Computational analysis of the phylogeny and thermodynamics of glycoside hydrolases

by

Blake Mertz

A dissertation submitted to the graduate faculty in partial fulfillment of the requirements for

the degree of

DOCTOR OF PHILOSOPHY

Major: Chemical Engineering

Program of Study Committee:
Peter J. Reilly, Major Professor
Mark S. Hargrove
James C. Hill
Monica H. Lamm
Xun Gu

Iowa State University

Ames, Iowa

2008
TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>TABLE OF CONTENTS</th>
<th>ii</th>
</tr>
</thead>
<tbody>
<tr>
<td>CHAPTER 1. Introduction</td>
<td>1</td>
</tr>
<tr>
<td>Rationale</td>
<td>1</td>
</tr>
<tr>
<td>Research questions</td>
<td>1</td>
</tr>
<tr>
<td>Organization of this study</td>
<td>3</td>
</tr>
<tr>
<td>References</td>
<td>3</td>
</tr>
<tr>
<td>CHAPTER 2. Theory and methods</td>
<td>6</td>
</tr>
<tr>
<td>Phylogenetic analysis</td>
<td>6</td>
</tr>
<tr>
<td>Theory and explanation of terms</td>
<td>6</td>
</tr>
<tr>
<td>PHYLIP</td>
<td>8</td>
</tr>
<tr>
<td>DIVERGE</td>
<td>9</td>
</tr>
<tr>
<td>Automated docking</td>
<td>10</td>
</tr>
<tr>
<td>Theory</td>
<td>10</td>
</tr>
<tr>
<td>Methodology</td>
<td>15</td>
</tr>
<tr>
<td>Molecular dynamics</td>
<td>16</td>
</tr>
<tr>
<td>References</td>
<td>17</td>
</tr>
<tr>
<td>CHAPTER 3. Phylogenetic analysis of family 6 glycoside hydrolases</td>
<td>19</td>
</tr>
<tr>
<td>Abstract</td>
<td>19</td>
</tr>
<tr>
<td>Introduction</td>
<td>19</td>
</tr>
<tr>
<td>Computational methods</td>
<td>22</td>
</tr>
<tr>
<td>Results and discussion</td>
<td>23</td>
</tr>
<tr>
<td>Phylogenetic analysis</td>
<td>23</td>
</tr>
<tr>
<td>Conservation of putative residues</td>
<td>30</td>
</tr>
<tr>
<td>Lengths of tunnel-forming loops</td>
<td>30</td>
</tr>
<tr>
<td>Conclusions</td>
<td>31</td>
</tr>
<tr>
<td>References</td>
<td>32</td>
</tr>
<tr>
<td>CHAPTER 4. Automated docking to explore subsite binding by glycoside hydrolase family 6 cellobiohydrolases and endoglucanases</td>
<td>35</td>
</tr>
<tr>
<td>Abstract</td>
<td>35</td>
</tr>
</tbody>
</table>
Introduction 36
Computational methods 39
Results 42
   Docking in the H. jecorina CBH active site 42
   Docking in the T. fusca EG active site 44
   Scissile glycosidic bond twisting 46
   Net processive forces 47
Discussion 47
Conclusions 52
References 52

CHAPTER 5. Analysis of functional divergence within two structurally related
   glycoside hydrolase families 58
Abstract 58
Introduction 59
Computational methods 61
   Sequence analysis 61
   Structural analysis 62
Results and discussion 63
   GH7 functional divergence 63
   GH16 functional divergence 71
Conclusions 83
References 83

CHAPTER 6. General conclusions 89
   General discussion 89
   Recommendations for future research 89
      Transition-state pathways of glycoside hydrolases 89
      Extension of functional divergence applications to cellulases 91
References 92

APPENDIX A. Functional divergence in glycoside hydrolase clan A enzymes that attack
   cellulosic complexes 94
APPENDIX B. Molecular dynamics simulations of Dicer ribonuclease 104

Specific aims 104
Rationale 108
Computational research 109
References 110

Supplemental Material is available upon request.
Chapter 1. Introduction

Rationale

Computational simulation plays an increasingly important role in the methodology that scientists use to answer research problems of many scales, from microscopic to atmospheric. Simulation has been used to examine the catalytic mechanisms of enzymes that could play a role as anticancer agents,\(^\text{1}\) to infer genetic relationships among groups of organisms,\(^\text{2}\) and to mimic the various processes that take place in a biodiesel plant.\(^\text{3}\)

One unique advantage of computational simulation is that it provides the opportunity to broaden the range of problems that can be examined in research. The conventional path in the research industry has been to develop a broad knowledge base to serve as the foundation for more detailed research in a specific area, in order to make incremental contributions to the scientific community's overall understanding of a research problem. With the increasing emphasis on interdisciplinary collaborative approaches to research, it is necessary to have a working knowledge of many fields, and computational simulations help to bridge the gap between applied and theoretical science, providing a testing environment for comparison to physical data. An excellent example of this is the use of molecular dynamics (MD) and hybrid quantum mechanics/molecular mechanics (QM/MM) to simulate the molecular interactions in cellular systems, which allows identification of essential components in reaction mechanisms.\(^\text{4}\) These insights lead to development of methods in the lab for attenuating or amplifying specific reactions.

Research Questions

There are widespread initiatives in the United States to develop alternative energy sources,\(^\text{5,6}\) and chief among them is the search for a viable alternative to gasoline. Over 60% of U.S. oil supplies are imported annually,\(^\text{7}\) creating an economic dependence on countries with unstable geopolitical climates. Efforts have been made to follow Brazil's lead in creating a self-sufficient transportation industry by developing ethanol production and use, and although the U.S. has surpassed Brazil in overall ethanol production,\(^\text{8}\) America also consumes 9.4 times as much fuel as Brazil,\(^\text{9}\) along with converting 14% of the national corn crop to 4.3 million gallons of ethanol annually.\(^\text{10}\) Cellulose can potentially supplant corn as the main
feedstock for ethanol, but supplying the group of enzymes necessary for cellulose hydrolysis remains the cost-prohibitive step in developing a product that can compete with oil. Glycoside hydrolase family 6 (GH6) cellobiohydrolases (CBHs) and endoglucanases (EGs) are two of the common enzymes found in the naturally-occurring protein ensemble used to degrade cellulose.\textsuperscript{11} The hydrolysis mechanism is understood, but identification of a conserved nucleophile/base amino acid has remained elusive.\textsuperscript{12} Our first research objective is to delineate the evolutionary relationships between GH6 enzymes and to use these results to establish the connection between the primary and tertiary structures that dictates substrate specificity. This will be addressed in Chapters 3 and 4.

Another potential avenue for increasing cellulase efficiency is through the use of bioinformatics to identify amino acid residues responsible for functional divergence between two groups of phylogenetically related enzymes. Cellulase engineering typically employs either rational design or directed evolution to improve enzymatic efficiency,\textsuperscript{13} but each method has its limitations. Rational design uses knowledge of protein structure and function, along with site-directed mutagenesis and screening, to attempt to produce a desired characteristic in the engineered protein. It suffers from a reliance on the availability of enzyme structure and function information, which does not always occur, and even when it does, it may not lead to the desired change in function.\textsuperscript{14} Another shortcoming of rational design is the large time requirement for mutagenesis and screening; an enzyme active site is the typical focus of a rational design experiment, but quite often protein function and stability is affected by residues outside the active site, as shown by a case study on the $10^6$-fold increase in $k_{cat}/K_m$ of aspartate aminotransferase that involved a 17-residue mutation – only one amino acid was located in the active site.\textsuperscript{15} With the exponential growth of primary protein sequences,\textsuperscript{16} bioinformatics methods can statistically identify amino acid residues responsible for functional divergence based on primary structure\textsuperscript{17} are becoming very advantageous as a tool to improve the efficiency of rational design. Coupling functional divergence results with MD studies would allow researchers to carry out an initial mutagenesis screening process for enzyme stability and potential function, thus making the rational design process even more efficient. Our second research objective is to implement functional divergence analysis
in order to identify additional leads for use in cellulase engineering. This will be addressed in Chapters 5 and Appendix A.

Organization of This Study

Research aims have followed the growth in computational knowledge of our research group. Initially we conducted analyses of glycosidic bonds in various glycoside hydrolase substrates using MM3, and eventually we began docking simulations of enzyme-substrate complexes using AutoDock. This was chosen because it fit well with the computational resources available to our group and had proven to give comparable results to more robust molecular mechanics and MD programs. Studies of cellulosic hydrolysis naturally extended from our research in starch hydrolysis, forming the basis for examining enzyme-ligand interactions in GH6 (Chapter 4) and other cellulase families (Chapter 5 and Appendix A).

Phylogenetic analysis has given us insight into identification of catalytic residues and delineation of appropriate subfamilies within glycoside hydrolase families (Chapters 3 and 5). PHYLIP is a phylogenetic program that can analyze a multiple sequence alignment by the maximum parsimony, maximum likelihood, or neighbor-joining methods and produce a phylogenetic tree that identifies potential subfamilies and shows evolutionary distances between species for a given protein. DIVERGE is a program that extends our phylogenetic analysis by identifying gene clusters which have undergone functional divergence due to gene duplication or speciation events and also by locating amino acid residues responsible for changes in functional constraint.

References


Chapter 2. Theory and Methods

Phylogenetic Analysis

Theory and Explanation of Terms. The aim of phylogenetic analysis is to map the evolution of a protein family from one common ancestor to the present homologous set. Underlying this analysis is the assumption that motifs responsible for protein function are conserved throughout the course of evolution, and point mutations of specific amino acid residues help determine the extent to which evolution has occurred within a species. Both the proximity and grouping of a protein family are illustrated in a phylogenetic tree.

