Large-Scale Analyses of Glycosylation in Cellulases

Fengfeng Zhou[1,2], Victor Olman[1,2], and Ying Xu[1,2]*

[1] *Computational Systems Biology Laboratory, Department of Biochemistry and Molecular Biology / Institute of Bioinformatics, University of Georgia, Athens, GA 30602-7229, USA;* [2] *BioEnergy Science Center, Oak Ridge National Laboratory, Oak Ridge, TN 37830-8050, USA.*

*Corresponding author. E-mail: xyn@bmb.uga.edu

DOI: 10.1016/S1672-0229(08)60049-2

Cellulases are important glycosyl hydrolases (GHs) that hydrolyze cellulose polymers into smaller oligosaccharides by breaking the cellulose β (1→4) bonds, and they are widely used to produce cellulosic ethanol from the plant biomass. N-linked and O-linked glycosylations were proposed to impact the catalytic efficiency, cellulose binding affinity and the stability of cellulases based on observations of individual cellulases. As far as we know, there has not been any systematic analysis of the distributions of N-linked and O-linked glycosylated residues in cellulases, mainly due to the limited annotations of the relevant functional domains and the glycosylated residues. We have computationally annotated the functional domains and glycosylated residues in cellulases, and conducted a systematic analysis of the distributions of the N-linked and O-linked glycosylated residues in these enzymes. Many N-linked glycosylated residues were known to be in the GH domains of cellulases, but they are there probably just by chance, since the GH domain usually occupies more than half of the sequence length of a cellulase. Our analysis indicates that the O-linked glycosylated residues are significantly enriched in the linker regions between the carbohydrate binding module (CBM) domains and GH domains of cellulases. Possible mechanisms are discussed.

Key words: glycosylation, cellulase, large-scale analyses

Introduction

Photosynthesis fixes the gaseous carbon into plant biomass consisting mainly of cellulose and hemicellulose, providing the basic energy source for all plants (*1*, *2*). Cellulose can be degraded by cellulases that are generally encoded in cellulolytic fungi and bacterial genomes (*3*–*6*). This process closes the carbon cycle, one of the most important material flows on earth. Numerous other carbohydrate hydrolysis enzymes, such as xylanases, are also encoded in cellulolytic fungi and bacterial genomes to degrade other components of the plant cell walls (*7*, *8*). Considering the world-wide shortage of energy, it is scientifically interesting to study the properties of these carbohydrate hydrolysis enzymes.

Cellulases are glycosyl hydrolases that can break the β (1→4) bonds in cellulose, and are the key enzymes in the production of cellulosic ethanol from the plant biomass. They are classified into three major families, endoglucanases (EC 3.2.1.4), cellobiohydrolases (EC 3.2.1.91) and β-glucosidases (EC 3.2.1.21) (*4*, *9*). Cellulases exert their catalytic activities through the following two major ways (*10*). First, some anaerobic bacteria and fungi degrade cellulose by secreting a multi-protein complex, called *cellulosome*, consisting of one or more catalytically inactive scaffoldin proteins and cellulosome-dependent cellulases (CDCs). Cellulosome was first observed in *Clostridium thermocellum* (*11*), which can efficiently hydrolyze crystalline cellulose using as many as 100 assembled components (*12*). The scaffoldin protein has multiple cohesin domains that bind to the dockerin domains in the CDCs and one or more cellulose-specific carbohydrate binding modules (CBMs). Besides the dockerin domains, the CDCs also carry at least one catalytic glycoside hydrolase (GH) domain to break the β (1→4) bonds in cellulose. Second, "free acting" cellulases (FACs), encoded in some thermophilic bacterial genomes, bind to the carbohydrate using their own CBMs. An FAC may have multiple CBM and GH domains, and can efficiently hydrolyze

different substrates (*10*). A cellulase is usually secreted out of the cell before acting on the target cellulose, and carries a signal peptide on the N-terminal for subcellular localization (*10*, *12*).

Glycosylation is one of the most common and important post-translational modifications of proteins, which is known to play an essential role in the function, structural folding and the stability of proteins (*13*–*15*). Two major types of glycosylation, N-linked and O-linked, were frequently observed in cellulases. The most extensively studied cellulase, Cel7A encoded in *Trichoderma reesei* (*Hypocrea jecorina*), carries a highly O-linked glycosylated linker peptide between its GH domain and CBM domain (*16*). Multiple N-linked glycosylated residues were also identified in the GH domain of *T. reesei* Cel7A, and an altered level of N-linked glycosylation was observed to have significant impacts on the activity and cellulose binding affinity of *T. reesei* Cel7A (*3*, *17*).

