Available online at www.sciencedirect.com
SCIENCE DIRECT®
Biochemical and Biophysical Research Communications 322 (2004) 277–280
BBRC
www.elsevier.com/locate/ybbrc

Principle component analysis in F/10 and G/11 xylanase☆

Liangwei Liu*, Jue Zhang, Bin Chen, Weilan Shao

The Key Laboratory of Industrial Biotechnology, Ministry of Education, Southern Yangtze University[1], 170 Huihe Road, Wuxi 214036, Jiangsu, PR China

Received 6 July 2004

Abstract

A bioinformatics method was used to analyze F/10 and G/11 xylanase basing on principle component analysis, and a model was made to classify between these two folds with an ideal result. The principle components were predicated to be secondary structures, the components were analyzed with the architecture of each family, and found comparable with $(\beta/\alpha)_8$-barrel of F/10 xylanase and right-hand structure of G/11 xylanase. Compared with sequence similarities, this method gave discriminating features a clear meaning. The largest component did not appear in the model, which revealed no difference between these two families.

Keywords: Classification; Computation; Principle component analysis; Structure; Xylanase

The importance of xylanase (EC 3.2.1.8) lies in the recycling of biomass and its wide usage in biotechnology, such as in pulp bleaching, baking industry, and manufacturing of animal feed [1–3]. More and more attention was paid to it, and many new xylanases were cloned and expressed. There were 67 xylanases in Swiss-Prot (Release 43.5 of 07-Jun-2004) and 447 xylanases in Tremble, certainly there were many fragments in Tremble, which were known as redundancy. This is the difference between Swiss-Prot and Tremble, because Swiss-Prot was checked by experts, whereas Tremble was not checked.

According to molecular size and pH of xylanase, Wong first classified xylanase into two families [4], which showed that one family had high molecular weight and low p*I* value and the other xylanase had low molecular weight and high p*I* value. In 1989, using amino acid sequence similarities and hydrophobic cluster analysis, Henrissat [5] classified cellulases and xylanases into six families (A–F); subsequently, the families were upgraded to 11 [6,7]; xylanases were then subdivided into F and G families, which were analogous to glycohydrolase families 10 and 11[8], comprising high and low molecular weight xylanase, respectively. After that, this method was widely used as a classification system for glycosyl hydrolases to complement the I.U.B [9], many years have passed and left it unchanged. Basing on principle component analysis, the amino acid contents were computed and used as factor to discriminate F/10 xylanase from G/11 xylanase, the result was very encouraging; the related principle components were predicted as secondary structures and the real difference of amino acid components between these two families of xylanase was found.

Materials and methods

Data set construction. As mentioned above, xylanases were downloaded from Swiss-Prot (http://au.expasy.org) (Release 43.5 of 07-Jun-2004), for the consideration of its non-redundancy. Its accuracy is protected for having been checked by experts, this is an important factor in analyzing sequence. There were 67 xylanases, 30 belong to F/10 family

☆ *Abbreviations:* PDB, Protein Data Bank; SAS, statistic analysis system; FORTRAN: formula translation; p_r, probability; P1, P2, P3, P4, P5, P6, P7: the upper 7 principle components. The single and three alphabet codes of amino acids used in the article conform to the common rules of the IUPAC-IUB Commission on Biochemical Nomenclature.

* Corresponding author. Fax: +86 25 83598838.
E-mail address: llw321@yahoo.com.cn (L. Liu).

[1] Former name was known as: Wuxi light industry university.

0006-291X/$ - see front matter © 2004 Elsevier Inc. All rights reserved.
doi:10.1016/j.bbrc.2004.07.116

and 31 belong to G/11 family, 2 fragments (P80717, P80718), 2 belong to 43 family (P48791,P45796); 1 belongs to 62 family (P23031); and 1 belongs to 16 family (Q53317). We each selected 25 xylanases with relatively the same length from the F/10 and G/11 xylanase, respectively, as base data set to make discriminating model basing on principle component analysis:

F/10 xylanase (P29417, P09850, Q00177, P26514, P45703, P33559, P40943, P56588, P07528, P23556, P23360, P40942, Q12603, Q60041, O59859, P23557, P49942, P23551, P07529, P48789, O60206, P07986, P26223, P23030, and P51584).

