Functional Residue Prediction by Multiple Sequence Alignment for Carbohydrate Binding Modules

WI Chou[1,3], WY Chou[2], SC Lin[1], TY Jiang[1], CY Tang[2], Margaret DT Chang[1*]

[1]Institute of Molecular and Cellular Biology & Department of Life Science, [2]Department of Computer Science, National Tsing Hua University, Hsinchu, Taiwan 300, Republic of China. [3]Simpson Biotech Co., Ltd, Taoyuan County, Taiwan 333, Republic of China.
*lscmdt@life.nthu.edu.tw

Abstract

Multiple sequence alignment is often used to locate consensus sequence stretches with evolutionary and functional conservation. However, when sequence similarity among the queries becomes low, sequence alignment tools generate extremely diverse results. The aim of this study is to incorporate relevant biological knowledge and assumptions to improve quality of general alignment on low similarity sequences. Since recognition of key features in carbohydrate binding module (CBM) family does not apply to general models, a more accurate weighted entropy function employing secondary-structure-based and key-residue-weighted algorithms for alignment was designed to approach this goal. The results indicate that the proposed method is able to detect the known ligand-binding residues and to predict unknown functional residues in cellulose binding domains (CBDs) and xylooligosaccharides binding domains (XBDs) in terms of three-dimensional structures. Our results contribute molecular basis of CBDs and XBDs and potential application in development of alternative energy for future needs.

Keywords: multiple sequence alignment, carbohydrate binding domain.

1. Background

In the post genomic and proteomic era, more than 50,000 protein structures are solved and released in Protein Data Bank (PDB) (http://www.rcsb.org/). With such huge information in the database, acquisition and analysis of the data prior to biological experiments become important. Multiple sequence alignment reveals evolutionary conservation and correlations among a set of query sequences. Unfortunately, it has been proven to be an NP-hard problem, and is thus extremely time-consuming [1]. Most researchers believe that the polynomial time solution for an NP-hard problem may not exist and thus heuristic and approximate multiple sequence alignment have been designed and proposed [2].

Progressive and iterative alignments are two main strategies for multiple sequence alignment. For efficiency and simplicity, the former sacrifices a little accuracy [3], while the latter improves the alignment quality by using more computational cost through all the iterative processes [4]. For implementations, ClustalW [5] and T-Coffee [6] are two well-developed processes based on residue-to-residue observations. MISA [7] is a segment-to-segment comparison method based on identified local similar segments. DIALIGN [4] is another segment-to-segment comparison method based on neighboring residues. Constrained alignment strategies attempt to align assigned consecutive residue combinations [3]. However, most alignment tools assume that the query sequences possess a certain level similarity. When the similarity among the query sequences becomes lower than 30%, different sequence alignment tools generate

Proceedings of the 11th Joint Conference on Information Sciences (2008)
Published by Atlantis Press

extremely diverse results. In addition, complex biological mechanisms make it difficult to design a general mathematical model to solve all kinds of problems. For example, the CBM family is a set of proteins capable of binding to different carbohydrates or polysaccharides. Interestingly, even though the primary sequence similarities is extremely low, in general less than 20%, their secondary and tertiary structures, as well as the key functional residues are still well-conserved.

In this study, biological knowledge is incorporated to enhance the identification of characteristic motifs and to rule out the noise. For most CBM family members, the most important feature is that aromatic residues play a major role as key ligand-binding residues. In addition, aromatic residues are generally rather hydrophobic, however, an aromatic residue surrounded by polar residues can be exposed on the structural surface to achieve ligand-binding functions. Moreover, the β-sheet topologies of most CBMs are well conserved. Based on the aforementioned knowledge and observations, a weighted entropy measure was designed to construct multiple sequence alignment process in this study.

