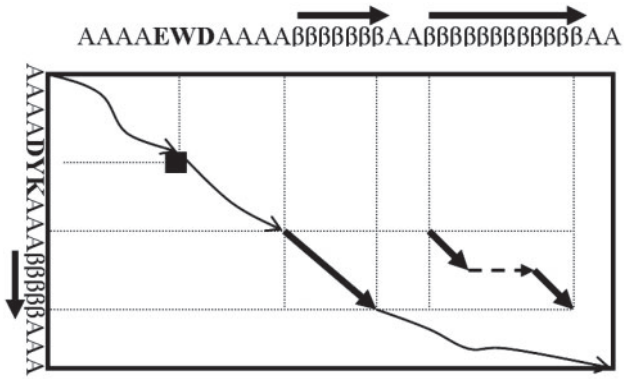

Fig. 2. Illustration of pairwise alignment process related to FIA. The position in square and the sub-path in bold arrow are weighted while the double-penalized indel in the β-sheet is highlighted in dotted arrow.

3.2 Prediction and verification of ligand-binding residues in CBM45 and CBM53

The ultimate goal of this study is not only to devise an accurate alignment program but also to predict conserved ligand-binding residues for protein sequences in CBMs based on confirmed ligand-binding residues. Then, FIA-based ligand-binding prediction was derived from conserved aromatic residues in hydrophilic regions. From the CBM45 and CBM53 family members, two target sequences of *At*CBM45 and *At*CBM53 were aligned pairwise against five template sequences to predict their conserved ligand- binding residues. Figure 3 demonstrates two FIA pairwise alignments between *At*CBM45 and *An*|1ac0A00, as well as between *At*CBM53 and *Ro*|2v8mA00. The FIA calculation led to sequence identities of 22.2 and 29.2% for these two cases. As expected, the key aromatic residues and β-stranded elements were well aligned in columns. Trp^{139} and Trp^{194} in *At*CBM45 and Trp^{340}, Trp^{366}, Phe^{383} and Tyr^{394} in *At*CBM53 were predicted as ligand-binding residues according to the consistent performance of the five pairwise alignments. (The comprehensive analysis can be found in Supplementary Figure S2.) To confirm significance at the structural level, the two alignments were applied to generate hypothetical structures. Figure 4 illustrates the two modeled structures from Swiss-Model by incorporating the two FIA alignments in Figure 3. Note that Swiss-Model originally failed to simulate using its first automated mode. The reported ligand-binding residues in the template structures are highlighted in color sticks and the predicted ligand-binding aromatic residues are highlighted in the corresponding color sticks in the two hypothetical structures. Interestingly, Trp^{139} and Trp^{194} in *At*CBM45 (Mikkelsen *et al.*, 2006) and Trp^{366} and Tyr^{394} in *At*CBM53 (Wayllace *et al.*, 2010) have been experimentally confirmed as true ligand-binding residues. Even the functional role of Trp^{340} and Phe^{383} in *At*CBM53 have not been reported, these two aromatic residues are evidently located on the outer surface in an orientation similar to Tyr^{94} and Tyr^{83}

target-> >gi|13124867:92-207 starch associated protein R1 *Solanum tuberosum* [*Arabidopsis thaliana*] (*At*CBM45)
template-> *An*|1ac0A00
sequence similarity->43.5%, sequence identity->22.2%, bit score->54.68, e value->4.33e–013

target-> gi|4835785:268-406 soluble-starch-synthase precursor (SSIII) from *Solanum tuberosum* [*Arabidopsis thaliana*] (*At*CBM53)
template->*Ro*|2v8mA00
sequence similarity->48.1%, sequence identity->29.2%, bit score->70.09, e value->1.17e–017

Fig. 3. The FIA pairwise alignments of two domain sequences from CBM45 and CBM53 against two annotated structures. The predicted β-stranded elements in *At*CBM45 and *At*CBM53 were shaded by grey boxes. '+' and ':' indicate exact and approximate match, respectively. The reported ligand-binding residues in the templates are numbered. The hydrophilic Phe, Trp and Tyr are represented by J, O and Z, respectively. '−' represents insertion or deletion gap in alignment.