G/11 xylanase (P55331, P36218, P17137, Q06562, P36217, P18429, P00694, P48793, P33557, P55328, P55333, P26220,P45705, P09850, P55332, P55334, P55329, P35809, O43097, Q06562, P55330, P55335, P83513,P81536, and P29127).

In order to check the accuracy of the model, the xylanases from Swiss-Prot and Tremble were selected as checking data set. In this set, only those xylanases with full length were used, although those not annotated as fragment may also be selected. There were 317 xylanases in the set, according to the annotation of Swiss-Prot and Tremble, 110 belonged to G/11 xylanase, and 207 belonged to F/10 xylanase.

Calculating the contents of amino acids and principle components. Program was compiled by Compaq visual Fortran (version 6.5) to calculate the contents of 20 amino acids in each xylanase sequence, and the contents of residues were used as variant to calculate principle components by the principle component analysis procedure of SAS (version 8.1). In this way, the principle components of amino acid in each family can be determined.

Finding discrimination function. Using the above principle components to regress with the two families by the stepwise procedure of SAS (version 8.1), an optimal function can be got to discriminate between these two xylanases.

Checking the function. Using the checking data set to check the validity of the function, and the calculation result can be checked with the annotations in the Swiss-Prot and Tremble to find out if the calculation was right.

Results and discussion

Principle components

After principle component analysis of xylanase in each family, the upper 7 components were found as the following in Table 1 (cumulative proportion of variance >90%, coefficient of each variant was retained with one decimal place accuracy, and the coefficient below 0.2 was omitted for simplicity of the function).

Discrimination function

F/10 was assigned a value of 1, and G/11 was assigned a value of 5, using the stepwise procedure of SAS, a discriminating function was got as the following (coefficient of each variant was retained with two decimal place accuracy, each coefficient reached significant $p_r < 0.15$, and the model reached significant $p_r < 0.001$, $R^2 = 0.92$): $F = 0.15 * P2 - 0.13 * P3 + 0.1 * P4 - 0.1 * P6 + 2.02$.

To discriminate F/10 xylanase from G/11 xylanase, we set the value as 2.12 according to the calculation value in the base data set of the known F/10 xylanase value, if the F value below 2.12 was considered as F/10 xylanase, and if the F value above 2.12 was considered as G/11 xylanase. The square regression coefficient of 0.92 revealed a significant relationship between principle components and the difference of the two families.

The validity of the function

Using the above function, we check its accuracy by calculating xylanase in the checking data set collected from Swiss-Prot and Tremble, among the 231 F/10 xylanases (quoted from the annotations from Swiss-Prot and Tremble Release 43.5 of 07-Jun-2004), 11 were calculated and predicted as the G/11 xylanase, the correct rate was 95.2%. Among the 135 G/11 xylanases, there was only 1 calculated and predicted as F/10 xylanase, the correct rate was 99.3%; the incorrect result is shown in Table 2. From the table, one can see that these xylanases all have some problems, such as having CBD