2. Results

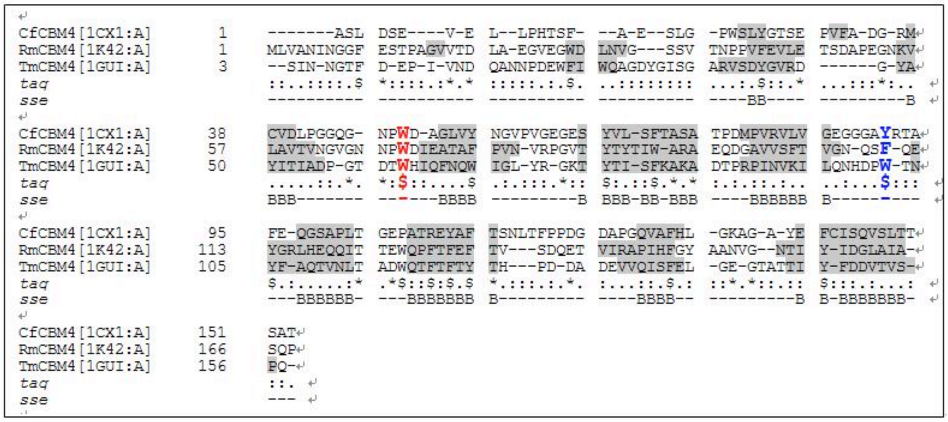
2.1. Datasets

CBMs have been classified into 52 families in ligand specificity, (for detailed classification, refer to http://afmb.cnrs-mrs.fr/CAZY/) [12]. CBM promotes the interaction between the substrate and glycan hydrolases (GHs), and increases local substrate concentration at the active sites of the catalytic domain. Currently 18 and 9 different CBM families have been identified to contain functional cellulose binding domains (CBDs) and xylooligosaccharides binding domains (XBDs), respectively. The three-dimensional structures of the resolved CBDs and XBDs consist of several β-strands forming two β-sheets. In order to verify the alignment by structure significance, only the proteins with resolved structures were selected. The primary sequences were fetched from PDB, and the functional domains were extracted by SCOP [9] and PDP [10]. The secondary structures were calculated by DSSP [11]. Only partial sequences of selected domains are used as the query sequences.

By CATH classification [12], two superfamilies classified from CBDs and XBDs were selected. The cluster of CATH code *2.60.120.260* containing both CBDs and XBDs was taken as the validation case since the ligand-binding residues in CBM4 were reported [13]. In the alignment of CBM4, the sequences are from *Cellulomonas fimi* (*Cf*CBM4), *Rhodothermus marinus* (*Rm*CBM4) and *Thermotoga maritima* (*Tm*CBM4). For the other case, we attempted to predict functional ligand-binding residues from the cluster of CATH code *2.80.10.50* whose sequences are rather dissimilar. In the alignment of this dataset, the sequences are from *Clostridium botulinum* (*Cb*CBM13), *Streptomyces lividans* (*Sl*CBM13), *Abrus precatorius* (*Ap*CBM13), *Cucumaria echinata* (*Ce*CBM13) and *Ricinus communis* (*Rc*CBM13). In order to compare the proposed method with existing tools, ClustalW and DIALIGN were performed.

2.2. Known Functional Residues Validations

Based on biological observations in CBMs, we hypothesized that strong correlation could be obtained not only from relative β-stranded structures, but also from key ligand-binding residues on loop regions. As shown in Fig. 1, the secondary structural elements marked by grey boxes were well-aligned in terms of relative positions and lengths, and key aromatic residues were found to be conserved spotted by *'$'*. Interestingly, the reported two key aromatic residues were successfully aligned and highlighted in red and blue respectively. To demonstrate the significance at structure level, the corresponding functional residues are also highlighted in red and blue in Fig. 2. It is clear that these three structures contain common cavities in the upper part of the structures and the functional residues are located on surface marked in red and blue. Those results confirm that our proposed method is capable of capturing the known key functional residues.



```
CfCBM4[1CX1:A]     1    -------ASL DSE----V-E L--LPHTSF- --A-E--SLG -PWSLYGTSE PVFA-DG-RM
RmCBM4[1K42:A]     1    MLVANINGGF ESTPAGVVTD LA-EGVEGWD LNVG---SSV TNPPVFEVLE TSDAPEGNKV
TmCBM4[1GUI:A]     3    --SIN-NGTF D-EP-I-VND QANNPDEWFI WQAGDYGISG ARVSDYGVRD ------G-YA
taq                     ::..::::.$ *::::.:*.* :::::.:.$. ..:::::::: ...:.$::.* ...:::*:..
sse                     ---------- ---------- ---------- ---------- ----BB---- ---------B

CfCBM4[1CX1:A]    38    CVDLPGGQG- NPWD-AGLVY NGVPVGEGES YVL-SFTASA TPDMPVRVLV GEGGGAYRTA
RmCBM4[1K42:A]    57    LAVTVNGVGN NPWDIEATAF PVN-VRPGVT YTYTIW-ARA EQDGAVVSFT VGN-QSF-QE
TmCBM4[1GUI:A]    50    YITIADP-GT DTWHIQFNQW IGL-YR-GKT YTI-SFKAKA DTPRPINVKI LQNHDPW-TN
taq                     ......::.*. *:$::....$ .:.:::.*:: $:.::$.*.* :.:.::.:.. ..:...$:::
sse                     BBB------- ---------BBBB ---------B BBB-BB-BBB ----BBBBBB B---------

CfCBM4[1CX1:A]    95    FE-QGSAPLT GEPATREYAF TSNLTFPPDG DAPGQVAFHL -GKAG-A-YE FCISQVSLTT
RmCBM4[1K42:A]   113    YGRLHEQQIT TEWQPFTFEF TV---SDQET VIRAPIHFGY AANVG--NTI Y-IDGLAIA-
TmCBM4[1GUI:A]   105    YF-AQTVNLT ADWQTFTFTY TH---PD-DA DEVVQISFEL -GE-GTATTI Y-FDDVTVS-
taq                     $.:......:* .*$::$:$.$ *.:::.:.*. :...::.$.: ::*.*::.:: $::.:...:
sse                     ---BBBBBB- ---BBBBBBB B--------- ----BBBB-- ---------B B-BBBBBBB-

CfCBM4[1CX1:A]   151    SAT
RmCBM4[1K42:A]   166    SQP
TmCBM4[1GUI:A]   156    PQ-
taq                     ::.
sse                     ---
```