We have computationally annotated the functional domains and glycosylated residues in all known cellulases, and conducted a systematic analysis of the distributions of the N-linked and O-linked glycosylated residues in these enzymes. We have made a number of interesting observations based on this analysis. To the best of our knowledge, this work represents the first such systematic analysis of the distributions of the N-linked and O-linked glycosylations in cellulases and their possible impacts on the function and the stability of cellulases. We found from our analysis that the N-linked glycosylated residues mainly appear in the GH domains, but they may appear there just by chance. We also observed that the O-linked glycosylated residues are significantly enriched in the linker regions between the GH and the CBM domains, confirming a previous observation that the linker regions are highly O-linked glycosylated and they are protected from proteolytic degradation through the O-linked glycosylation (*18*).

Results

N-linked glycosylated residues in catalytic domains of cellulases

The GH domain of *T. reesei* Cel7A contains multiple N-linked glycosylation motifs, and the majority of them are experimentally confirmed to be N-linked glycosylated (*3*, *16*, *17*). The level of N-linked glycosylation significantly affects the catalytic activity and the cellulose binding affinity of *T. reesei* Cel7A (*17*). Therefore, we studied whether the N-linked glycosylated residues are significantly enriched in the GH domains from the dataset of 216 cellulases (**Table S1**). Among them, we analyzed all the 19 proteins annotated to have both the GH domains and N-linked glycosylated residues in UniProtKB/Swiss-Prot (*19*), with their annotation details given in **Table S2**. The GH domains occupy at least 68.18% of the total sequence length of each cellulase under consideration. From Table S2, we noticed that 15 out of the 19 (78.95%) proteins have *Nvalid*=100%, and the other 4 proteins have their *Nvalid* at 50% or more, where *Nvalid* represents the percentage of the N-linked glycosylated residues within the GH domains out of the total number of such residues in each cellulase. None of the 19 proteins has N-linked glycosylated residues (statistically) significantly enriched in the GH domains with $P \leq 0.05$ (Table S2), since the GH domains are relatively large within each cellulase. It is interesting to further study this problem using computationally annotated domains.

In our combined dataset of the experimentally known and computationally predicted annotations, 195 out of 216 cellulases (90.28%) have both N-linked glycosylated residues and the GH domains (**Table S3**). We noticed that at least half of the N-linked glycosylated residues in 143 out of 195 (73.33%) cellulases are within the GH domains (*Nvalid*≥50%), and all the N-linked glycosylated residues in 92 out of 195 (47.18%) cellulases are within the GH domains (*Nvalid*=100%) (**Figure 1**). Since the GH domains occupy at least half of the sequence lengths in 196 out of 216 (90.74%) cellulases, it is important to check whether the N-linked glycosylated residues appear in the GH domains of the cellulases just by chance or are statistically enriched in those regions. This can be done by checking the following null hypothesis that the N-linked glycosylated residue appears in a position within and outside the GH domains with the same probability, which will be rejected on the 5% level of the type-1 classification error.

Among all the cellulases under study, only four (BGLX_ECOLI, BGLX_SALTY, GUN21_ARATH and GUND_CLOCE) have a P value ≤0.05 with an enrichment ratio (ER) >1 (see Materials and Methods), hence the null hypothesis could not be rejected for the other cellulases, suggesting that the N-linked glycosylation does not have preference to modify the residues in the GH domains of cellulases. Similar patterns were observed for the three families of cellulases,