in *Ro*|2v8mA00. These findings suggest that these two conserved hydrophilic aromatic residues are very likely to be potential ligand-binding residues to be further investigated. In addition, direct comparison of the performance between our FIA-based method and FRpred was carried out and summarized in Supplementary Table S3. FRpred is a machine learning methodology that combines information from the conservation at each site, its amino acid distribution, as well as its predicted secondary structure and relative solvent accessibility. Nine CBMs with experimentally confirmed ligand-binding residues were collected as test dataset. While both FIA-based method and FRpred can identify true ligand-binding residues, our FIA-based method achieved a significantly higher positive predictive value (PPV) of 56.3% than FRpred with PPV of 35.3%. Hence less try-and-error experimental efforts would be needed for hands on site-directed mutagenesis.

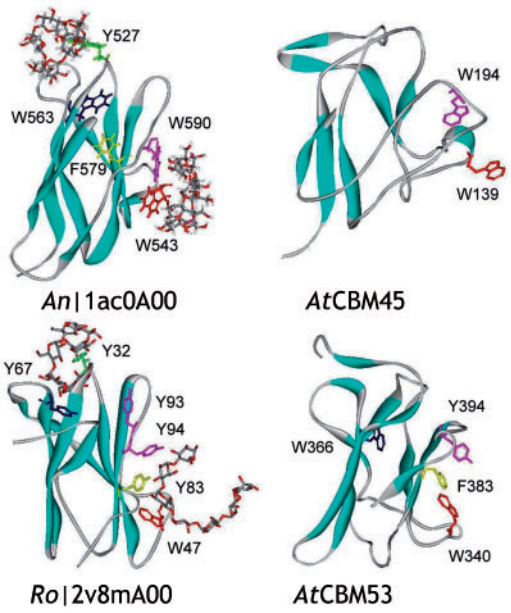

Fig. 4. Structure modeling of *At*CBM45 and *At*CBM53. The two experimentally resolved structures with bound ligands are placed on the left and the two hypothetical structures are located on the right. The reported ligand-binding residues are highlighted in color sticks in the resolved structures on the left, and the corresponding aromatic residues are highlighted in sticks in the hypothetical structures on the right.

3.3 Alignment accuracy comparison

In the aspect of computational performance, we also compared the alignment accuracy for sequence similarity and identity among FIA and six leading alignment programs. Table 1 summarizes the average sequence similarities and sequence identities produced from FIA, DIALIGN-TX, MUSCLE, T-COFFEE, ClustalW2, ProbCons and MAFFT. A total of 808 sequences without structural information were classified into 11 families, and each sequence was aligned to the five template structures in a pairwise manner. That is, the performance of all seven tools was calculated from 4040 (808*5) pairwise alignments. FIA achieved the highest average sequence similarity (45.1%) and average sequence identity (27.2%), whereas DIALIGN-TX gave the lowest average sequence similarity of 29.9% and ClustalW2 showed the lowest average sequence identity of 16.6%. Hence FIA improved the sequence similarity and identity estimation in these CBM families. The superior alignment accuracy indicates that the incorporated conservative properties are informative under the condition of low sequence identity in CBMs. Following the statistical analysis in MUSCLE and ProbCons, the statistically different significance was determined by *P*-values of the Friedman rank test in terms of sequence identity and sequence similarity as listed in Supplementary Table S4. Referring to the performance comparison listed in Table 1, it is clear that FIA obviously and significantly outperformed the other six well-known tools with *P*-values less than 1.0E–10 in all CBM families tested. Besides, the performance of FIA on benchmark protein sequences was also tested using BAliBASE in comparison with the same six alignment tools in terms of sum-of-pair score. Supplementary Table S4 indicated an average of 5.28% variation between FIA and the other methods. Due to the advantage of incorporating specific features for comparison of query sequences with low sequence identity, FIA outperformed ClustalW2, T-COFFEE and DIALIGN-TX in analyzing group ref3 data in BAliBASE with sequences sharing <25% identity, as expected.