Construction of the tree begins with a multiple sequence alignment (MSA) of the proteins under consideration. Typically an alignment program begins by creating a guide tree that determines the order in which primary sequences are added to the MSA. Gaps in a sequence are introduced and/or extended to obtain higher homology with other sequences, but each of these actions incur a penalty, to retain the genetic code as much as possible.

Gap creation carries a much greater penalty than gap extension because it has an effect that can extend to secondary structures – breaking up an α-helix, β-strand, or active site loop would infer poor homology between function-conserved proteins and result in a poor MSA. Adjustment by hand follows the initial MSA, since programs will commonly intersperse small deletions at different positions in a motif sequence area for proteins with good homology. The final MSA is then used as the input for generation of the phylogenetic tree.

Building a tree can be achieved several ways, most thoroughly by discrete data methods, due to their comparison analysis of every column in the MSA. Examples of discrete data methods are maximum parsimony and maximum likelihood methods. Maximum parsimony (MP) takes the entire tree topology and infers an ancestral sequence, measuring the number and degree of residue substitutions that occur at each position. Each bootstrapped tree is analyzed, and the topology containing the fewest residue substitutions, or shortest tree length, is chosen as the best tree. MP operates under the assumption of Occam's Razor, which states that the simplest solution is the best solution. The advantage of using MP is that it lacks the assumptions for residue substitution inherent in likelihood methods (to be discussed in the following paragraphs), producing a more reliable tree when sequence divergence is low. Unfortunately, this is not always the case; when divergence is relatively high or substitution
rates vary, branches in the final tree will tend to attract each other or even merge together, a phenomenon called long- (or short-) branch attraction. Due to the large number of possible topologies that can be generated from an MSA of greater than 20 sequences (as given by \[ \frac{(2m-4)!}{2^{(m-3)}(m-3)!} = 4.04 \times e^{91} \text{ possibilities}, \] heuristic search methods must be employed in which a small percentage of all possible trees is analyzed. To increase the likelihood of finding the best tree, a stepwise addition algorithm is used in conjunction with branch-swapping to create a provisional MP tree. Programs such as PHYLIP create the provisional tree beginning with the first three sequences in the MSA, and (in the case of PHYLIP) then proceed to randomly add sequences from the rest of the MSA. The provisional MP tree is found for each sequence addition until all sequences have been used, and then tree bisection-reconnection (TBR) is employed, where the provisional MP tree is bisected into two subtrees and then reconnected, testing all possible pairs of branches (Figure 2.1), culminating in the tree with the minimum tree length. Since there is usually more than one tree with the minimum length, a majority rule consensus tree is constructed that represents the tree with a branching pattern that occurs more than 50% of the time in all trees.

Maximum likelihood (ML) uses branch length estimations based on the probability of positional residue mutation (from \(i\) to \(j\)) over time to calculate the ML tree, given by the highest ML value. The likelihood function \(L\) for any ML topology can be expressed as \(L = f(x; \theta)\), where \(x\) is the MSA sequences and \(\theta\) represents branch length, residue frequency, and substitution matrix parameters. The Jones-Taylor-Thornton (JTT) matrix is a common substitution matrix used; similar amino acids (such as valine and isoleucine, which are similar in structure and differ by only by one codon in three instances) have a higher probability for a point mutation from one residue to the other, while dissimilar amino acids have a smaller chance of occurring.

Figure 2.1. The tree bisection-reconnection (TBR) method of branch-swapping. Bisection of the provisional MP tree (center tree) allows for sampling of larger numbers of topologies (outer trees).
The JTT model also assumes that residue frequencies are in equilibrium and remain constant throughout evolution. ML's advantage over other phylogenetic methods is that it gives a more realistic estimate of branch lengths for a given tree, but this increased accuracy also potentially sacrifices the search for the overall best topology.

Several characteristics are notable when examining a phylogenetic tree. They can be either rooted or unrooted; rooted trees splay outward from the common ancestor, whereas unrooted trees radiate from a central set of nodes. The distance between two proteins is a function of the homogeneity (or percent identity) of the two sequences, and can be calculated using a mutation matrix.

Bootstrapping is used to verify the reproducibility and accuracy of the original phylogenetic tree. What occurs is a random sampling with replacement of amino acid residue combinations at each position in the MSA, generating new alignments on the order of \(1 \times 10^2\) or more. Trees are created for each of these new MSAs, and each branch of the original tree is assigned a percentage frequency with which it recurred in the bootstrapped trees (a consensus tree). A bootstrapping frequency of 70% or higher has been shown to indicate a 95% representation of the correct tree.\(^6\)

**PHYLIP.** PHYLIP\(^7\) is a program package used to generate phylogenetic trees from MSAs. Each program carries out a specific task, and they are designed to be used sequentially to produce the best possible phylogenetic tree.\(^8\) PROTPARS utilizes MP under the previously described principles, along with making several assumptions that merge traditional methods, the most important of them being that change in different sites and lineages is independent, as well as having a much higher chance for a synonymous base change in an amino acid than a non-synonymous one.\(^9,10\) PROML is based on ML principles and several of the PROTPARS assumptions. A Hidden Markov Method is also implemented, freeing the ML search from assuming constant evolutionary rates by inferring varying evolution rates at different amino acid positions.\(^11\) PROTDIST is the branch length program of PHYLIP, and it takes the MSA and uses the Dayhoff PAM 001, JTT, or PMB matrices, the Kimura distance model, or Felsenstein's Categories distance model to calculate branch length between two species. Bootstrapping analysis is done in SEQBOOT, which allows the user to perform a bootstrap, jackknife, or permutation of the MSA data set. CONSENSE will
take the pool of trees generated from the MP or ML programs and create a consensus tree. Two tree-building options are available in CONSENSE: 1) a strict consensus tree, where visible branches are those present in all the trees; 2) a majority-rule consensus tree, in which all branches are shown that occur in at least 50% of the trees. The majority-rule method can be user-defined for a value from 50-100%, allowing for a wide range of possible consensus trees.

**Figure 2.2.** Description of types I and II functional divergence (FD). Evolutionary rate sharply increases after gene duplication (early stage). As the rate decreases, type II FD will return to the original evolutionary rate, whereas type I FD will show either an increased or decreased evolutionary rate.13

DIVERGE. DIVERGE\textsuperscript{12} is a program designed to identify amino acid residues that have undergone a change in functional constraints due to a gene duplication event. This identification is achieved by performing a two-step procedure of statistically testing an MSA for presence of functional divergence (FD) by a likelihood ratio test and then doing a posterior analysis to find the probability of a specific residue being in a state of FD.\textsuperscript{12} A phylogenetic tree along with display of identified amino acid residues in a tertiary structure is shown as the output in DIVERGE.

FD is the relationship two orthologous (derived from a common ancestor) sequences have with respect to their evolutionary rates, or $\lambda$. Gene duplication and speciation are the two sources of distinguishable functional differences within gene families, and Gu developed a stochastic model that estimates the extent of FD and helps to identify amino acid residues responsible for these events.\textsuperscript{13} The model is based on the hypothesis that gene duplication is responsible for increasing $\lambda$ immediately after the duplication event, and that purifying selection controls $\lambda$ during the later stages of evolution (Figure 2.2). Two types of FD exist: type I FD occurs when an altered functional constraint exists in an orthologous gene pair, due to either an increase or decrease in $\lambda$ when compared to the original rate of $\lambda$, and type II FD
shows no change in functional constraints during the total evolutionary timescale. The
dependence of FD, functional constraint, and $\lambda$ provides the basis for evaluation of type I FD.

The extent of functional divergence can be expressed by

$$\theta_A = 1 - \frac{\sigma_{12}}{\sqrt{(V_1 - D_1)(V_2 - D_2)}}$$

where $\theta_A$ is the coefficient of FD, $D_1$ and $D_2$ are the mean value of the number of changes in
gene clusters 1 and 2, $V_1$ and $V_2$ are the variances of the number of changes in each cluster,
and $\sigma_{12}$ is the covariance over sites between clusters 1 and 2. Statistical significance of the
coefficient of FD is verified by testing the null hypothesis $\theta_A = 0$, which depends on knowing
the number of changes at each site in the gene clusters, given by $X_i$, where $i = 1, 2$. Gu and
Zhang developed a phylogeny method combining inference of an ancestral sequence with a
ML estimation to calculate $X_i$; obtaining $X_i$ also allows for calculation of a posterior
probability ($P(S_i|X)$, where $S_i$ represents a state of being in type I FD) which correlates to the
presence of a site-specific altered functional constraint. Identification of amino acids that are
responsible for the functional-structural differences of protein families is invaluable, due to
the time saved by computational prediction and the relative ease with which the functional
importance of these residues can be verified experimentally.

**Automated docking**

**Theory.** AutoDock 3.0 uses molecular mechanics to analyze the binding energies in an
enzyme-ligand complex, based on the principles of quantitative structure-activity
relationships (QSAR). It is designed to rapidly and efficiently sample the conformational
space of enzyme-ligand complexes using a combination of grid-based energy evaluation and
an empirical Gibbs free energy function. The free energy function is broken down into five
components, based on the following equation:

$$\Delta G = \Delta G_{vdW} \sum_{i,j} \left( \frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}} \right) + \Delta G_{hbond} \sum_{i,j} E(t) \left( \frac{C_{ij}}{r_{ij}^{12}} - \frac{D_{ij}}{r_{ij}^{10}} + E_{hbond} \right) +$$

$$\Delta G_{elec} \sum_{i,j} \frac{q_i q_j}{\epsilon(r_{ij}) r_{ij}} + \Delta G_{tor} N_{tor} + \Delta G_{solv} \sum_{q_i} S_i V_f e^{-\frac{r_{ij}^2}{2\sigma^2}}$$

(2)
Each Gibbs free energy term on the right hand side of the equation is an empirical weighting factor, determined by linear regression against a training set of protein-ligand complexes with known experimental binding constants; these weighting terms are not used, due to the fact that Morris et al.’s training set was performed with few protein-carbohydrate systems. All summations except for the solvation term pertain to pairs of ligand atoms, \( i \), and protein atoms, \( j \), along with non-bonded ligand atoms separated by at least three bonds. Solvation summation is performed over ligand carbon atoms, \( i \), and protein atoms, \( j \). The five components are the van der Waals, hydrogen-bonding, and electrostatic interactions, the six global rotational and translational torsions as well as internal torsions of heavy atom ligand pendant groups, and the change in solvation energy due to ligand binding and the hydrophobic effect.