Fig. 1: Multiple sequence alignment of CATH code *2.60.120.260* cluster (CBM4 only): The β-sheets are highlighted in gray boxes. The aromatic residues highlighted in red and blue are corresponding to functional residues on surface. Extra *tag* and *sse* sequences denote the quality of the alignment regarding to residue and β-sheet conservations respectively. The *'$'* in a column in the extra *tag* sequence represents the conservations of aromatic residues.

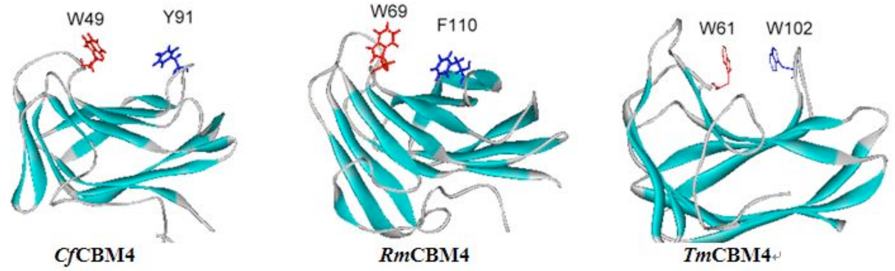
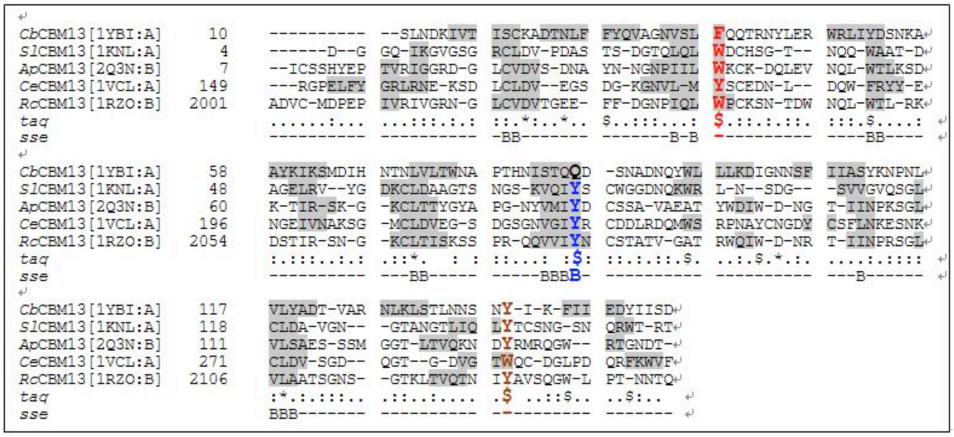


Fig. 2: Structure features of CATH code *2.60.120.260* cluster: Wherever possible, structures are oriented similarly. The aromatic residues predicted by structure-based sequence alignment to involve in ligand-binding are shown in sticks in red and blue. *Cf*CBM4 is from *Cellulomonas fimi*; PDB code: 1CX1. *Rm*CBM4 is from *Rhodothermus marinus*; PDB code: 1K42. *Tm*CBM4 is from *Thermotoga maritima*; PDB code: 1GUI.