Table 1
Meaning of principle components in xylanase

	Proportion (%)	Amino acids (positive)	Amino acids (negative)	Secondary structure
F/10				
P1	55	0.5S 0.4A 0.2G 0.2T 0.2Q	0.4E 0.4K 0.2L	Coil
P2	11	0.4R 0.4E 0.3S	0.6K 0.4A	Helix
P3	8	0.6S 0.2I 0.2K 0.2Y	0.4Q 0.4T	Strand
P4	7	0.6N 0.3Y	0.5A 0.3D 0.3L 0.2S 0.2E	Turn
P5	4	0.6I 0.3Q 0.3V 0.2H	0.4C 0.3L 0.3T 0.2F 0.2Y	Strand
P6	3	0.5Q 0.3D 0.3L	0.4T 0.3P 0.2A 0.2E 0.2F	Helix
P7	3	0.5P 0.3Q 0.2M 0.2S 0.2V	0.5L 0.3I 0.3T 0.3N	Turn
G/11				
P1	39	0.7S 0.3A 0.2V	0.3G 0.3K 0.2N 0.2Q 0.2R	Coil
P2	14	0.5G 0.3N 0.3V 0.2Q 0.2S	0.5K 0.4D 0.2C 0.2E 0.2F 0.2I	Turn
P3	13	0.4Q 0.3F 0.2E 0.2N	0.6T 0.4G 0.2D 0.2W	Strand
P4	10	0.5N 0.4K 0.3S 0.2I 0.2T	0.3D 0.3Y 0.2E 0.2G	Turn
P5	9	0.4G 0.4S 0.2E 0.2I	0.6V 0.3D 0.2W	Turn
P6	4	0.5A 0.2L 0.2N	0.6Q 0.2T 0.2D	Helix
P7	3	0.4L 0.3D 0.3I 0.3V 0.2P	0.4M 0.2A 0.2F 0.2R 0.2T 0.4Y	Strand

Table 2
Result of calculation

Accession No.	Annotation	Calculation family	Comment
Q8WZJ4	7	G/11	
Q9L8L8	F/10	G/11	
P96988	F/10 (fragment)	G/11	CBD-3
Q8CK14	43	G/11	
Q8RVD6	—	G/11	Hypothetical
Q8TVR8	Xylanase/deacetylase	G/11	Predicted
Q93AQ5	Putative (fragment)	G/11	CBD-6
Q7UM13	F/10	G/11	Probable
Q7UP58	Xylanase/deacetylase	G/11	Probable
Q7UU02	Xylanase/deacetylase	G/11	Probable
Q7WTN6	F/10	G/11	CBD-4-9-2
Q7M836	Xylanase/deacetylase	F/10	

Accession No., the no assigned by Swiss-Prot or Tremble; annotation, annotation quoted from Swiss-Prot or Tremble; calculation family, assigned by present work according to calculation value of principle; and comment, quoted from Swiss-Prot or Tremble.

(P96988, Q93AQ5, and Q7WTN6) or being fragments (P96988, Q93AQ5), or being annotated as probable xylanases (Q8RVD6, Q8TVR8, Q7UM13, Q7UP58, and Q7UU02), or non-specific xylanase (Q8TVR8, Q7UP58, Q7UU02, anf Q7M836), for they were also annotated as chitin deacetylase. To Q7M836, because it was only annotated as xylanase/chitin deacetylase without being annotated as G/11 xylanase or F/10 xylanase, it can be proposed as a member of F/10 xylanase. From this result we can see that the classification of xylanase is not an absolute standard as being black or white, for there were some xylanases which could be the intermediates. As was indicated xylanase II from the alkaliphilic thermophilic *Bacillus* having a distinctly different structure from other xylanases, for the polypeptide, was predicted to be formed primarily by β-strands; but the sequence homology revealed similar identity with families F/10 and G/11 of xylanases [10]; because there was often domain shuffling and gene transferring among F/10 xylanase [11,12], the result revealed that F/10 family had more xylanase deviated from the common amino acid composition than G/11 xylanase. G/11 xylanase mostly had one catalysis domain, so the result showed that only one xylanase deviated from the common value.

Meaning of the principle components

According to the protein secondary structure prediction method of Chou and Fasman [13], we assigned the meaning of the principle components as in Table 1. From Table 1, it can be seen that the largest proportion of the component P1 made no difference between these two families, because they all belong to coils of secondary structure in proteins, this kind of structure makes no characteristic features to discriminate itself from other folds. It is also clear that coils comprise most of the structure of a protein, whether it is in F/10 or G/11 xylanase, because P1 had 55% in F/10; and 39% in G/11 xylanase. The architecture of F/10 xylanase was known to be made of $(\beta/\alpha)_8$ barrel [14]. From the result shown in Table 1, principle components of P2, P3, P5, and P6 were seen as helixes and strands, coinciding with the $(\beta/\alpha)_8$ structure of F/10 xylanase. Well, turns (P4 and P7) were also found to be among the larger proportions, because many regions needed to connect helixes and