The van der Waals term is a measure of the dispersion/repulsion forces, and is calculated with Lennard-Jones 12-6 parameters \( (A_{ij} \text{ and } B_{ij}) \); the hydrogen-bonding term is calculated by directional 12-10 terms \( (C_{ij} \text{ and } D_{ij}) \), where \( E(t) \) is a directional weight based on the angle, \( t \), between the probe atom of the protein grid and the ligand atom and \( E_{hbond} \) models the average hydrogen-bonding energy of polar atoms. Different parameters for metal atoms and Gibbs free energy function weighting parameters from the default values were developed by the Reilly group to accurately simulate carbohydrate interactions. Electrostatic contributions come from screened Coulombic electrostatic potentials, where \( \varepsilon(r_{ij}) \) is a sigmoidal distance-dependent dielectric function and \( q_i \) and \( q_j \) are atomic charges. \( N_{tor} \) represents the number of \( sp^3 \) bonds in the ligand and is used to account for the entropic effects that occur upon ligand binding, due to restriction of conformational degrees of freedom.

Desolvation effects are modeled after a pairwise, volume-based method that can utilize AutoDock’s grid for protein-ligand interactions, and they were split between the hydrogen-bonding term and solvation term, to account for polar and non-polar desolvation effects, respectively. The application is based on Hess's law of heat summation, which states that the change in free energy between two states is path independent. This relationship is expressed in the following equation:

\[
\Delta G_{\text{binding,solution}} = \Delta G_{\text{binding,vacuo}} + \Delta G_{\text{solvation}(E\!\!\!\!\!\!\text{I})} + \Delta G_{\text{solvation}(E\!\!\!\!\!\!\text{II})}
\] (3)
whose terms are illustrated in Figure 2.3. AutoDock can calculate $\Delta G_{\text{binding,vacuo}}$ and also estimate the free energy change upon solvation for both bound and unbound complexes, thus finding the contribution to the overall free energy due to desolvation. Calculations are limited to aliphatic and aromatic carbon atoms in the ligand because nitrogen and oxygen atoms have a negligible effect on desolvation energies. Ligand internal energy is neglected in the free energy calculation because there is no noticeable difference when it is included, leading to the assumption that the ligand has equivalent internal energies both in complex and in solution.

AutoDock's grid map feature helps speed calculations of protein-ligand interactions by generating a look-up table, in the form of protein atoms as a three-dimensional box of regularly spaced points functioning as probe atoms storing the potential energy of all atoms in the macromolecule (Figure 2.4). Each probe atom's energy is based upon the parameters given for a specific atom type and encompasses all the protein atoms within a non-bonded
cutoff radius of the interacting ligand atom. Individual grid maps are generated for each atom type in the docked ligand, along with a grid map for electrostatic calculations.

The Lamarckian Genetic Algorithm (LGA) is a search method used to sample protein-ligand interaction energies, providing a more efficient search over previous algorithms used in AutoDock such as Monte Carlo or simulated annealing. Genetic algorithms (GAs) use genetics theory and nomenclature to implement a search scheme designed to locate the global minimum on (in this case) the energy landscape of molecular interactions. Just as in the ecological domain, several genetic events occur while increasing a population's fitness. An enzyme-ligand complex has a set of characterizing state variables, or 'genes', which the GA manipulates: the orientation, which is a quaternion defining the ligand's rotation (the roll, pitch, and yaw, similar to Euler angles); the translation, representing the locational change in $xyz$ space; and ligand torsions, one for each freely rotatable heavy atom bond (Figure 2.5). The sum of genes

![Figure 2.4](image1.png)

**Figure 2.4.** Gridmap simulation of enzyme-ligand interactions in AutoDock 3.0. Energy calculations are based on interactions between the probe atom, which represents all enzyme atoms, and the ligand. Calculations are performed for each ligand atom type and for electrostatic interactions.

![Figure 2.5](image2.png)

**Figure 2.5.** Variable genes of a ligand in AutoDock. A ligand’s genotype is characterized by its orientation (a quaternion consisting of a unit vector of roll, pitch, and yaw, plus a magnitude), translation in $xyz$ coordinate space, and rotatable bonds between heavy atom pendant groups. Cellobiose is shown as an example.
makes up a ligand's genotype, whereas the phenotype is the ligand's atomic coordinates and fitness. AutoDock calculates a ligand's fitness as the sum of the intermolecular interactions between the enzyme and ligand plus the intramolecular 1,3 and larger interactions between ligand atoms. Crossover occurs when a set of two ligands (the 'parents') exchange genes to produce offspring (a 'child'), while mutations are the random change from one gene to another. Attainment of the global docked minimum, in theory, is achieved through selection of the fittest population members for survival to subsequent generations.15

The LGA's increased efficiency stems from its ability to incorporate a local search into the global search of the conformational energy landscape typical of most GAs.15 Simulated annealing (SA) uses a local search somewhat similar to the LGA; SA's use of local searching is more prevalent as the system temperature decreases because larger moves up the energy landscape become less probabilistic. The drawback of SA is that there is a greater likelihood of becoming trapped in a local minimum as temperature decreases. LGA addresses this issue by using the Solis and Wets (SW) local search for each generation during a docking run. The uniqueness of LGA is based upon the incorrect evolutionary model of Lamarck, who stated that phenotypic traits acquired by an organism can be passed on to the genotype of its offspring. In the context of AutoDock, the LGA employs inverse developmental mapping. However, most GAs are based on Darwinian evolution and Mendelian genetics, where developmental mapping occurs from the genotype to the phenotype (Figure 2.6). Local searches in a GA can occur only if an inverse mapping function

Figure 2.6. Comparison of the Lamarckian genetic algorithm (LGA) and standard GAs. The fitness function, or total docked energy of an enzyme-ligand interaction, is shown as f(x). Genetic mutation is the most effective way for standard GAs to randomly sample the energy landscape, due to the dependence of a standard GA on developmental mapping of the genotype to the phenotype. Local searches of the phenotypic space allows the LGA to arrive at a lower local minimum before passing on the parent's genes to the offspring because of inverse mapping abilities.15
exists, due to the fact that fitness improvement from generation to generation is passed on through the genotype and not the phenotype. Thus, the LGA can make a more efficient search of the conformational space than can a normal GA, with the added advantage that the SW local search is performed in genotypic space, eliminating the need for an inverse mapping function.

**Methodology.** AutoDock's LGA search loops through a user-defined number of generations and energy evaluations to locate the global minimum of the protein-ligand complex being studied. Each iteration creates a specified population of random individuals. These individuals are then mapped from their genotypes to phenotypes and fitnesses are calculated, based on equation (2). Fitness evaluation allows for selection of individuals to produce offspring; the amount of offspring is proportional to fitness. Two-point crossover is the method used to produce two offspring from parents, and these offspring replace the parents to keep the population constant. Once this is accomplished, mutation occurs in a user-defined percentage of the population, and is calculated by adding a randomly generated number with a Cauchy distribution to one of the individual's state variables. Elitism can be used to ensure a given number of the fittest individuals survive to the next generation, and is implemented into selection after crossover and mutation take place. AutoDock ranks conformational clusters once the maximum number of generations or energy evaluations is reached.\textsuperscript{15}

**Molecular dynamics**

Molecular dynamics (MD) is the computational simulation method that studies the thermodynamic interactions between molecules by analyzing Newton’s laws of motion, as follows:\textsuperscript{20}

1. A body retains its inertia unless acted upon by some net force.

2. $\mathbf{F} = ma$, or the force on an object equals the rate change of momentum, where $\mathbf{F}$ is force, $m$ is mass, and $a$ is the acceleration.

3. For every action, there is an equal and opposite reaction.

What results in MD is a time-dependent trajectory of Newton’s second law of motion, as described by the differential equation:
\[
\frac{d^2 x_i}{dt^2} = \frac{F_{x_i}}{m_i}
\] (4)

The right-hand side of the equation describes the force on the particle \(m_i\) along a coordinate \(x_i\). When examining the force on a particle dependent on interactions with other particles, analytical calculations become impossible, due to the coupling of all particle motions (this is called the **continuous potential** model). The continuous potential model of intermolecular interactions solves Newton’s laws of motion through finite difference methods, in which integration of atomic positions and dynamical properties is broken down into a finite number of timesteps, with each step defined by \(\delta t\). Expression of these variables is embodied in a Taylor series expansion, as shown for the position vector, \(\mathbf{r}\):

\[
\mathbf{r}(t + \delta t) = \mathbf{r}(t) + \delta t \mathbf{v}(t) + \frac{1}{2} \delta t^2 \mathbf{a}(t) + \frac{1}{6} \delta t^3 \mathbf{b}(t) + \ldots
\] (5)

where \(\mathbf{v}\) represents the first time derivative of the position vector, or velocity, \(\mathbf{a}\) represents acceleration, \(\mathbf{b}\) represents the position vector’s third time derivative, and so on. Many algorithms have been developed to calculate the various properties in an MD simulation, but all of them truncate the Taylor series to some degree.\(^\text{20}\)

Ensembles are used to specify properties held constant in MD simulations, and are chosen according to the phenomena that are desired to be examined. The microcanonical ensemble, or NVE (for constant moles (N), volume (V), and energy (E)), is synonymous with an adiabatic process with no heat exchange, and is traditionally used in MD. Two other commonly used ensembles in MD are the NVT and NPT ensembles. NVT, or canonical, ensembles hold the volume (V) and temperature (T) constant, and are useful for examining things such as protein unfolding processes. Temperature is held constant through the use of a heat bath, such as the Nosé-Hoover thermostat. The NPT (isobaric-isothermal) ensemble is often used for comparison to experimental data, since experiments are often done at ambient temperature and pressure.\(^\text{21}\)

**References**

harbor Laboratory Press, Cold Spring Harbor, NY.