2.3. Binding Residue Prediction

In CATH code *2.80.10.50* cluster, the conserved binding residues are unknown. Mul tiple sequence alignment of the five structures classified into CATH code *2.80.10.50* cluster is presented in Fig. 3. As the same representation in Fig. 1, the secondary structural elements were well-aligned in terms of relative positions and lengths, and key aromatic residues were found to be conserved. In particular, the well-aligned aromatic residues, when exposed on a surface, were further labeled in red, blue and brown.



```
CbCBM13[1YBI:A]     10    ----------  --SLNDKIVT  ISCKADTNLF  FYQVAGNVSL  FQQTRNYLER  WRLIYDSNKA
SlCBM13[1KNL:A]      4    ------D--G  GQ-IKGVGSG  RCLDV-PDAS  TS-DGTQLQL  WDCHSG-T--  NQQ-WAAT-D
ApCBM13[2Q3N:B]      7    --ICSSHYEP  TVRIGGRD-G  LCVDVS-DNA  YN-NGNPIIL  WKCK-DQLEV  NQL-WTLKSD
CeCBM13[1VCL:A]    149    ---RGPELFY  GRLRNE-KSD  LCLDV--EGS  DG-KGNVL-M  YSCEDN-L--  DQW-FRYY-E
RcCBM13[1RZO:B]   2001    ADVC-MDPEP  IVRIVGRN-G  LCVDVTGEE-  FF-DGNPIQL  WPCKSN-TDW  NQL-WTL-RK
tag                       ......:...  ...:::.:.:  ::.*:..*..  $..::::...:  $.::.:.::.  :::.$....:
sse                       ----------  ----------  -BB-------  -------B-B  ----------  ----BB----

CbCBM13[1YBI:A]     58    AYKIKSMDIH  NTNLVLTWNA  PTHNISTQQD  -SNADNQYWL  LLKDIGNNSF  IIASYKNPNL
SlCBM13[1KNL:A]     48    AGELRV--YG  DKCLDAAGTS  NGS-KVQIYS  CWGGDNQKWR  L-N--SDG--  -SVVGVQSGL
ApCBM13[2Q3N:B]     60    K-TIR-SK-G  -KCLTTYGYA  PG-NYVMIYD  CSSA-VAEAT  YWDIW-D-NG  T-IINPKSGL
CeCBM13[1VCL:A]    196    NGEIVNAKSG  -MCLDVEG-S  DGSGNVGIYR  CDDLRDQMWS  RPNAYCNGDY  CSFLNKESNK
RcCBM13[1RZO:B]   2054    DSTIR-SN-G  -KCLTISKSS  PR-QQVVIYN  CSTATV-GAT  RWQIW-D-NR  T-IINPRSGL
tag                       :.:::..:.:  .::*.  : :  ::...: :$:  ::.:.::.$.  ..:.$.*.:.  ....:.::::
sse                       ----------  ---BB-----  -----BBBB-  ----------  ----------  ---B------

CbCBM13[1YBI:A]    117    VLYADT-VAR  NLKLSTLNNS  NY-I-K-FII  EDYIISD
SlCBM13[1KNL:A]    118    CLDA-VGN--  -GTANGTLIQ  LYTCSNG-SN  QRWT-RT
ApCBM13[2Q3N:B]    111    VLSAES-SSM  GGT-LTVQKN  DYRMRQGW--  RTGNDT-
CeCBM13[1VCL:A]    271    CLDV-SGD--  QGT--G-DVG  TWQC-DGLPD  QRFKWVF
RcCBM13[1RZO:B]   2106    VLAATSGNS-  -GTKLTVQTN  IYAVSQGW-L  PT-NNTQ
tag                       :*.:.:::..  .::..:....  .$ ..::$..  ..$:..
sse                       BBB-------  ----------  ----------  -------
```

Fig. 3: Multiple sequence alignment of CATH code *2.80.10.50* cluster: The β-sheets are highlighted in gray boxes. The aromatic residues highlighted in red, blue and brown are corresponding to the aromatic residues on surface. Extra *tag* and *sse* sequences denote the quality of the alignment regarding to residue and β-sheet conservations respectively. The *'$'* in a column in the extra *tag* sequence represents the conservations of aromatic residues.