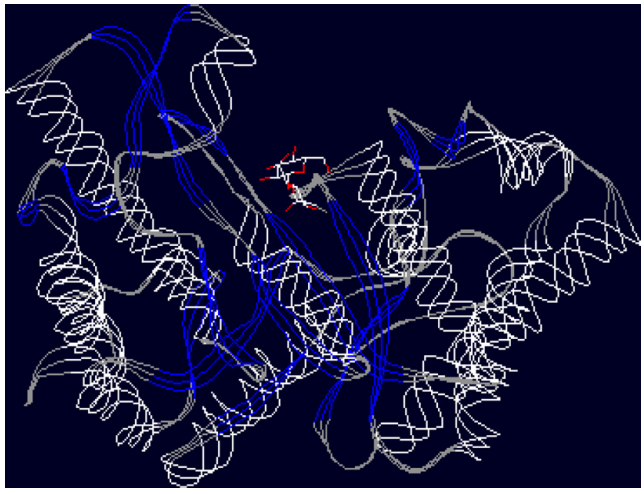

Fig. 1. Structure of F/10 xylanase (produced by Swiss-pdbViewer of accession number of 1EXP1 in PDB).

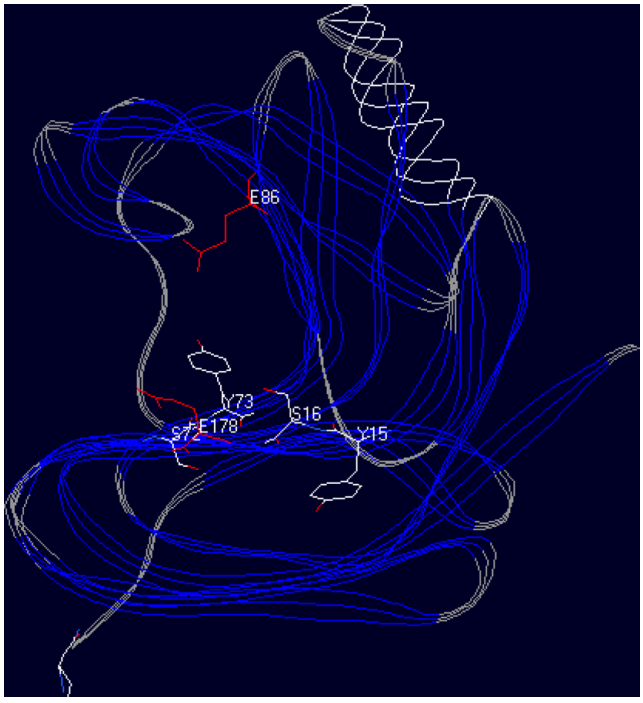

Fig. 2. Structure of G/11 xylanase (produced by Swiss-pdbViewer of accession number of 1YNA in PDB).

strands; and it seems that P4 and P7 belong to different turns, for they use different amino acid compositions. This architecture constructs 10% of the natural proteins [15], and it was also found to be the most common protein folding pattern [14], so computation of the principle components in the structure is important. A direct view of the $(\beta/\alpha)_8$ barrel structure of F/10 xylanase can be seen in Fig. 1.

As to the architecture of G/11 xylanase, it was reported resembling a right-hand structure [16], composed of predominantly β-strands and only one α-helix [17]. From the result in Table 1, it can be seen that principle components P3 and P7 were β-strands, comparable to the predominantly β-strands; the component P6 was assigned as α-helix, coinciding with the α-helix of the structure. Well, there was also large amount of turns in this structure, as it can be seen from components P2, P4, and P5; from the direct view of the structure in Fig. 2, we can understand this architecture also needs many turns to connect each of the β-strands and with the only α-helix.

Conclusion

Computational method was used to analyze the principle components in F/10 xylanase and G/11 xylanase, the components were assigned to the secondary structures of protein, and compared with the structure of each family, providing a new thought for analyzing the fold difference of protein. It is useful to discriminate between different folds of protein.