Chapter 3. Phylogenetic Analysis of Family 6 Glycoside Hydrolases

A paper published in Biopolymers\(^1\)

Blake Mertz,\(^2,3\) Robert S. Kuczenski,\(^2,4\) Robert T. Larsen,\(^2,4\) Anthony D. Hill,\(^2,5\) Peter J. Reilly\(^2,6,*\)\(^∀\)

Abstract

Multiple sequence alignment separates members of glycoside hydrolase Family 6 into eight subfamilies: one of mainly actinobacterial endoglucanases (EGs), one of ascomycotal EGs, one of chytridiomycotal EGs and cellobiohydrolases (CBHs), one of actinobacterial and proteobacterial CBHs, one of chytridiomycotal CBHs, two of ascomycotal CBHs, and one of basidiomycotal CBHs. Each also has some proteins of unknown function. Multiple sequence alignment also extends to all of Family 6 the observation that lengths of loops that can form the active-site tunnel in CBHs vary among subfamilies and, along with loop conformations, determine enzyme function.

Keywords: cellobiohydrolase, endoglucanase, glycoside hydrolase family 6, multiple sequence alignment, phylogenetic tree

Introduction

Glycoside hydrolase Family 6 (GH6),\(^1,2\) formerly cellulase Family B, comprises many cellobiohydrolases (β-1,4-glucan cellobiohydrolases, EC3.2.1.91, CBHs, often called CBH

\(^1\) Reprinted with permission from Biopolymers, Vol. 79, 197-206 (2005)
\(^2\) Department of Chemical and Biological Engineering, Iowa State University, Ames, IA 50011
\(^3\) Primary researcher and author.
\(^4\) Performed preliminary phylogenetic analysis of a smaller subset of GH6.
\(^5\) Consulted for methodology and accuracy of analysis.
\(^6\) Associate Professor.
\(^\ast\) Correspondence to: Peter J. Reilly, Department of Chemical and Biological Engineering, 2114 Sweeney Hall, Iowa State University, Ames, IA 50011-2230. Phone: 515-294-5968; fax: 515-294-2689; E-mail: reilly@iastate.edu
\(^∀\) Contract grant sponsor and numbers: U. S. Department of Agriculture through the Biotechnology Byproducts Consortium, 2004-34188-15067; National Science Foundation, BES-0313878.
II’s to differentiate them from the CBH I’s in GH7) and endoglucanases (β-1,4-glucan glucanohydrolases, EC3.2.1.4, EGs). CBHs cleave the ends of cellulose and cellooligosaccharide chains to yield mostly cellobiose, while EGs produce a range of cellooligosaccharides from the same substrates, attacking throughout their chains.

There must be significant differences in tertiary structures for GH6 family members to yield different product patterns from the same substrates. In fact, catalytic domain tertiary structures are known for five GH6 enzymes: *Hypocrea jecorina* (formerly *Trichoderma reesei*) QM9414 CBH, \(^3\) \(^4\) \(^5\) \(^6\) *Thermobifida* (formerly *Thermomonospora*) *fusca* E2 EG, \(^7\) *Humicola insolens* CBH, \(^8\) \(^9\) \(^10\) which is very similar to *H. jecorina* CBH, *H. insolens* EG, \(^11\) and *Mycobacterium tuberculosis* *H37Rv* EG. \(^12\) The GH6 catalytic domain is a distorted (β/α)\(_8\) barrel, with the active site being formed by a sandwich of seven parallel β-strands connected by five α-helices and one loop (Figure 3.1). CBH active sites are mainly enclosed in approximately 20-Å long, 4-Å diameter tunnels formed by two loops, \(^3\) \(^8\) one from each side, connected to each other by a single hydrogen bond. \(^8\) EGs are reported to occur when amino acid residue deletions shorten the C-terminal loop and bending of the N-terminal loop pulls it away from the tunnel roofs found in CBHs. \(^7\) \(^11\) A cleft rather than a tunnel results, allowing cellulose chains to enter the active site at any point rather than from one end. The role of the tunnel in conferring exo-type specificity on CBHs was confirmed by deleting the C-terminal loop of *Cellulomonas fimi* CBH, increasing its EG activity. \(^13\)

To aid further classification and study of the GH6 family, we constructed a multiple sequence alignment (MSA) and a phylogenetic tree of its members. This will allow researchers to more accurately infer the properties of closely related enzymes while cautiously inferring properties of more distantly related ones. The MSA also allows assessment of the conservation of various proposed catalytic bases and of the effect of lengths of the loops enclosing the active site on enzyme action patterns. Tsai *et al.* performed a limited MSA and phylogenetic analysis of GH6 enzymes. \(^14\) Their phylogenetic tree contained 30 enzymes and delineated four major subfamilies with one outlier. Varrot *et al.* presented a phylogenetic tree of 77 enzymes, of which 20 were identified, grouped into four subfamilies with two outliers. \(^12\) Until this article these were the only two GH6 phylogenetic studies in the open literature.
Figure 3.1. Tertiary structures of (a) M. tuberculosis H37Rv EG (Subfamily 1), (b) T. fusca EG (Subfamily 1), (c) H. insolens EG (Subfamily 2), (d) H. jecorina QM9414 CBH (Subfamily 8), and (e) H. insolens CBH (Subfamily 8). N-terminal (violet) and C-terminal (blue) loops enclosing the active-site cleft in CBHs are based on multiple sequence alignments using the assignments of residues that initiate and terminate the loops in H. jecorina CBH.
**Computational Methods**

GH6 amino acid sequences were collected from the Swiss-Prot\textsuperscript{15} and GenBank\textsuperscript{16} databases using the CAZy database\textsuperscript{1,2}.

An MSA of the catalytic domains of all the family members was constructed and optimized with ClustalX 1.81\textsuperscript{17} using the Gonnet 250 substitution matrix.\textsuperscript{18} Pairwise gap opening, multiple alignment gap opening, and multiple alignment divergent sequence penalty parameters were set at 30, 30, and 30\%, respectively.

A phylogenetic tree was constructed with PHYLIP 3.62 beta,\textsuperscript{19} allowing GH6 to be further divided into subfamilies. The MSA was bootstrapped using SEQBOOT, with molecular sequence and bootstrap chosen, block size = 1, replicates = 250, and input sequences interleaved. PROTPARS was used, with the best tree searched for, the sequence input order randomized, 250 multiple data sets analyzed, and input sequences interleaved. A consensus tree was found using CONSENSE, with majority rule extended for consensus type.

PROTDIST uses the sequences from MSAs to generate branch lengths, utilizing the JTT substitution matrix.\textsuperscript{20} Input sequences were interleaved; the MSA produced in ClustalX was used to calculate the distances between all 66 GH6 catalytic domain sequences. Putative amino acid sequences of the nodes were computed by PROTPARS. The internal node and catalytic domain sequences formed the input for PROTDIST, which generated distances used as branch lengths in the phylogenetic tree. Output distances were given in units of 100 PAMs (accepted point mutations), representing the amino acid mutations accepted by natural selection. Subfamilies were identified by qualitative inspection of the tree. Mean distances within and between subfamilies were found by averaging the given data set for each group, and standard deviations (\(\sigma\)) were calculated.

A z-test was performed to confirm the validity of the subfamily groupings. Variances (\(\sigma^2\)) were pooled into four groups; those of subfamilies 1, 3, and 4 were pooled individually because their \(\sigma\) values were 2–10 times larger than those of the other five subfamilies, which were pooled together into one average variance. The equation for calculating z values is...
\[ z = \frac{\bar{x}_{ij} - (\bar{x}_{ii} + \bar{x}_{jj})/2}{\sqrt{\frac{\sigma_i^2}{n_{ij}} + \frac{\sigma_{ii,jj}^2}{n_{ii,jj}}}} \]  

where \( \bar{x}_{ii}, \bar{x}_{ij}, \) and \( \bar{x}_{jj} \) are the inter- and intra-subfamily distance means calculated from PROTDIST results; \( n_{ii}, n_{ij}, \) and \( n_{jj} \) are the total number of data points for each \( \bar{x} \) value; and \( \sigma_i^2 \) and \( \sigma_{ii,jj}^2 \) are the pooled intra- and inter-subfamily \( \sigma^2 \) values. The second term in the denominator had different forms for specific subfamily comparisons:

\[ \frac{\sigma_{ii,jj}^2}{n_{ii,jj}} = \frac{\sigma_i^2}{n_{ii}} + \frac{\sigma_{jj}^2}{n_{jj}} \quad \text{(for} \sigma^2 \text{ pools between subfamilies 1, 3, and 4)} \]

\[ \frac{\sigma_{ii,jj}^2}{n_{ii,jj}} = \frac{\sigma_{avg}^2}{n_{ii}} + \frac{\sigma_{avg}^2}{n_{jj}} \quad \text{(for} \sigma^2 \text{ pools between subfamilies 2, 5, 6, 7, and 8)} \]

MSAs were then computed for members of the individual subfamilies, again using the Gonnet 250 substitution matrix. Pairwise gap opening, multiple alignment gap opening, and multiple alignment divergent sequence penalty parameters were 30, 30, and 30%, respectively, for Subfamilies 2, 3, 4, 6, and 8; 30, 30, and 75% for Subfamily 1; 10, 25, and 75% for Subfamily 5; and 10, 10, and 75% for Subfamily 7.