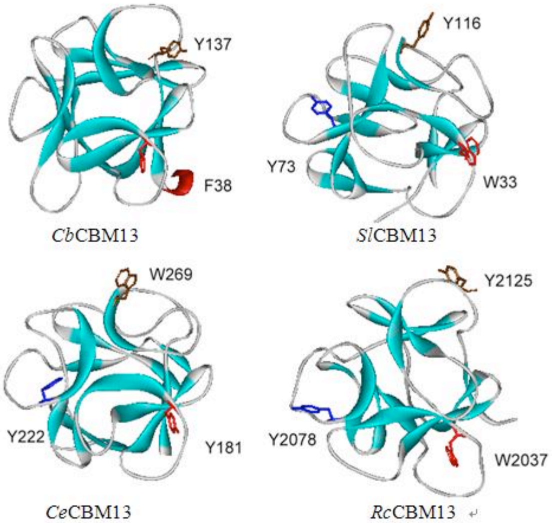


Fig. 4: Structure features of CATH code 2.80.10.50 cluster: Wherever possible, structures are oriented similarly. The aromatic and polar residues predicted by structure-based sequence alignment to be involved in ligand-binding are shown as sticks. CbCBM13 is from *Clostridium botulinum*; PDB code: 1YBI. SlCBM13 is from *Streptomyces lividans*; PDB code: 1KNL. ApCBM13 is from *Abrus precatorius*; PDB code: 2Q3N. CeCBM13 is from *Cucumaria echinata*; PDB code: 1VCL. RcCBM13 is from *Ricinus communis*; PDB code: 1RZO.

The corresponding residues were also highlighted in the structures using the same color annotations as displayed in Fig. 4, in which all five structures had similar topology. The highlighted residues were all located at corresponding positions. Interestingly, we found that even though *Cb*CBM13 shared similar β-sheet and loop structures on the left hand of the structures, it lacks of the aromatic residue highlighted in blue in other structures. These results suggest that the three aligned aromatic residues on surface are more likely to have ligand-binding functions.

In order to demonstrate the strength of our proposed methodologies, two well-established sequence alignment tools named ClustalW and DIALIGN were also performed on the same dataset. The alignment result of ClustalW revealed that one aromatic residue was misaligned in *Cb*CBM13 and some other residues were poorly aligned (data not shown). On the other hand, the alignment result of DIALIGN revealed that too many gaps were inserted such that there were many giant shifts. The three putative functional aromatic residues were scattered without any alignment pattern in DIALIGN (data not shown). The alignment qualities of the proposed method, ClustalW and DIALIGN were also compared on entropy measure on CATH code *2.80.10.50* cluster. Since β-stranded structures are not available in ClustalW and DIALIGN, without bias, the E_n measure defined in Equation 1 was taken as the criterion. Table 1 presents the sum of E_n values in columns of the alignments. The proposed manner possessed the lowest sum of E_n at 152.01, while DIALIGN produced almost two times of that. These data indicate that the proposed method outperformed ClustalW and DIALIGN in terms of both structural significances and E_n criterion.

	The proposed method	ClustalW	DIALIGN
sum of E_w values	**152.01**	155.33	302.31

Table 1: Comparisons of sum of Ew values: The values are from the proposed method, ClustalW and DIALIGN on CATH code 2.80.10.50 cluster.

3. Conclusions

In viewpoint of computation, we can successfully align key functional ligand-binding residues in CBDs and XBDs very well, but the particular biological knowledge is required in advance. There are more protein families which possess similar structure properties, but the sequence similarity may be low. It is worthy to apply our method to cope with different kinds of functionally related low similarity sequences. On the other hand, in viewpoint of biology, carbohydrates are important materials for the biofuel, food and beverage applications. Detailed knowledge of both carbohydrate manipulation, *via* studies of catalytic domains of carbohydrate-active enzymes, and carbohydrate binding, *via* studies of CBMs, is vital to an overall understanding of carbohydrate-active enzymes. This study therefore contributes molecular basis for further engineering of novel carbohydrate-active enzymes to meet future needs.

4. Methods

$$E_n(a,b) = E(a,b) - p_{gap}*[(1-p_{gap})*\ln(p_{gap}) + p_{gap}*\ln(1/n)] \quad (1)$$

In Equation 1, p_{gap} stands for the occurrence probability of gaps in a column and n denotes the total number of characters in a putative aligned column. The *1/n* represents that all gaps are treated as different characters. When p_{gap} increases, the weighting coefficient of ln(p_{gap}) decrease, while the coefficient of ln(*1/n*) increases. Moreover, the major features of the proposed system focus on the incorporation of knowledge-based scoring function. In this scenario, entropy values were weighted or penalized based on appearance of particular biological characteristics. The weighted entropy measurement of two aligned residues is defined as follows:

$$E_w(a,b) = E_n(a,b) + \beta\_bonus + aromatic\_bonus + polar\_surrounded\_bonus \quad (2)$$