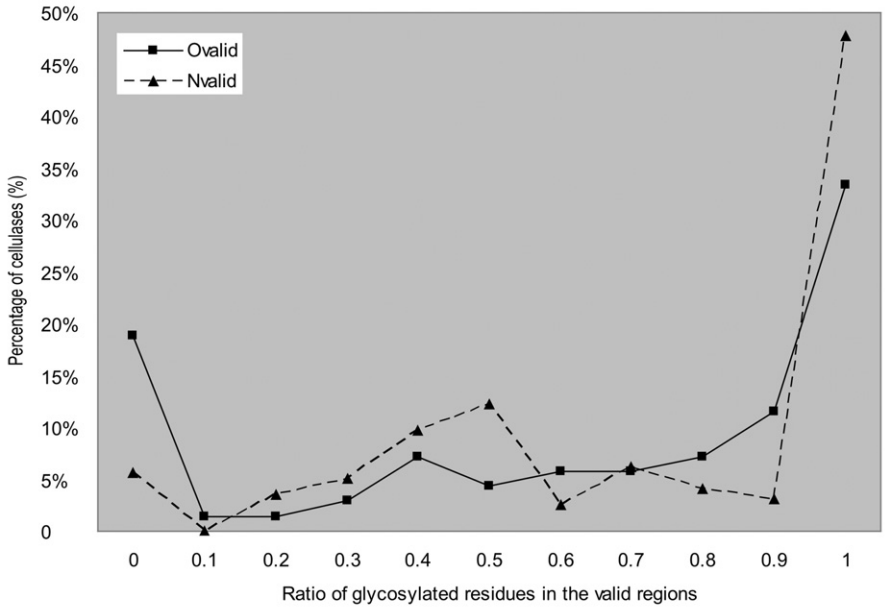

Figure 1 Ratios of N-linked glycosylated residues within the GH domains and O-linked glycosylated residues within the linker regions of cellulases in the dataset combined from known and predicted glycosylated residues. These regions are called the valid regions for N-linked (Nvalid) and O-linked (Ovalid) glycosylated residues, respectively.

i.e., endoglucanases, cellobiohydrolases and β-glucosidases, respectively.

O-linked glycosylated residues in linker regions of cellulases

Linker regions between the GH domains and the CBM domains of known cellulases were reported to be highly O-linked glycosylated (*16*) and protected from proteolysis by the O-linked glycosylation (*18*). Only one protein, the *T. reesei* Cbh2 (GUX2_TRIRE), was found to contain both GH domain, CBM domain and O-linked glycosylated residue in the combined UniProtKB/Swiss-Prot database (*19*), as shown in **Figure 2** and Table S2. However, none of the seven O-linked glycosylated residues in *T. reesei* Cbh2 is within the linker region, rather they are all in the N-terminal region of 40 amino acids (aa) of the GH domain. This suggests that the flanking regions of the linker may also play a role in protecting the protein from proteolysis by the O-linked glycosylation. Compared with the length of 471 aa of *T. reesei* Cbh2, the seven O-linked glycosylated residues in this protein are enriched with a ratio of 5.81 [(7/81)/(7/471)] in the region consisting of the 41-aa linker and the 40-aa flanking region in the catalytic domain. The P value of 5.28e-09 was calculated using the formula as shown in Materials and Methods. This P value was significantly smaller than the commonly used threshold 0.05, suggesting that the O-linked glycosylated residues are significantly enriched in the region consisting of the linker and its flanking 40-aa regions.

We further studied if O-linked glycosylated residues are enriched in the linker regions between the GH domains and the CBM domains of cellulases using computationally annotated domains and glycosylated residues. Our dataset was combined from the experimentally known and computationally predicted annotations, in which 69 out of 216 cellulases (31.94%) have GH domains, CBM domains and O-glycosylated residues together (Table S3). We verified our hypothesis in these 69 cellulases with linker regions.

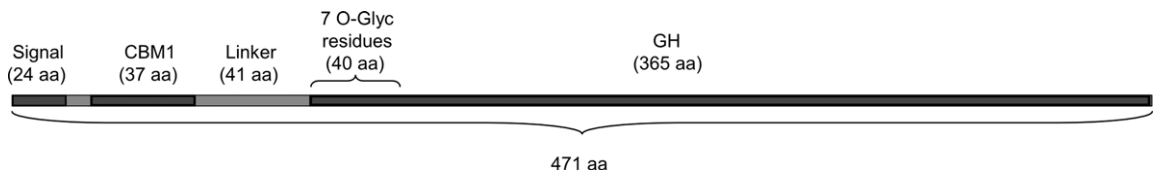

Figure 2 Domain architecture of *T. reesei* Cbh2 (GUX2_TRIRE) and the experimentally verified O-linked glycosylated residues in it.

Figure 3 Domain architecture of *T. reesei* Cbh2 (GUX2_TRIRE) and the predicted plus experimentally verified 4 N-linked (in red vertical lines) and 25 O-linked (in green vertical lines) glycosylated residues in it.

The linker regions occupy less than 17% of the total lengths of all the 69 cellulases except for two (GUNB_PSEFL and GUNA_CELFI), whereas at least 25% of the O-linked glycosylated residues in 54 out of the 69 (78.26%) cellulases appear in the linker regions. Actually the O-linked glycosylated residues are enriched in the linker regions of 54 (78.26%) cellulases ($P \leq 0.05$ and ER>1), except for the 15 endoglucanases.