**Results and Discussion**

**Phylogenetic Analysis.** We aligned 66 proteins and further defined them phylogenetically. An MSA of the catalytic domains of GH6 family members yields eight subfamilies (Table 3.1, plus Supplemental Material). This is a significant expansion over the four subfamilies found earlier.\textsuperscript{12,14}

Mean distances between different subfamilies were much larger than within subfamilies, and values of \( z \) ranged from 6 to 34 (Table 3.2), giving \( p \)-values of < 0.0001. This clearly demonstrates that all inter-subfamily distances are significantly higher than all intra-subfamily distances.

An unrooted phylogenetic tree (Figure 3.2) was constructed from the MSA. The diversity of the GH6 family is readily apparent.
Table 3.1. Summary of the Members and Properties of Each Subfamily.

<table>
<thead>
<tr>
<th>Producing Organism</th>
<th>Class</th>
<th>Type</th>
<th>Location</th>
<th>GenBank Number</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Subfamily 1</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1. <em>Cellulomonas fimi</em> #2</td>
<td>Ac</td>
<td>EG</td>
<td>N</td>
<td>2</td>
</tr>
<tr>
<td>2. <em>Eremothecium gossypii</em></td>
<td>A</td>
<td>UP</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>3. <em>Micromonospora cellulosoliticum</em></td>
<td>Ac</td>
<td>EG</td>
<td>C</td>
<td>2</td>
</tr>
<tr>
<td>4. <em>Mycobacterium bovis</em> AF2122/97</td>
<td>Ac</td>
<td>UP</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>5. <em>Mycobacterium tuberculosis</em> CDC1551</td>
<td>Ac</td>
<td>UP</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>6. <em>Mycobacterium tuberculosis</em> H37Rv&lt;sup&gt;b&lt;/sup&gt;</td>
<td>Ac</td>
<td>EG</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>7. <em>Mycobacterium tuberculosis</em> MA-4680</td>
<td>P</td>
<td>EG</td>
<td>N</td>
<td>4_9</td>
</tr>
<tr>
<td>8. <em>Streptomyces avermitilis</em> MA-4680</td>
<td>Ac</td>
<td>UP</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>9. <em>Streptomyces avermitilis MA-4680</em></td>
<td>Ac</td>
<td>UP</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>10. <em>Streptomyces coelicolor</em> A3(2)</td>
<td>Ac</td>
<td>UP</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>11. <em>Streptomyces coelicolor</em> A3(2)</td>
<td>Ac</td>
<td>UP</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>12. <em>Streptomyces halstedii</em></td>
<td>Ac</td>
<td>EG</td>
<td>C</td>
<td>2</td>
</tr>
<tr>
<td>13. <em>Streptomyces nanchangensis</em> NS3226</td>
<td>Ac</td>
<td>UP</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>14. <em>Streptomyces sp. KSM-9</em></td>
<td>Ac</td>
<td>EG</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>15. <em>Thermobifida fusca</em>&lt;sup&gt;b&lt;/sup&gt;</td>
<td>Ac</td>
<td>EG</td>
<td>C</td>
<td>2</td>
</tr>
<tr>
<td>16. <em>Thermobispora bispora</em></td>
<td>Ac</td>
<td>EG</td>
<td>C</td>
<td>2</td>
</tr>
<tr>
<td>17. <em>Xylanimicrobium pachnodae</em></td>
<td>Ac</td>
<td>EG</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td><strong>Subfamily 2</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1. <em>Humicola insolens</em>&lt;sup&gt;b&lt;/sup&gt;</td>
<td>A</td>
<td>EG</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>2. <em>Magnaporthe grisea</em> 70-15</td>
<td>A</td>
<td>UP</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>3. <em>Neurospora crassa</em> OR74A</td>
<td>A</td>
<td>UP</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td><strong>Subfamily 3</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1. <em>Neocallimastix patriciarum</em></td>
<td>Ch</td>
<td>CBH</td>
<td>N</td>
<td>1</td>
</tr>
<tr>
<td>2. <em>Orpinomyces sp. PC-2</em></td>
<td>Ch</td>
<td>EG/CBH</td>
<td>N</td>
<td>D2/D2</td>
</tr>
<tr>
<td>3. <em>Orpinomyces sp. PC-2</em></td>
<td>Ch</td>
<td>EG/CBH</td>
<td>N</td>
<td>D2/D2</td>
</tr>
<tr>
<td>4. <em>Orpinomyces sp. PC-2</em></td>
<td>Ch</td>
<td>EG/CBH</td>
<td>N</td>
<td>1</td>
</tr>
<tr>
<td>5. <em>Orpinomyces sp. PC-2</em></td>
<td>Ch</td>
<td>CBH</td>
<td>N</td>
<td>D2/D2</td>
</tr>
<tr>
<td>6. <em>Piromyces rhizinflatus</em> 2301</td>
<td>Ch</td>
<td>EG</td>
<td>N</td>
<td>D2/D2</td>
</tr>
<tr>
<td>7. <em>Piromyces rhizinflatus</em> 2301</td>
<td>Ch</td>
<td>CBH</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>8. <em>Piromyces rhizinflatus</em> 2301</td>
<td>Ch</td>
<td>CBH</td>
<td>C</td>
<td>D2/D2</td>
</tr>
<tr>
<td>9. <em>Piromyces rhizinflatus</em> 2301</td>
<td>Ch</td>
<td>CBH</td>
<td>N</td>
<td>D2/D2</td>
</tr>
<tr>
<td>10. <em>Piromyces sp. KS11</em></td>
<td>Ch</td>
<td>UP</td>
<td>C</td>
<td>D2/D2</td>
</tr>
<tr>
<td><strong>Subfamily 4</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1. <em>Cellulomonas fimi</em> #1</td>
<td>Ac</td>
<td>CBH</td>
<td>C</td>
<td>2</td>
</tr>
<tr>
<td>2. <em>Ralstonia solanacearum</em> GMI1000</td>
<td>P</td>
<td>UP</td>
<td>C</td>
<td>2</td>
</tr>
<tr>
<td>3. <em>Streptomyces avermitilis</em> SAV1853</td>
<td>Ac</td>
<td>UP</td>
<td>N</td>
<td>2</td>
</tr>
<tr>
<td>4. <em>Streptomyces avermitilis</em> SAV557</td>
<td>Ac</td>
<td>UP</td>
<td>C</td>
<td>3</td>
</tr>
<tr>
<td>5. <em>Streptomyces coelicolor</em> A3(2)</td>
<td>Ac</td>
<td>UP</td>
<td>N</td>
<td>2</td>
</tr>
<tr>
<td>6. <em>Streptomyces sp. M23</em></td>
<td>Ac</td>
<td>CBH</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>7. <em>Thermobifida fusca</em></td>
<td>Ac</td>
<td>CBH</td>
<td>N</td>
<td>2</td>
</tr>
<tr>
<td>8. <em>Thermobifida fusca</em></td>
<td>Ac</td>
<td>UP</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>9. <em>Xanthomonas campestris</em> ATCC33913</td>
<td>P</td>
<td>UP</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>10. <em>Xanthomonas campestris</em> ATCC33913</td>
<td>P</td>
<td>UP</td>
<td>—</td>
<td>—</td>
</tr>
</tbody>
</table>

(continued)
### Table 3.1. (continued)