Table 1. Comparison of average sequence similarity and identity between FIA and various alignment methods

	No. of seqs	FIA	MUSCLE	ClustalW2	DIALIGN-TX	T-COFFEE	ProbCons	MAFFT
CBM4	38	45.3 (27.3)	37.2 (18.5)	33.1 (15.4)	28.8 (17.0)	34.6 (18.6)	33.9 (18.9)	38.9 (22.8)
CBM6	63	42.4 (25.9)	34.8 (17.5)	31.3 (14.6)	26.5 (15.8)	33.0 (17.7)	31.9 (18.1)	36.6 (21.7)
CBM9	36	50.5 (31.2)	39.9 (19.9)	36.8 (17.6)	29.4 (17.4)	37.1 (20.1)	35.9 (19.9)	42.3 (25.1)
CBM20	131	47.4 (28.9)	40.3 (21.7)	38.6 (20.0)	34.4 (21.2)	37.9 (21.5)	37.4 (21.8)	41.7 (24.9)
CBM21	66	45.7 (26.2)	38.4 (19.4)	35.0 (17.2)	32.4 (19.1)	36.0 (19.7)	35.4 (20.0)	39.6 (23.2)
CBM25	15	44.4 (27.7)	36.7 (19.6)	34.6 (16.7)	29.3 (18.3)	34.7 (20.0)	33.7 (19.1)	38.0 (22.7)
CBM26	5	48.4 (31.2)	37.0 (20.8)	37.5 (18.4)	32.6 (19.8)	36.1 (21.5)	34.9 (21.0)	40.7 (25.3)
CBM34	59	46.1 (27.4)	38.5 (20.1)	35.8 (18.0)	33.7 (19.8)	33.6 (20.1)	36.3 (20.5)	40.2 (23.2)
CBM41	65	42.7 (26.6)	34.5 (17.4)	31.9 (15.2)	27.4 (16.7)	32.4 (17.8)	31.0 (17.7)	35.8 (21.7)
CBM48	317	44.3 (26.6)	35.5 (17.6)	33.1 (15.5)	28.2 (17.0)	33.2 (17.9)	32.4 (18.1)	37.2 (21.9)
CBM53	13	44.3 (24.3)	36.5 (17.5)	33.9 (15.0)	29.2 (16.7)	34.2 (17.5)	34.4 (18.0)	38.8 (21.3)
Average	808	45.1 (27.2)	36.9 (18.8)	34.3 (16.6)	29.9 (18.0)	34.4 (19.0)	33.9 (19.2)	38.6 (22.8)

In *s*(*i*) format, *s* and *i* represent sequence similarity and identity, respectively.

3.4 Application of FIA on non-CBM proteins

FIA has been demonstrated in successful identification of hydrophilic aromatic residues and correlation of key functional residues in CBMs with main β-stranded structures. A more challenging task is to investigate whether it is applicable to proteins with different structural architectures. Here two non-CBM protein families were collected for demonstrating the generalization of this idea. As illustrated in Supplementary Figure S3, FIA analysis of HSP20-like chaperones superfamily possessing β-sheet-rich domains leads to identification of Trp^{48} in 1GME:A as a conserved hydrophilic aromatic residue. This Trp^{48} has been proven as a true ligand-binding residue (van Montfort *et al.*, 2001). In addition, for E2 regulatory transactivation domain comprising of α-helix and β-sheet domains, FIA also successfully identifies the experimentally confirmed ligand-binding residues of Tyr^{19} and Tyr^{178} (Abbate *et al.*, 2006). Hence *in silico* FIA prediction result is practically applicable in correlation with *in vitro* functional characterization of a variety of proteins.