A case study of Cbh2 in *T. reesei*

There are 4 N-linked and 25 O-linked glycosylated residues in *T. reesei* Cbh2 based on the predicted and experimentally verified data. *T. reesei* Cbh2 has a signal peptide, a CBM domain and a GH domain, as shown in **Figure 3**. All except for one (38*Asn*) N-linked glycosylated residues are within the GH domain, but the P value (0.55) cannot reject the null hypothesis that the N-linked glycosylated residues appear within and outside the GH domain of *T. reesei* Cbh2 with the same probability. There are 17 (68%) O-linked glycosylated residues in the linker region of *T. reesei* Cbh2. They are significantly enriched in the linker region of *T. reesei* Cbh2 with a P value of 1.95e-25. There are seven more O-linked glycosylated residues in the linker's 40-aa flanking regions, and the P value to reject the null hypothesis within the link and its 40-aa flanking regions could reach 4.71e-26.

Discussion

Despite the essential role of glycosylation in the cellulases (*3*, *16–18*), there has not been any systematic study on the distribution and enrichment of glycosylated residues within these enzymes, as far as we know. The lack of systematic analyses is mainly due to the limited annotations of functional domains and glycosylated residues in cellulases. For example, only 8.8% out of the 216 known full-length cellulases have annotations of both the GH domains and N-linked glycosylated residues, and only one cellulase can be used to verify the distribution of O-linked glycosylated residues, as shown in Table S2. It is time consuming and labor intensive to identify the functional domains and all glycosylated residues in the 216 cellulases using experimental techniques. So we carried out computational annotations for the cellulases and a systematic analysis of the distributions of N-linked and O-linked glycosylated residues in these cellulases.

Many N-linked glycosylated residues appear in the GH domains of cellulases, and removing the N-linked glycans will impact the catalytic efficiency and cellulose binding affinity of cellulases (*3*, *17*). It is necessary to investigate whether the N-linked glycosylated residues appear in the GH domains of cellulases just by chance or tend to be within these regions to exert their regulatory role. Our analysis based on Tables S2 and S3 suggests that N-linked glycosylated residues do not have any tendency to be within these domain regions. Our analysis also suggests that N-linked glycosylations carry out their regulatory functions through modifying the 3D structures of cellulases by attaching N-linked glycans to the largely uniformly distributed amino acid *Asn* (N) in cellulases.

O-linked glycosylation was proposed to protect the linker regions of cellulases from proteolysis (*16*). However, it is impossible to verify this based on the available annotations in Table S2, since there is only one cellulase (GUX2_TRIRE) with annotations of a CBM domain, a GH domain and O-linked glycosylated residues, and all the known O-linked glycosylated residues in GUX2_TRIRE are not in the linker region between the CBM domain and the GH domain. We have annotated the domains and the O-linked glycosylated residues in the 216 cellulases, and investigated the distribution of O-linked glycosylated residues in the cellulases based on the comprehensive dataset combined from the predicted and experimentally verified annotations. Table S3 shows that the O-linked glycosylated residues are significantly enriched in the linker regions of cellulases, and at least a quarter of the O-linked glycosylated residues in 54 cellulases are within the linker regions. The data support the previous hypothesis that O-linked glycosylation protects the linker regions from proteolysis mainly through competing access to some *Ser* residues in the linker regions with proteases by attaching O-linked glycans to those residues (*16*).

There may be underestimation of the above analysis, due to the limitations of the computational anno-

tations of protein domains and glycosylated residues. The Pfam annotation system (*20*) has been widely used to identify protein domains (*21–23*), but many protein domains are not included in this system, and 25% of the proteins in the UniProt Knowledgebase do not have a curated Pfam domain (*20*, *24*). The N-linked and O-linked glycosylated residues were predicted using one of the best predictors for the glycosylated residues, EnsembleGly (*13*), which reaches 94% in prediction accuracy and do not have known bias for the distribution of the predicted glycosylated residues. So the computational annotations in this work are statistically significant enough to support our hypothesis, and it would be interesting to verify the hypothesis in those cellulases without recognizable GH domains and CBM domains in the current version of the Pfam system.

Fossil fuels are rapidly depleting and it is essential to develop sustainable next-generation fuel sources. Cellulolytic ethanol, one of the most promising fuel sources, is generated from biomass through cellulases and other glycosyl hydrolases (*25*). The large-scale analysis in this work greatly improves our understanding of the role of glycosylations in the stabilization and regulation of cellulases. We are working on the engineered efficient cellulases based on the knowledge in this work and the literature.