<table>
<thead>
<tr>
<th>Producing organism</th>
<th>Class</th>
<th>Type</th>
<th>CBM Location</th>
<th>CBM Family</th>
<th>Swiss-Prot or GenBank Number</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>11. Xylella fastidiosa 9a5c</strong></td>
<td>P</td>
<td>CBH</td>
<td>C</td>
<td></td>
<td>Q9PDW2</td>
</tr>
<tr>
<td><strong>12. Xylella fastidiosa Temecula 1</strong></td>
<td>P</td>
<td>UP</td>
<td>C</td>
<td></td>
<td>Q87E00</td>
</tr>
<tr>
<td><strong>Subfamily 5</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>1. Orpinomyces sp. PC-2</strong></td>
<td>Ch</td>
<td>UP</td>
<td>N</td>
<td>D2/D2</td>
<td>Q96V98</td>
</tr>
<tr>
<td><strong>2. Orpinomyces sp. PC-2</strong></td>
<td>Ch</td>
<td>UP</td>
<td>N</td>
<td>D2/D2</td>
<td>Q96V97</td>
</tr>
<tr>
<td><strong>3. Piromyces equi</strong></td>
<td>Ch</td>
<td>CBH</td>
<td>N</td>
<td>D2/D2</td>
<td>Q7Z7X6</td>
</tr>
<tr>
<td><strong>4. Piromyces sp. E2</strong></td>
<td>Ch</td>
<td>CBH</td>
<td>N</td>
<td>D2/D2</td>
<td>Q874K3</td>
</tr>
<tr>
<td><strong>Subfamily 6</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>1. Magnaporthe grisea 70-15</strong></td>
<td>A</td>
<td>UP</td>
<td>—</td>
<td>—</td>
<td>XM_362054</td>
</tr>
<tr>
<td><strong>2. Neurospora crassa OR74A</strong></td>
<td>A</td>
<td>UP</td>
<td>—</td>
<td>—</td>
<td>Q7RXI7</td>
</tr>
<tr>
<td><strong>Subfamily 7</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>1. Agaricus bisporus</strong></td>
<td>B</td>
<td>CBH</td>
<td>N</td>
<td>1</td>
<td>P49075</td>
</tr>
<tr>
<td><strong>2. C. oriolus versicolor</strong></td>
<td>B</td>
<td>UP</td>
<td>N</td>
<td>1</td>
<td>AF233583</td>
</tr>
<tr>
<td><strong>3. Lentinula edodes L54</strong></td>
<td>B</td>
<td>CBH</td>
<td>N</td>
<td>1</td>
<td>Q9C1R4</td>
</tr>
<tr>
<td><strong>4. Lentinula edodes Stamets CS-2</strong></td>
<td>B</td>
<td>UP</td>
<td>N</td>
<td>1</td>
<td>Q96VU2</td>
</tr>
<tr>
<td><strong>5. Phanerochaete chrysosporium</strong></td>
<td>B</td>
<td>CBH</td>
<td>N</td>
<td>1</td>
<td>Q02321</td>
</tr>
<tr>
<td><strong>6. Pleurotus sajor-caju</strong></td>
<td>B</td>
<td>CBH</td>
<td>N</td>
<td>1</td>
<td>Q96TP4</td>
</tr>
<tr>
<td><strong>7. Vovariella volvacea V14</strong></td>
<td>B</td>
<td>CBH</td>
<td>N</td>
<td>1</td>
<td>Q9Y894</td>
</tr>
<tr>
<td><strong>Subfamily 8</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>1. Acremonium cellulolyticus Y-94</strong></td>
<td>A</td>
<td>CBH</td>
<td>N</td>
<td>1</td>
<td>Q93837</td>
</tr>
<tr>
<td><strong>2. Aspergillus nidulans FGSC A4</strong></td>
<td>A</td>
<td>UP</td>
<td>—</td>
<td>—</td>
<td>AACD01000017</td>
</tr>
<tr>
<td><strong>3. Aspergillus nidulans FGSC A4</strong></td>
<td>A</td>
<td>UP</td>
<td>N</td>
<td>1</td>
<td>AACD01000093</td>
</tr>
<tr>
<td><strong>4. Cochliobolus heterostrophus C4</strong></td>
<td>A</td>
<td>CBH</td>
<td>—</td>
<td>—</td>
<td>Q8J215</td>
</tr>
<tr>
<td><strong>5. Fusarium oxysporum</strong></td>
<td>A</td>
<td>EG</td>
<td>N</td>
<td>1</td>
<td>P46236</td>
</tr>
<tr>
<td><strong>6. Humicola insolens</strong></td>
<td>A</td>
<td>UP</td>
<td>N</td>
<td>1</td>
<td>Q9C1S9</td>
</tr>
<tr>
<td><strong>7. Humicola insolens</strong></td>
<td>A</td>
<td>CBH</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td><strong>8. Hypocrea jecorina QM9414</strong></td>
<td>A</td>
<td>CBH</td>
<td>N</td>
<td>1</td>
<td>P07987</td>
</tr>
<tr>
<td><strong>9. Hypocrea koningii</strong></td>
<td>A</td>
<td>CBH</td>
<td>N</td>
<td>1</td>
<td>AF315681</td>
</tr>
<tr>
<td><strong>10. Magnaporthe grisea 70-15</strong></td>
<td>A</td>
<td>UP</td>
<td>N</td>
<td>1</td>
<td>XM_360146</td>
</tr>
<tr>
<td><strong>11. Neurospora crassa OR74A</strong></td>
<td>A</td>
<td>UP</td>
<td>N</td>
<td>1</td>
<td>Q872J7</td>
</tr>
<tr>
<td><strong>12. Talaromyces emersonii</strong></td>
<td>A</td>
<td>CBH</td>
<td>N</td>
<td>1</td>
<td>Q8NIB5</td>
</tr>
<tr>
<td><strong>13. Trichoderma viride CICC 13038</strong></td>
<td>A</td>
<td>CBH</td>
<td>N</td>
<td>1</td>
<td>AY368688</td>
</tr>
</tbody>
</table>

*Abbreviations: A: ascomycote; Ac: actinobacterium; C: C-terminal; Cel6ACBH: cellobiohydrolase II; Ch: chytridiomycote; D2: Dockerin type 2; EG: Endoglucanase; N: N-terminal; P: proteobacterium; UP: unknown protein.; —: missing or unknown.

*Tertiary structure available.

Subfamily 1 contains eight actinobacterial EGs, one proteobacterial EG, seven actinobacterial hypothetical proteins or proteins of unknown function, hereafter called unknown proteins (UPs), and one ascomycotal UP. The subfamily is loosely grouped, but with almost identical proteins from *Eremothecium gossypii* (since eliminated from the CAZy database) and *Streptomyces halstedii*, the only ascomycotal fungus and an actinobacterium, respective-
ly, and from *Mycobacterium bovis subsp. bovis* and two strains of *M. tuberculosis*, all actinobacteria. Five of these enzymes have carbohydrate-binding modules (CBMs), two N-terminal and three C-terminal. Four actinobacterial CBMs are from CBM Family 2 (CBM2), \(^1,^2\) almost unique to bacterial enzymes, and are linked to their catalytic domains by proline-serine/threonine boxes, while the proteobacterial *Myxococcus xanthus* has a CBM4_9. Two of the five known GH6 tertiary structures are of *T. fusca* EG\(^7\) and *M. tuberculosis* EG, \(^12\) both Subfamily 1 members.

Subfamily 2 has one ascomycotal EG and two ascomycotal UPs. The EG, from *Humicola insolens*, has a known tertiary structure. \(^11\) None has a CBM.

### Table 3.2. Mean JTT Distances and z-values (Italicized) within and between Different Subfamilies.

<table>
<thead>
<tr>
<th>Subfamilies</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.11 ± 0.28(^a)</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>2</td>
<td>1.87 ± 0.15</td>
<td>0.66 ± 0.09</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>3</td>
<td>1.72 ± 0.15</td>
<td>1.42 ± 0.07</td>
<td>0.63 ± 0.29</td>
<td>16.4</td>
<td>11.7</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>4</td>
<td>1.84 ± 0.17</td>
<td>1.48 ± 0.10</td>
<td>1.44 ± 0.08</td>
<td>0.64 ± 0.23</td>
<td>24.3</td>
<td>14.4</td>
<td>15</td>
<td>—</td>
</tr>
<tr>
<td>5</td>
<td>2.00 ± 0.16</td>
<td>1.48 ± 0.06</td>
<td>1.35 ± 0.06</td>
<td>1.17 ± 0.09</td>
<td>0.19 ± 0.07</td>
<td>31.2</td>
<td>18.9</td>
<td>16.7</td>
</tr>
<tr>
<td>6</td>
<td>1.96 ± 0.22</td>
<td>1.17 ± 0.07</td>
<td>1.34 ± 0.09</td>
<td>1.37 ± 0.06</td>
<td>1.02 ± 0.02</td>
<td>27.0 ± 0.00</td>
<td>25.3</td>
<td>12.9</td>
</tr>
<tr>
<td>7</td>
<td>1.84 ± 0.16</td>
<td>1.15 ± 0.07</td>
<td>1.37 ± 0.10</td>
<td>1.28 ± 0.07</td>
<td>1.14 ± 0.05</td>
<td>0.88 ± 0.05</td>
<td>0.36 ± 0.12</td>
<td>34.7</td>
</tr>
<tr>
<td>8</td>
<td>1.86 ± 0.19</td>
<td>1.18 ± 0.06</td>
<td>1.34 ± 0.09</td>
<td>1.31 ± 0.06</td>
<td>1.09 ± 0.04</td>
<td>0.85 ± 0.05</td>
<td>0.69 ± 0.09</td>
<td>0.44 ± 0.12</td>
</tr>
</tbody>
</table>

\(^a\) Standard deviation.

Subfamily 3 is reported in articles \(^14,^21,^22,^23\) and by the CAZy database to have five chytridiomycotal CBHs, three chytridiomycotal EGs/CBHs, one chytridiomycotal EG, and one chytridiomycotal UP. Homology among the members is quite high except for one *Piromyces rhizinflatus* CBH. CBHs from *Neocallimastix patriciarum* and *Orpinomyces sp. PC-2* are the only two Subfamily 3 members to have CBM1s rather than dockerins, while a *Piromyces rhizinflatus* 2301 CBH and a *Piromyces sp. KS11* UP are the only two having dockerins located at their C-termini instead of dockerins or at their N-termini. The linkers between catalytic domains and CBMs or dockerins are high in serine and threonine residues,
although *N. patriciarum* CBH and an *Orpinomyces* sp. EG/CBH (Q874D8) have Asn-rich linkers (Supplemental Material).

Subfamily 4 contains three actinobacterial CBHs, four actinobacterial UPs, one proteobacterial CBH, and four proteobacterial UPs. Seven of the twelve subfamily members have CBM2s and one has a CBM3. Three of these are to the N-terminal side of the catalytic domain and five are to the C-terminal side, all linked to it by proline-serine/threonine boxes.

Subfamily 5 has two chytridiomycotal CBHs and two chytridiomycotal UPs, each with double N-terminal dockerins and linked to it with proline-serine/threonine boxes.

Subfamily 6 consists of two ascomycotal UPs, neither with a CBM.

