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

Wei-Yao Chou1,2,3, Wei-I Chou2,3,4, Tun-Wen Pai5,6, Shu-Chuan Lin2,3, Ting-Ying Jiang2,3, Chuan-Yi Tang1 and Margaret Dah-Tsyr Chang2,3,∗

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

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 (Gurman-Maldonado and Paredes-Lopez, 1995; Schmidt and Daenenhauer, 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
alignment implementation, the global optimal pairwise alignment can be solved in $O(n^3)$, 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 (Bah et al., 2001).

However, one common deficiency in sequence alignment is that the alignment accuracy relies heavily on sequence identity (Yang and Hug, 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; Poryxi 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 $\beta$-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 $\beta$-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.

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 $\beta$-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 Stemberg, 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, $\alpha$-helix, and $\beta$-sheet for primary sequences. When the reported probability of an amino acid in a $\beta$-sheet was higher than 60%, the amino acid would be annotated as ‘in $\beta$-strand’ in this study. In addition, the predicted $\beta$-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 $W_{\alpha}$, $F_{\alpha}$, and $Y_{\alpha}$, 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
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 \( R_s \) 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 90% 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.

Consistent rate(\( R_p, R_s \)) = \frac{\text{frequency of } R_p \text{ aligning to } R_s}{\text{frequency of } R_s \text{ in templates}} \times 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 >50% were filtered out and only five CBM structures including the starch-binding domains derived from *Rhizopus oryzae* glucosylase (Rs2v8mA02) (Tung et al., 2008), *Aspergillus niger* glucoamylase (An1ad0A01) (Sonirama et al., 1997), *Thermococcyscyes vulgaris α-amylase-II* (Tv1uh3A01) (Kamitori et al., 1999), and *Sulfobolus sajor-caju α-amylase-II* (Sl1eh9A03) (Fees 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 (AtCBM45) and starch synthase (AtCBM53) 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 ArchCBM catalyzes the transfer of the β-phosphate of adenosine triphosphate (ATP) in either C-3 or C-6 of the glucose residue for starch phosphorylation (Mikkelsen et al., 2006). The ArchCBM 53 contains a three repeated starch-binding domains at the N-terminal portion of starch synthase III (SSIII) involved in plant starch synthetase 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, 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,
In FIA. As shown in Figure 2, the square and the bold arrows values of the conserved properties in CBMs are specifically weighted identity, such positions become more difficult to align. To increase physiochemical properties. When two sequences share low sequence is to flexibly anchor the key conserved regions according to their et al., 2002; Subramanian, 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)].

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

![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.](image)

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 ArCBM45 and ArCBM53 were aligned pairwise against five template sequences to predict their conserved ligand-binding residues. Figure 3 demonstrates two FIA pairwise alignments between ArCBM45 and Arac000A00, as well as between ArCBM53 and Ro2v8mA00. 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. Trp139 and Trp194 in ArCBM45 and Trp340, Trp306, Phe383 and Tyr394 in ArCBM53 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, Trp139 and Trp340 in ArCBM45 (Mikkelsen et al., 2006) and Trp306 and Tyr394 in ArCBM53 (Wayllace et al., 2010) have been experimentally confirmed as true ligand-binding residues. Even the functional role of Trp340 and Phe383 in ArCBM53 have not been reported, these two aromatic residues are evidently located on the outer surface in an orientation similar to Tyr394 and Tyr31.

![Fig. 3. The FIA pairwise alignments of two domain sequences from CBM45 and CBM53 against two annotated structures. The predicted β-stranded elements in ArCBM45 and ArCBM53 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 I, O and Z, respectively. ‘−’ represents insertion or deletion gap in alignment.](image)
W.-Y. Chou et al.

in Ro28mAM0. 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.

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.