References

[1] L. Viikari, A. Kantellinen, J. Sundquist, M. Linko, Xylanases in bleaching: from an idea to the industry, FEMS Microbiol. Rev. 13 (1994) 335–350.
[2] R.A. Prade, Xylanases: from biology to biotechnology, Biotechnol. Genet. Eng. Rev. 13 (1996) 101–131.
[3] K. Poutanen, Enzymes: an important tool in the improvement of the quality of cereal foods, Trends Food Sci. Technol. 9 (1997) 300–306.
[4] K.K. Wong, L.U. Tan, J. Saddler, Multiplicity of β-1,4-xylanase in microorganisms: functions and applications, Microbiol. Mol. Biol. Rev 52 (1988) 305–317.
[5] B. Henrissat, M. Claeyssens, P. Tomme, L. Lemesle, J.P. Mornon, Cellulase families revealed by hydrophobic cluster analysis, Gene 81 (1989) 83–95.
[6] N.R. Gilkes, D.G. Kilburn, R.C. Miller Jr., J. Sundquist, Domains in microbial B-1,4-glycanases: sequence conservation, function, and enzyme families, Microbiol. Mol. Biol. Rev. 55 (1991) 303–315.
[7] B. Henrissat, A classification of glycosyl hydrolases based on amino acid sequence similarities, Biochem. J. 280 (1991) 309–316.
[8] B. Henrissat, A. Bairoch, New families in the classification of glycosyl hydrolases based on amino acid sequence similarities, Biochem. J. 293 (1993) 781–788.
[9] Anonymous, Enzyme Nomenclature: Recommendations of the Nomenclature Committee of the International Union of Biochemistry on the Nomenclature and Classification of Enzyme-Catalysed Reactions, in: Anonymous, London and New York, 1984.
[10] N. Kulkarni, M. Lakshmikumaran, M. Rao, Xylanase II from an alkaliphilic thermophilic *Bacillus* with a distinctly different structure from other xylanases: evolutionary relationship to alkaliphilic xylanases, Biochem. Biophys. Res. Commun. 263 (1999) 640–645.
[11] N. Kulkarni, A. Shendye, M. Rao, Molecular and biotechnological aspects of xylanases, FEMS Microbiol. Rev. 23 (1999) 411–456.
[12] K.E. Nelson, R.A. Clayton, S.R. Gill, M.L. Gwinn, R.J. Dodson, D.H. Haft, E.K. Hickey, J.D. Peterson, W.C. Nelson, K.A. Ketchum, L. McDonald, T.R. Utterback, J.A. Malek, K.D. Linher, M.M. Garrett, A.M. Stewart, M.D. Cotton, M.S. Pratt, C.A. Phillips, D. Richardson, J. Heidelberg, G.G. Sutton, R.D. Fleischmann, J.A. Eisen, O. White, S.L. Salzberg, H.O. Smith, J.C. Venter, C.M. Fraser, Evidence for lateral gene transfer between Archaea and Bacteria from genome sequence of *Thermotoga maritima*, Nature 399 (1999) 323–329.
[13] P.Y. Chou, G.D. Fasman, Empirical predictions of protein conformation, Ann. Rev. Biochem. 47 (1978) 251–276.
[14] L. Lo Leggio, S. Kalogiannis, M.K. Bhat, R.W. Pickersgill, High resolution structure and sequence of *T. aurantiacus* xylanase I: implication for the evolution of thermostability in family 10 xylanases and enzymes with Barrel architecture, Proteins 36 (1999) 295–306.
[15] R. Vadrevu, C.J. Falzone, C.R. Matthews, Partial NMR assignments and secondary structure mapping of the isolated subunit of *Escherichia coli* tryptophan synthase a 29-kDa TIM barrel protein, Protein Sci. 12 (2003) 185–191.
[16] A. Torronen, A. Harkki, J. Rouvinen, Three-dimensional structure of endo-1,4-beta-xylanase II from *Trichoderma reesei*: two conformational states in the active site, EMBO J. 13 (1994) 2493–2501.
[17] U. Krengel, Three-dimensional structure of endo-1,4–xylanase I from *Aspergillus niger*: molecular basis for its low pH optimum, J. Mol. Biol. 263 (1996) 70–78.