4 DISCUSSION

The original motivation of this study was to locate conserved regions possessing hydrophilic aromatic residues and β-stranded structures in CBMs. Based on accurate alignment results from FIA, we further applied the idea to predict ligand-binding residues for CBM45 and CBM53 family proteins lacking resolved structures. Figure 3 demonstrates the strength of FIA and in that β-stranded structures and aromatic residues appeared to be well aligned. Low sequence identity makes it difficult to recognize the core regions in most cases, which implies that not every region in the query sequence is equally important for molecular recognition. Alternatively, if the key conserved regions can be well aligned or anchored, the alignment quality will not drop substantially. Indeed FIA successfully predicts four experimentally confirmed ligand-binding residues in *At*CBM45 and *At*CBM53. In addition, sequence alignment is a preliminary analysis tool and is the foundation of advanced research topics such as classification (CATH), secondary structure prediction (DSC) and structure modeling (Swiss-Model). As shown in Figure 4, Swiss-Model alone originally failed to simulate both of the *At*CBM45 and *At*CBM53 cases based on its automated procedures. Combination of FIA and Swiss-Model indeed led to successful generation of hypothetical structures, which in turn might facilitate further investigation of the functional roles of Trp^{340} and Phe^{389} in *At*CBM53. Moreover, FIA achieved PPV of 56.3%, while FRpred gave only 35.3%, from experimentally confirmed ligand-binding residues as indicated in Supplementary Table S3. The higher PPV derived from FIA is beneficial for biologists by substantial reduction of experimental efforts and expense in terms of finding crucial functional residues. Moreover, Table 1 lists the performance comparison of FIA and the other six outstanding alignment tools. FIA achieved the highest average sequence similarity and identity among all tested families. FIA significantly improved sequence similarity and sequence identity at a high confidence level by Friedman rank test. These improvements suggest that the incorporated biological features are informative for increasing alignment accuracy. Besides, an excellent alignment tool should insert fewer gaps to minimize alignment length. As shown in Table 1, with the exception of FIA, MAFFT outperforms the other five tools, by which five more averaged gaps were inserted as compared with FIA (refer to Supplementary Table S2). These findings suggest that FIA reports high alignment accuracy with minimized alignment length. A similar knowledge-incorporated approach was applied in constrained multiple sequence alignment (Tang *et al.*, 2003). The problem was that the constrained positions rely on experts' prior knowledge. In contrast, FIA is based on biological properties inside sequence content contained within CBMs. Nevertheless, the hydrophilic aromatic residues in other protein families including HSP20-like chaperones superfamily and E2 regulatory transactivation domain can be also identified and correlated with functional roles. Thus, FIA is easy to be generalized without specialist intervention. Furthermore, advanced tools usually rely on plain alignment, and the alignment quality directly affects the performance of future studies. The advanced tools require accurate alignment to increase their robustness. This requirement emphasizes the importance of alignment accuracy. In practice, FIA is based on the experimentally confirmed ligand-binding residues. More annotated sequences trend to increase the PPV. Furthermore, the template structures were preclassified by CATH. A more plentiful and accurate structure classification is believed to extend protein categories that FIA can predict. For implementation, FIA requires secondary structure prediction. When more accurate secondary structure prediction tool is available, the performance can be further improved.

ACKNOWLEDGEMENTS

We thank Chia-Han Chu, Tsan-Huang Shih and Liang-Cheng Chang for computational analysis suggestions and the verifications as well as Chien-Jung Chen and Sim-Kun Ng for biological data collection.

Funding: National Tsing Hua University (98N2903E1 to M.D.-T.C.); National Science Council (ROC NSC98-2622-B-007-001-CC1 to M.D.-T.C. and NSC98-2627-B-019-003 to T.-W.P.)

Conflict of Interest: none declared.

REFERENCES

Abbate,E.A. *et al.* (2006) Structure of the papillomavirus DNA-tethering complex E2:Brd4 and a peptide that ablates HPV chromosomal association. *Mol. Cell*, **24**, 877–889.