Subfamily 7 contains five basidiomycotal CBHs and two basidiomycotal UPs. The least homologous member of this subfamily is *Vovariella volvacea* CBH, which may have been mis sequenced and whose sequence was not included in calculations of branch distances. Two *Lentinula edodes* CBHs are essentially identical. All Subfamily 7 members have CBM1s, which are almost exclusively found in fungi. They are located to the N-terminal side of the catalytic domain and are connected to it by serine- and threonine-rich linkers of varying lengths.

Subfamily 8 contains seven ascomycotal CBHs, one ascomycotal EG, and five
ascomycotal UPs, of which CelAs from *H. jecorina, Hypocrea koningii*, and *Trichoderma viride* are essentially identical. An *H. insolens* CBH has no GenBank or Swiss-Prot sequence, but may be classified in Subfamily 8 because its tertiary structure is so close to that of *H. jecorina* CBH (Figure 3.1). Ten members have N-terminal CBM1s attached to their catalytic domains with serine-and threonine-rich linkers.

Figure 3.3 shows an MSA of ten catalytic domains, including single EGs from Subfamilies 1 and 2, a CBH and an EG/CBH from Subfamily 3, single CBHs or UPs from Subfamilies 4–7, and a CBH and EG from Subfamily 8. It shows the wide variation among different subfamilies.

Seven of the eight subfamilies appear to contain either EGs (Subfamilies 1 and 2) or CBHs (Subfamilies 4–8) except for a single Subfamily 8 EG, from *Fusarium oxysporum*, tentatively identified by a genomic study, but little different from other subfamily members (Figures 3.2 and 3.3, Supplemental Material). This presupposes that those enzymes now classified as UPs will eventually be assigned the same specificities as those in the same subfamily already classified, and that the two Subfamily 6 UPs are really CBHs, as are those in the adjacent subfamilies. The substantial similarities among Subfamily 3 members (Table 3.2, Figures 3.2, and 3.3, and Supplemental Material) suggest that, upon extensive testing with sufficiently sensitive assays, most or all its members will be classified as mixed EG/CBHs, as three already have.

The eight subfamilies identified here may be compared to the four subfamilies of Tsai *et al.* and Varrot *et al.*. Three of the first four fungal cellulases of Tsai *et al.* are part of Subfamily 7, while the other three belong to Subfamily 8. Of their first five bacterial cellulases, three are part of Subfamily 4 and one belongs to Subfamily 1. The rumen fungal enzyme outlier, from *P. rhizinflatus*, is part of Subfamily 3, as are all six grouped rumen fungal cellulases. Of their remaining eleven bacterial cellulases, eight belong to Subfamily 1 and one is part of Subfamily 4. They classified no members of Subfamilies 2, 5, or 6. Of the 21 members of bacterial Subfamily 2 of Varrot *et al.*, fifteen are part of our Subfamily 1 and one belongs to our Subfamily 4. They have not identified the members of their fungal Subfamilies 1 and 3 or their bacterial Subfamily 4, but a tentative comparison of numbers and placements suggests that their Subfamily 1 is our Subfamily 3, that their Subfamily 3 is
Figure 3.3. Multiple sequence alignment of catalytic domains of *T. fusca* EG (Subfamily 1), *H. insolens* EG (Subfamily 2), *N. patriciarum* CBH (Subfamily 3), *Orpinomyces* sp. PC-2 EG/CBH (Subfamily 3), *C. fimi* CBH (Subfamily 4), *Piromyces* sp. E2 CBH (Subfamily 5), *M. grisea* UP (Subfamily 6), *A. bisporus* CBH (Subfamily 7), *H. jecorina* QM9414 CBH (Subfamily 8), and *Fusarium oxysporum* EG (Subfamily 8). (*) Total conservation; (:) total conservation of similar residues; (l) partial conservation. Ruler positions: 104 and 121—beginning and end of N-terminal loop in *H. jecorina* CBH, respectively; 377 and 428—beginning and end of C-terminal loop in *H. jecorina* CBH, respectively; 107—Asp\(^175\); 156—Asp\(^221\); 202—Asp\(^263\); 385—Asp\(^401\).
our Subfamilies 5–8, and that their Subfamily 4 is our Subfamily 4. We have not classified their two metazoan outliers.

**Conservation of Putative Catalytic Residues.** The catalytic acid/proton donor, Asp$^{221}$, is conserved in all sequenced members of GH6 except for an unknown residue in *Piromyces sp. KS11* that would be Asp except for one unidentified nucleotide. Asp$^{175}$ and Asp$^{263}$ are conserved throughout, while Asp$^{401}$ is conserved in all but *V. volvacea* CBH, the least homologous member of Subfamily 7.$^{3,4,5,7}$

**Lengths of Tunnel-Forming Loops.** The suggestion that CBHs and EGs can be differentiated by the lengths of their N- and C-terminal loops, particularly the latter, that form tunnels in CBHs, as well as by the folding back of one or both loops in EGs,$^{7,11}$ can be tested by inspection of the overall MSA (Figure 3.3) and the MSAs of the eight subfamilies (Supplementary Material). The N-terminal loop in *H. jecorina* CBH extends from Pro$^{172}$ to Asp$^{189}$ (18 residues), and its C-terminal loop encompasses Val$^{394}$ to Ala$^{429}$ (36 residues).$^3$ Based on homology with this enzyme, the same numbers of loop residues are present in all other members of Subfamilies 5–8, except for the 37 residues in the C-terminal loop of *M. grisea* UP in Subfamily 8. Subfamily 4 N-terminal loops all have 17 residues, while their C-terminal loops have 49 to 52 residues. All Subfamily 3 members have 16 N-terminal loop residues except *Piromyces rhizinflatus 2301* CBH (O93680), which has 18. Their C-terminal loops consist of 34 residues, except for one member with 31 and one with 36. Subfamily 2 members have 18 N-terminal loop residues and 21 C-terminal loop residues. Members of Subfamily 1 have 15 or 16 residues in their N-terminal loops and 15–21 in their C-terminal loops, except for *X. pachnodae* EG, which has 32 in the latter.

Subfamilies 1 and 2 include all of the unquestioned EGs in GH6 and no CBHs. Their enzymes have not only the shortest C-terminal loops, but N-terminal loops of the same or shorter lengths than Subfamilies 4–8 members, which have no EGs. This supports the observation drawn from the tertiary structures of *H. insolens* and *H. jecorina* CBHs and *T. fusca* and *H. insolens* EGs (Figure 3.1),$^{3,7,8,11}$ buttressed by mutation of *C. fimi* CBH,$^{13}$ that shortening the tunnel-forming loops leads to endo activity. The little variation in loop lengths within individual subfamilies suggests, as previously, that most or all of the UPs in Subfamilies 1 and 2 will eventually be identified as EGs, while those in Subfamilies 4–8 will
eventually be classified as CBHs.

A substantial complication is introduced by the recent publication of the tertiary structure of Subfamily 1 *M. tuberculosis* EG with a cleft lightly closed by C- and N-terminal loops\textsuperscript{12} (Figure 3.1a), even though they are short, as is typical of other Subfamily 1 EGs. One would expect that they must be pulled back during substrate binding to allow EG action. There are a sufficient number of mutations within the loops to easily allow such folding.

Subfamily 3 contains enzymes classified as CBHs and mixed EG/CBHs, with a single EG. We suggested above that most or all would eventually be classified as mixed EG/CBHs. This is tentatively supported by the observation that nearly all its members have C- and N-terminal loops slightly shorter than those found in members of Subfamilies 4–8, but C-terminal loops substantially longer than those of Subfamily 1 and 2 members.

These results extend earlier observations of the effect of loop length and folding on enzyme product patterns from the five enzymes with crystal structures to all 67 GH6 family members.

**Conclusions**

A phylogenetic analysis of GH6 enzymes has identified eight subfamilies to which they belong. Aligning amino acid sequences of new GH6 members with those of existing family members and assigning them to subfamilies will assist in inferring their structures and modes of action before the difficult tasks of determining their tertiary structures and performing detailed mechanistic studies on them is undertaken.

The GH6 MSA shows two of the four Asp residues, Asp\textsuperscript{175} and Asp\textsuperscript{263}, identified experimentally as being important for catalysis, are completely conserved, while Asp\textsuperscript{221} may not be conserved in one sequence and Asp\textsuperscript{401} is not conserved in a second sequence diverging significantly from the other members of its subfamily.

Lengths of one of the two loops that form the active-site tunnel in the GH6 enzymes are well correlated with the identities of their products through all 67 aligned enzymes. Long C-terminal loops are found in CBHs; much shorter loops are found in EGs. Slightly shorter C-terminal loops are found in a subfamily of mixed EGs and CBHs. EG activity in an enzyme with a lightly closed tunnel must be explained by dissociation of the opposing loops during substrate binding.
The authors thank Professor Derrick Rollins for his advice on statistics on subfamily separation. They are grateful to the U.S. Department of Agriculture through the Biotechnology Byproducts Consortium, a grouping of Iowa State University and the University of Iowa, and to the National Science Foundation for funding this project.

References


7. Spezio, M.; Wilson, D. B.; Karplus, P. A. Crystal structure of the catalytic domain of a


Chapter 4. Automated Docking to Explore Subsite Binding by Glycoside Hydrolase Family 6 Cellobiohydrolases and Endoglucanases

A paper published in Biopolymers¹

Blake Mertz,² Anthony D. Hill,³ Chandrika Mulakala,⁴ Peter J. Reilly⁵♣∇

Department of Chemical and Biological Engineering, Iowa State University, Ames, IA 50011

Abstract

Cellooligosaccharides were computationally docked using AutoDock into the active sites of the glycoside hydrolase Family 6 enzymes Hypocrea jecorina (formerly Trichoderma reesei) cellobiohydrolase and Thermobifida fusca endoglucanase. Subsite –2 exerts the greatest intermolecular energy in binding β-glucosyl residues, with energies progressively decreasing to either side. Cumulative forces imparting processivity exerted by these two enzymes are significantly less than by the equivalent glycoside hydrolase Family 7 enzymes studied previously. Putative subsites –4, –3, +3, and +4 exist in H. jecorina cellobiohydrolase, along with putative subsites –4, –3, and +3 in T. fusca endoglucanase, but they are less important than subsites –2, –1, +1, and +2. In general, binding adds 3–7 kcal/mol to ligand intramolecular energies because of twisting of scissile glycosidic bonds. Distortion of β-glucosyl residues to the $^3S_0$ conformation by binding in subsite –1 adds ~7 kcal/mol to substrate intramolecular energies.