![Fig. 4. Structure modeling of AtCBM45 and AtCBM53. 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.](image)

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

<table>
<thead>
<tr>
<th>No. of seqs</th>
<th>FIA</th>
<th>MUSCLE</th>
<th>ClustalW2</th>
<th>DIALIGN-TX</th>
<th>T-COFFEE</th>
<th>ProbCons</th>
<th>MAFFT</th>
</tr>
</thead>
<tbody>
<tr>
<td>CBM4</td>
<td>38</td>
<td>45.3 (27.3)</td>
<td>37.2 (18.5)</td>
<td>33.1 (15.4)</td>
<td>28.8 (17.0)</td>
<td>34.6 (18.6)</td>
<td>33.9 (18.9)</td>
</tr>
<tr>
<td>CBM6</td>
<td>63</td>
<td>42.4 (25.9)</td>
<td>34.8 (17.5)</td>
<td>31.3 (14.6)</td>
<td>26.5 (15.8)</td>
<td>33.0 (17.7)</td>
<td>31.9 (18.1)</td>
</tr>
<tr>
<td>CBM9</td>
<td>36</td>
<td>50.5 (31.2)</td>
<td>39.9 (19.9)</td>
<td>36.8 (17.6)</td>
<td>29.4 (17.4)</td>
<td>37.1 (20.1)</td>
<td>35.9 (19.9)</td>
</tr>
<tr>
<td>CBM20</td>
<td>131</td>
<td>47.4 (28.9)</td>
<td>40.3 (21.7)</td>
<td>38.6 (20.0)</td>
<td>34.4 (21.2)</td>
<td>37.9 (21.5)</td>
<td>37.4 (21.8)</td>
</tr>
<tr>
<td>CBM21</td>
<td>66</td>
<td>45.7 (26.2)</td>
<td>38.4 (19.4)</td>
<td>35.0 (17.2)</td>
<td>32.4 (19.1)</td>
<td>36.0 (19.7)</td>
<td>35.4 (20.0)</td>
</tr>
<tr>
<td>CBM25</td>
<td>15</td>
<td>44.4 (27.7)</td>
<td>36.7 (19.6)</td>
<td>34.6 (16.7)</td>
<td>29.3 (18.3)</td>
<td>34.7 (20.0)</td>
<td>33.7 (19.1)</td>
</tr>
<tr>
<td>CBM26</td>
<td>5</td>
<td>48.4 (31.2)</td>
<td>37.0 (20.8)</td>
<td>37.5 (18.4)</td>
<td>32.6 (19.8)</td>
<td>36.1 (21.5)</td>
<td>34.9 (21.0)</td>
</tr>
<tr>
<td>CBM34</td>
<td>59</td>
<td>46.1 (27.4)</td>
<td>38.5 (20.1)</td>
<td>35.8 (18.0)</td>
<td>33.7 (19.8)</td>
<td>33.6 (20.1)</td>
<td>36.3 (20.5)</td>
</tr>
<tr>
<td>CBM41</td>
<td>65</td>
<td>42.7 (26.6)</td>
<td>34.5 (17.4)</td>
<td>31.9 (15.2)</td>
<td>27.4 (16.7)</td>
<td>32.4 (17.6)</td>
<td>31.0 (17.7)</td>
</tr>
<tr>
<td>CBM48</td>
<td>317</td>
<td>44.3 (26.6)</td>
<td>35.5 (17.6)</td>
<td>33.1 (15.5)</td>
<td>28.2 (17.0)</td>
<td>33.2 (17.9)</td>
<td>32.4 (18.1)</td>
</tr>
<tr>
<td>CBM53</td>
<td>13</td>
<td>44.3 (24.3)</td>
<td>36.5 (17.5)</td>
<td>33.9 (15.0)</td>
<td>29.2 (16.7)</td>
<td>34.2 (17.5)</td>
<td>34.4 (18.0)</td>
</tr>
<tr>
<td>Average</td>
<td>808</td>
<td>45.1 (27.2)</td>
<td>36.9 (18.8)</td>
<td>34.3 (16.6)</td>
<td>29.9 (18.0)</td>
<td>34.4 (19.0)</td>
<td>33.9 (19.2)</td>
</tr>
</tbody>
</table>
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 Trp85 in 1GME:A as a conserved hydrophilic aromatic residue. This Trp85 has been proven as a true ligand-binding residue (van Montfort et al., 2001). In addition, for β2 regulatory transactivation domain comprising of α-helix and β-sheet domains, FIA also successfully identifies the experimentally confirmed ligand-binding residues of Tyr25 and Tyr12 (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 AtCBM45 and AtCBM53. In addition, sequence alignment is a preliminary analysis tool and is the foundation of advanced research topics such as classification (CATH), secondary structure prediction tool is available, 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