Materials and Methods

Data collection and *in silico* annotation

Cellulases were classified into three classes [endoglucanases (EC 3.2.1.4), cellobiohydrolases (EC 3.2.1.91) and β-glucosidases (EC 3.2.1.21)] (*4*, *9*) in the ENZYME database (release of July 1, 2008) (*26*), and their sequences and functional annotations were retrieved from the combined UniProtKB/Swiss-Prot database (*19*). We removed the partial or fragment peptides. The final dataset consists of 161, 30 and 25 proteins for endoglucanases, cellobiohydrolases and β-glucosidases, respectively. We used the program SignalP 3.0 (*27*) to identify the N-terminal signal peptides in the cellulases and the program Pfam_scan.pl based on the database Pfam (release 22.0 of July 2007) to annotate functional domains in the cellulases (*28*). The detailed information of the 216 proteins can be found in Table S1.

The N-linked and O-linked glycosylated residues in the enzymes were predicted using one of the best programs EnsembleGly (*13*).

Enrichment analyses

We investigated whether the N-linked glycosylated residues are significantly enriched in the catalytic GH domains of cellulases, using a statistical approach. We propose a null hypothesis that the probability, q, of the N-linked glycosylated residue appearing in any position of the GH domains of a given cellulase is the same as the probability, p, of this residue in any position outside the GH domains. In other words, we are testing the statistical hypothesis that occurrence of N-linked glycosylated residues satisfies the same Bernoulli process along the whole protein.

Let the length of the given cellulase and the length of its GH domains be N and M, respectively, and n be the number of N-linked glycosylated residues with m of them belonging to the GH domains. The enrichment ratio (ER) of N-linked glycosylated residues in the GH domains of this cellulase is defined as:

$$ER = \frac{m/M}{n/N}$$

The maximum likelihood estimators for p and q are $\hat{p} = \frac{n-m}{N-M}$ and $\hat{q} = \frac{m}{M}$. Based on the Central Limit Theorem, these estimators are approximately Gaussian, *i.e.*, $\hat{q} \sim N\left(q, \frac{q(1-q)}{M}\right)$ and $\hat{p} \sim N\left(p, \frac{p(1-p)}{N-M}\right)$. If the null hypothesis is true ($p = q$), then $\hat{q} - \hat{p} \sim N\left(0, p(1-p)(\frac{1}{M} + \frac{1}{N-M})\right)$. Substituting p with its estimator $\frac{n}{N}$ (if the hypothesis is true), we get $\hat{q} - \hat{p} \sim N\left(0, \frac{n(N-n)}{N(N-M)M}\right)$ with P value $= 1 - \phi\left(\varepsilon\sqrt{\frac{NM(N-M)}{n(N-n)}}\right)$, where $\varepsilon = \hat{q} - \hat{p}$, and $\phi()$ is a cumulative distribution function of the standard Gaussian distribution.

The ER and P value can be defined similarly for the null hypothesis that the probabilities of O-linked glycosylated residues appearing in the linker regions between the catalytic GH domains and the CBM domains and outside the linker regions are the same to each other.

Acknowledgements

We thank all the CSBL colleagues for their comments on this work, and also thank the reviewers for the helpful comments. This work was supported in part by the National Science Foundation of

USA (Grants DBI-0354771, ITR-IIS-0407204, DBI-0542119 and CCF0621700), the grant for the BioEnergy Science Center, and a Distinguished Scholar grant from the Georgia Cancer Coalition.

Authors' contributions

FZ collected the data, performed the analyses and wrote the manuscript. VO participated in the statistical analyses. YX supervised the project and co-wrote the manuscript. All authors read and approved the final manuscript.

Competing interests

The authors have declared that no competing interests exist.