In Equation 2, the β_bonus is earned if both a and b are located in their β-sheets, respectively. Otherwise, the β_bonus is set to be invalid. The aromatic_bonus is added if the occurrence probability of aromatic residues in a and b is greater than a threshold. Otherwise, the aromatic_bonus is set to be invalid. The polar_surrounded_bouns is calculated by the number of particular polar residues surrounding a and b. In the case that aromatic_bonus is invalid, the polar_surrounded_bouns is set to be zero directly. In particular, if a and '-' are aligned in a column, the requirements of the above three bonuses are never satisfied. However, based on the biological observations in CBMs, β-sheet structures are rather conserved. Gaps in β-sheets mean relevant diversities. Therefore, β_gap_penalty is turned to account if the inserted gap is in a β-sheet, and Equation 3 displays the formal formula.

$$E_w(a,'-') = E_n + \beta\_gap\_penalty \quad (3)$$

5. Acknowledgements

The authors thank Mr. Timothy Hwang for proofreading. This work is sponsored by National Science Council, R. O. C. grants NSC 97-2622-B-007-001-, 97-2627-B-007-017- and Congress of Agriculture, R. O. C. 97AS-1.2.1.-ST-a5 to Margaret Dah-Tsyr Chang.

6. References

[1] Bonizzoni P, Vedova GD, "The complexity of multiple sequence alignment with SP-score that is a metric," *Theoretical Computer Science* 2001, 259:63-79.

[2] Thompson JD, Plewniak F, Poch O, " A comprehensive comparison of multiple sequence alignment programs," *Nucleic Acids Res* 1999, 27(13):2682-2690.

[3] Tang CY, Lu CL, Chang MD, Tsai YT, Sun YJ, Chao KM, Chang JM, Chiou YH, Wu CM, Chang HT *et al*, " Constrained multiple sequence alignment tool development and its application to RNase family alignment," *J Bioinform Comput Biol* 2003, 1(2):267-287.

[4] Subramanian AR, Weyer-Menkhoff J, Kaufmann M, Morgenstern B, "DIALIGN-T: An improved algorithm for segment-based multiple sequence alignment**,**" *BMC Bioinformatics* 2005, 6:66.

[5] Thompson JD, Higgins DG, Gibson TJ, "CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice," *Nucleic Acids Res* 1994, 22(22):4673-4680.

[6] Notredame C, Higgins DG, Heringa J, " T-Coffee: A novel method for fast and accurate multiple sequence alignment," *J Mol Biol* 2000, 302(1):205-217.

[7] Chou WY, Pai TW, Lai ZC, Tzou WS, " Multiple indexing sequence alignment for group feature identification," *The 3rd Annual RECOMB Satellite Workshop on Regulatory Genomics* 2006:77-89.

[8] Boraston AB, Bolam DN, Gilbert HJ, Davies GJ, "Carbohydrate-binding modules: fine-tuning polysaccharide recognition," *Biochem J* 2004, 382(Pt 3):769-781.

[9] Murzin AG, Brenner SE, Hubbard T, Chothia C, "SCOP: a structural classification of proteins database for the investigation of sequences and structures," *J Mol Biol* 1995, 247(4):536-540.

[10] Alexandrov N, Shindyalov I, " PDP: protein domain parser**,**" *Bioinformatics* 2003, 19(3):429-430.

[11] Kabsch W, Sander C, "Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features," *Biopolymers* 1983, 22(12):2577-2637.

[12] Greene LH, Lewis TE, Addou S, Cuff A, Dallman T, Dibley M, Redfern O, Pearl F, Nambudiry R, Reid A *et al*, "The CATH domain structure database: new protocols and classification levels give a more comprehensive resource for exploring evolution," *Nucleic Acids Res* 2007, 35(Database issue):D291-297.

[13] Simpson PJ, Jamieson SJ, Abou-Hachem M, Karlsson EN, Gilbert HJ, Holst O, Williamson MP, "The solution structure of the CBM4-2 carbohydrate binding module from a thermostable Rhodothermus marinus xylanase," *Biochemistry* 2002, 41(18):5712-5719.

[14] Shannon CE, "The mathematical theory of communication," 1963. *MD Comput* 1997, 14(4):306-317.

[15] Liao H, Yeh W, Chiang D, Jernigan RL, Lustig B, "Protein sequence entropy is closely related to packing density and hydrophobicity," *Protein Eng Des Sel* 2005, 18(2):59-64.