Abe,A. *et al.* (2004) Complex structures of Thermoactinomyces vulgaris R-47 alpha-amylase 1 with malto-oligosaccharides demonstrate the role of domain N acting as a starch-binding domain. *J. Mol. Biol.*, **335**, 811–822.

Bahr,A. *et al.* (2001) BAliBASE (Benchmark Alignment dataBASE): enhancements for repeats, transmembrane sequences and circular permutations. *Nucleic Acids Res.*, **29**, 323–326.

Benson,D.A. *et al.* (2009) GenBank. *Nucleic Acids Res.*, **37**, D26–D31.

Berman,H. *et al.* (2007) The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. *Nucleic Acids Res.*, **35**, D301–D303.

Boraston,A.B. *et al.* (2004) Carbohydrate-binding modules: fine-tuning polysaccharide recognition. *Biochem J.*, **382**, 769–781.

Boutet,E. *et al.* (2007) UniProtKB/Swiss-Prot. *Methods Mol. Biol.*, **406**, 89–112.

Capra,J.A. and Singh,M. (2007) Predicting functionally important residues from sequence conservation. *Bioinformatics*, **23**, 1875–1882.

Chen,X.W. and Jeong,J.C. (2009) Sequence-based prediction of protein interaction sites with an integrative method. *Bioinformatics*, **25**, 585–591.

Chou,W.Y. *et al.* (2006) Multiple indexing sequence alignment for group feature identification. *The 3rd Annual RECOMB Satellite Workshop on Regulatory Genomics*, 77–89.

Chou,W.-Y. *et al.* (2009) Biological feature incorporated alignment for cross species analysis on carbohydrate binding modules. *IEEE International Conference on Bioinformatics & Biomedicine* Washington, DC.

Christiansen,C. *et al.* (2009) The carbohydrate-binding module family 20–diversity, structure, and function. *FEBS J.*, **276**, 5006–5029.

Do,C.B. *et al.* (2005) ProbCons: probabilistic consistency-based multiple sequence alignment. *Genome Res.*, **15**, 330–340.

Edgar,R.C. (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*, **5**, 113.

Feese,M.D. *et al.* (2000) Crystal structure of glycosyltrehalose trehalohydrolase from the hyperthermophilic archaeum Sulfolobus solfataricus. *J. Mol. Biol.*, **301**, 451–464.

Fischer,J.D. *et al.* (2008) Prediction of protein functional residues from sequence by probability density estimation. *Bioinformatics*, **24**, 613–620.

Gotoh,O. (1982) An improved algorithm for matching biological sequences. *J. Mol. Biol.*, **162**, 705–708.

Guzman-Maldonado,H. and Paredes-Lopez,O. (1995) Amylolytic enzymes and products derived from starch: a review. *Crit. Rev. Food Sci. Nutr.*, **35**, 373–403.

Hashimoto,H. (2006) Recent structural studies of carbohydrate-binding modules. *Cell Mol. Life Sci.*, **63**, 2954–2967.

Henikoff,S. and Henikoff, J.G. (1992) Amino acid substitution matrices from protein blocks. *Proc. Natl Acad. Sci. USA*, **89**, 10915–10919.

Kamitori,S. *et al.* (1999) Crystal structure of Thermoactinomyces vulgaris R-47 alpha-amylase II (TVAII) hydrolyzing cyclodextrins and pullulan at 2.6 A resolution. *J. Mol. Biol.*, **287**, 907–921.

Katoh,K. *et al.* (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.*, **30**, 3059–3066.

King,R.D. and Sternberg,M.J. (1996) Identification and application of the concepts important for accurate and reliable protein secondary structure prediction. *Protein Sci.*, **5**, 2298–2310.

Larkin,M.A. *et al.* (2007) Clustal W and Clustal X version 2.0. *Bioinformatics*, **23**, 2947–2948.