References

1. Lynd, L.R., *et al.* 2002. Microbial cellulose utilization: fundamentals and biotechnology. *Microbiol. Mol. Biol. Rev.* 66: 506-577.
2. Mellerowicz, E.J. and Sundberg, B. 2008. Wood cell walls: biosynthesis, developmental dynamics and their implications for wood properties. *Curr. Opin. Plant Biol.* 11: 293-300.
3. Eriksson, T., *et al.* 2004. Heterogeneity of homologously expressed *Hypocrea jecorina* (*Trichoderma reesei*) Cel7B catalytic module. *Eur. J. Biochem.* 271: 1266-1276.
4. Romero, M.D., *et al.* 1999. Cellulase production by *Neurospora crassa* on wheat straw. *Enzyme Microb. Technol.* 25: 244-250.
5. Béguin, P. and Lemaire, M. 1996. The cellulosome: an exocellular, multiprotein complex specialized in cellulose degradation. *Crit. Rev. Biochem. Mol. Biol.* 31: 201-236.
6. Mitchell, W.J. 1998. Physiology of carbohydrate to solvent conversion by clostridia. *Adv. Microb. Physiol.* 39: 31-130.
7. Beliën, T., *et al.* 2006. Microbial endoxylanases: effective weapons to breach the plant cell-wall barrier or, rather, triggers of plant defense systems? *Mol. Plant Microbe Interact.* 19: 1072-1081.
8. Sunna, A. and Antranikian, G. 1997. Xylanolytic enzymes from fungi and bacteria. *Crit. Rev. Biotechnol.* 17: 39-67.
9. Criquet, S. 2002. Measurement and characterization of cellulase activity in sclerophyllous forest litter. *J. Microbiol. Methods* 50: 165-173.
10. Blumer-Schuette, S.E., *et al.* 2008. Extremely thermophilic microorganisms for biomass conversion: status and prospects. *Curr. Opin. Biotechnol.* 19: 210-217.
11. Gilbert, H.J. 2007. Cellulosomes: microbial nanomachines that display plasticity in quaternary structure. *Mol. Microbiol.* 63: 1568-1576.
12. Ding, S.Y., *et al.* 2008. A biophysical perspective on the cellulosome: new opportunities for biomass conversion. *Curr. Opin. Biotechnol.* 19: 218-227.
13. Caragea, C., *et al.* 2007. Glycosylation site prediction using ensembles of support vector machine classifiers. *BMC Bioinformatics* 8: 438.
14. Imperiali, B. and O'Connor, S.E. 1999. Effect of N-linked glycosylation on glycopeptide and glycoprotein structure. *Curr. Opin. Chem. Biol.* 3: 643-649.
15. Bosques, C.J., *et al.* 2004. Effects of glycosylation on peptide conformation: a synergistic experimental and computational study. *J. Am. Chem. Soc.* 126: 8421-8425.
16. Harrison, M.J., *et al.* 1998. Modified glycosylation of cellobiohydrolase I from a high cellulase-producing mutant strain of *Trichoderma reesei*. *Eur. J. Biochem.* 256: 119-127.
17. Jeoh, T., *et al.* 2008. Implications of cellobiohydrolase glycosylation for use in biomass conversion. *Biotechnol. Biofuels* 1: 10.
18. MacLeod, A.M., *et al.* 1992. Streptomyces lividans glycosylates an exoglucanase (Cex) from *Cellulomonas fimi*. *Gene* 121: 143-147.
19. Boutet, E., *et al.* 2007. UniProtKB/Swiss-Prot. *Methods Mol. Biol.* 406: 89-112.
20. Mistry, J. and Finn, R. 2007. Pfam: a domain-centric method for analyzing proteins and proteomes. *Methods Mol. Biol.* 396: 43-58.
21. Wei, G., *et al.* 2009. A transcriptomic analysis of superhybrid rice LYP9 and its parents. *Proc. Natl. Acad. Sci. USA* 106: 7695-7701.
22. Bidargaddi, N.P., *et al.* 2008. Hidden Markov models incorporating fuzzy measures and integrals for protein sequence identification and alignment. *Genomics Proteomics Bioinformatics* 6: 98-110.
23. Zhao, X.Q., *et al.* 2006. Comparative analysis of eubacterial DNA polymerase III alpha subunits. *Genomics Proteomics Bioinformatics* 4: 203-211.
24. Wu, C.H., *et al.* 2006. The Universal Protein Resource (UniProt): an expanding universe of protein information. *Nucleic Acids Res.* 34: D187-191.
25. Mukhopadhyay, A., *et al.* 2008. Importance of systems biology in engineering microbes for biofuel production. *Curr. Opin. Biotechnol.* 19: 228-234.
26. Bairoch, A. 2000. The ENZYME database in 2000. *Nucleic Acids Res.* 28: 304-305.
27. Bendtsen, J.D., *et al.* 2004. Improved prediction of signal peptides: SignalP 3.0. *J. Mol. Biol.* 340: 783-795.
28. Finn, R.D., *et al.* 2006. Pfam: clans, web tools and services. *Nucleic Acids Res.* 34: D247-251.

Supporting Online Material
Tables S1–S3
DOI: 10.1016/S1672-0229(08)60049-2