Liu,Y.N. *et al.* (2007) Solution structure of family 21 carbohydrate-binding module from Rhizopus oryzae glucoamylase. *Biochem J.*, **403**, 21–30.

Mikkelsen,R. *et al.* (2006) A novel type carbohydrate-binding module identified in alpha-glucan, water dikinases is specific for regulated plastidial starch metabolism. *Biochemistry*, **45**, 4674–4682.

Milton,F. (1937) The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *J. Am. Stat. Assoc.*, **32**, 675–701.

Needleman,S.B. and Wunsch,C.D. (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.*, **48**, 443–453.

Notredame,C. *et al.* (2000) T-Coffee: a novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol.*, **302**, 205–217.

Orengo,C.A. *et al.* (1997) CATH–a hierarchic classification of protein domain structures. *Structure*, **5**, 1093–1108.

Pell,G. *et al.* (2003) Importance of hydrophobic and polar residues in ligand binding in the family 15 carbohydrate-binding module from Cellvibrio japonicus Xyn10C. *Biochemistry*, **42**, 9316–9323.

Ponyi,T. *et al.* (2000) Trp22, Trp24, and Tyr8 play a pivotal role in the binding of the family 10 cellulose-binding module from Pseudomonas xylanase A to insoluble ligands. *Biochemistry*, **39**, 985–991.

Schmidt,L.D. and Dauenhauer,P.J. (2007) Chemical engineering: hybrid routes to biofuels. *Nature*, **447**, 914–915.

Sorimachi,K. *et al.* (1997) Solution structure of the granular starch binding domain of Aspergillus niger glucoamylase bound to beta-cyclodextrin. *Structure*, **5**, 647–661.

Southall,S.M. *et al.* (1999) The starch-binding domain from glucoamylase disrupts the structure of starch. *FEBS Lett.*, **447**, 58–60.

Subramanian,A.R. *et al.* (2008) DIALIGN-TX: greedy and progressive approaches for segment-based multiple sequence alignment. *Algorithms Mol. Biol.*, **3**, 6.

Tang,C.Y. *et al.* (2003) Constrained multiple sequence alignment tool development and its application to RNase family alignment. *J. Bioinform. Comput. Biol.*, **1**, 267–287.

Thomsen,R. and Christensen,M.H. (2006) MolDock: a new technique for high-accuracy molecular docking. *J. Med. Chem.*, **49**, 3315–3321.

Tormo,J. *et al.* (1996) Crystal structure of a bacterial family-III cellulose-binding domain: a general mechanism for attachment to cellulose. *EMBO J.*, **15**, 5739–5751.

Tung,J.Y. *et al.* (2008) Crystal structures of the starch-binding domain from Rhizopus oryzae glucoamylase reveal a polysaccharide-binding path. *Biochem. J.*, **416**, 27–36.

Valdez,H.A. *et al.* (2008) Role of the N-terminal starch-binding domains in the kinetic properties of starch synthase III from Arabidopsis thaliana. *Biochemistry*, **47**, 3026–3032.

van Montfort,R.L. *et al.* (2001) Crystal structure and assembly of a eukaryotic small heat shock protein. *Nat. Struct. Biol.*, **8**, 1025–1030.

Wayllace,N.Z. *et al.* (2010) The starch-binding capacity of the noncatalytic SBD2 region and the interaction between the N- and C-terminal domains are involved in the modulation of the activity of starch synthase III from Arabidopsis thaliana. *FEBS J.*, **277**, 428–440.

Xie,H. *et al.* (2001) Role of hydrogen bonding in the interaction between a xylan binding module and xylan. *Biochemistry*, **40**, 5700–5707.

Yang,A.S. and Honig,B. (1999) Sequence to structure alignment in comparative modeling using PrISM. *Proteins*, (Suppl. 3), 66–72.

Yang,J.M. and Chen,C.C. (2004) GEMDOCK: a generic evolutionary method for molecular docking. *Proteins*, **55**, 288–304.