---

¹ Reprinted with permission from Biopolymers, Vol. 87, 249-260 (2007)
² Primary researcher and author.
³ Consulted for use of all computer programs and to discuss simulation results.
⁴ Consulted for discussion of force methodology and analysis of results.
⁵ Associate professor.
♣ Correspondence to: Peter J. Reilly, Department of Chemical and Biological Engineering, 2114 Sweeney Hall, Ames, IA 50011-2230. Phone: 515-294-5968; fax: 515-294-2689; e-mail: reilly@iastate.edu
∇ Contract grant sponsors and numbers: U. S. Department of Agriculture (USDA) through the Biotechnology Byproducts Consortium, 2004-34188-15067; National Science Foundation (NSF), BES-0313878.
Keywords: AutoDock; automated docking; cellulbiohydrolase; endoglucanase; glycoside hydrolase Family 6; Hypocrea jecorina; Thermobifida fusca; Trichoderma reesei

Introduction

Glycoside hydrolase Family 6 (GH6) \(^1\) comprises many cellbiohydrolases (β-1,4-glucan cellbiohydrolases, EC3.2.1.91, CBHIIs, CBHs) and endoglucanases (β-1,4-glucan glucanohydrolases, EC3.2.1.4, EGs). CBHs cleave the ends of cellulose and cellooligosaccharide chains to yield mostly cellobiose (Cel\(_2\)), while EGs produce a range of cellooligosaccharides from the same substrates, attacking throughout their chains.

The differences between CBHs and EGs in product patterns from cellulose hydrolysis can be partially elucidated by comparing their known tertiary structures. Five GH6 enzymes, Hypocrea jecorina (formerly Trichoderma reesei) QM9414 CBH, \(^2,3,4,5\) Humicola insolens CBH, \(^6,7,8\) (Figure 4.1a) Thermobifida (formerly Thermonospora) fusca E2 EG, \(^9,10\) H. insolens EG, \(^11\) and Mycobacterium tuberculosis H37Rv EG, \(^12\) (Figure 4.1b) have published crystal structures. The GH6 catalytic domain is a distorted β/α barrel, with the active site located at the C-terminal end of the barrel. CBH active sites are mainly enclosed in ~20-Å long, 4-Å diameter tunnels formed by two loops, \(^2,6\)
one from each side, connected to each other by a single hydrogen bond. Amino acid residue deletions shorten the C-terminal loops in EGs, while their N-terminal loops may be pulled back. A cleft rather than a tunnel results, allowing cellulose chains to enter the active site at interior glycosidic bonds. The role of the tunnel in conferring exo-type specificity on CBHs was confirmed by deleting the C-terminal loop of *Cellulomonas fimi* CBH, increasing its endo-type action. Loop flexibility changes relative endo or exo activity in some GH6 enzymes: *H. insolens* CBH showed EG activity by hydrolyzing β-1,4-glycosidic bonds in short celloligosaccharide chains having bulky substituents at one or both ends, tertiary structures from *M. tuberculosis* EG show both loops forming an active-site tunnel typical of CBHs, and the active-site loops in *H. jecorina* and *H. insolens* CBH crystal structures assume varying degrees of open and closed conformations depending on the presence of ligands or single mutations.

GH6 CBH and EG active sites are composed of four main subsites, labeled –2, –1, +1, and +2, with glycosidic bond cleavage occurring between β-glucosyl residues bound in subsites –1 and +1. Experimental and computational evidence suggests the existence of putative subsites +3 and +4, as well as subsites –3 and –4. However, these subsites bind β-glucosyl residues less strongly, as shown by the propensity of short ligands to bind in subsites –2 to +2 when crystallized with these enzymes, and by the high isotropic temperatures of these residues when modeled in subsites +3 and +4. The absolute requirement in *H. jecorina* CBH that subsite –2 but not subsite +2 be occupied before hydrolysis can occur is exemplified by the production of α-Cel2 from both Cel3 and β-cellobiosyl fluoride. Experiments on labeled Cel5 and other cellooligosaccharides show that they are attacked by CBHs at their nonreducing ends, implying that this end enters the active-site tunnel first. CBH inhibition from glucose binding in subsite –2 further emphasizes the importance of substrate occupation of subsite –2 to cause hydrolysis.

Various cellooligosaccharides and substrate analogs have been co-crystallized with GH6 enzymes, most with the glucosyl residue in subsite –1 distorted from the energetically favorable \( C_1 \) chair conformation to a \( S_0 \) skew-boat conformation, coupled with a twisting of the residue away from the plane of the cellulose substrate. The majority of *H. jecorina* and *H. insolens* CBH crystal structures with an occupied subsite –1 have the β-glucosyl residue in a
$^2S_O$ conformation, $^{2,4,8}$ which is close to the proposed $^2B$ transition-state conformation seen in another $H. insolens$ CBH crystal structure (1OCN) complexed with two isofagamine ligands. $^{23}$ Four EG crystal structures contain complexed ligands occupying subsite –1, but only the two wild-type structures are distorted from the $^4C_1$ conformation: both $M. tuberculosi$ EG complexed with an isofagamine residue $^{12}$ and $T. fusca$ EG complexed with a Cel$_4$ analog $^{10}$ have $^2S_O$ conformations.

Computational studies have provided insight into the mechanism of GH6 enzymes and the structures of substrates bound in their active sites. Taylor et al. $^{24}$ employed molecular mechanics and molecular dynamics (MD) to optimally position Cel$_4$ in the $T. fusca$ EG active site, finding that the $\beta$-glucosyl residue in subsite –1 was strongly twisted, as later found experimentally. $^{4,8}$ More recently, Koivula et al. $^5$ conducted a MD study on both the wild-type form and on two mutants of $H. jecorina$ CBH using Cel$_4$ and the $\alpha$-cellobiose and $\beta$-cellobiose products in the active site, confirming the propensity of the reactant residue in subsite –1 to be distorted into a $^2S_O$ conformation. They also confirmed the role of Asp221 as the catalytic acid, with Asp175 contributing stabilizing interactions between a water molecule acting as a potential nucleophile to facilitate hydrolysis. A follow-up large-body MD simulation on $T. fusca$ EG demonstrated the role that N-terminal loop flexibility had in bringing a potential catalytic base, Asp79, from $\sim 13 \text{ Å}$ to within $\sim 5 \text{ Å}$ of the scissile glycosidic bond. $^{25}$

Given the varying structures and functions of GH6 enzymes, the purpose of this project was to computationally dock the homologous series $b$-glucose through $\beta$-Cel$_6$ in the active sites of $H. jecorina$ CBH and $T. fusca$ EG to further investigate their substrate specificities and subsite binding energies and processivities, and by extension those of other family members. We chose $H. jecorina$ CBH because it is the only nonmutated GH6 CBH that has been crystallized with a substrate analog occupying the four main subsites, $-2/-1/+1/+2$, and $T. fusca$ EG because of its significant difference in active-site loop lengths compared to $H. jecorina$ CBH and other crystallized GH6 EGs. $^{26}$ We also chose these two enzymes because they are members of widely-separated GH6 subfamilies $^{26}$ and because of the extensive research on them.

Automated docking of ligands is a powerful tool to extend knowledge beyond that
elicited from crystal structures of enzymes complexed with ligands, since it allows the rapid screening and study of many different ligands, including substrates not readily crystallized in the enzyme active site. Another advantage is that docking provides estimates of binding energies for both individual residues and the entire docked ligand. We have used AutoDock 3.06 (Scripps Research Institute, La Jolla, CA) to study conformations of enzyme-ligand complexes and to further elucidate the mechanisms of the glycoside hydrolases glucoamylase, β-amylase, Cel7A, Cel7B, and α-1,2-mannosidase.

Computational Methods

Docking protocol. H. jecorina CBH and T. fusca EG crystal structures, 1QK2 and 2BOD, respectively, were obtained from the Protein Data Bank. All water atoms were removed, and hydrogen atoms were added to the structures using the WHAT IF web interface. AutoDock’s myqpdb.awk shell script assigned charges according to Cornell et al. The all-atom parameters of AutoDock 1.0 based on the AMBER force field were used for nonbonded interactions, since they best reproduced crystal ligand conformations. Implicit solvation parameters were added with AddSol, part of the AutoDock Tools package. Nonpolar hydrogen atoms were designated as atom type ‘X’. AutoGrid, also part of AutoDock, computed four docking grids for carbon, oxygen, and polar and nonpolar hydrogen atoms, plus a fifth grid for electrostatic potential, at 0.375-Å spacing. A sixth grid, this one for sulfur, was constructed when β-D-glucopyranosyl-(1→4)-β-D-glucopyranosyl-(1→4)-4-thio-β-D-glucopyranosyl-(1→4)-1-methyl-β-D-glucopyranoside (methyl celllobiosyl-4-thio-β-celllobioside, MCTC) was docked. Grid maps contained 111 x 41 x 41 points for Cel5 and Cel6 ligands to constrain them within the active site; for all other ligands, a 71 x 71 x 71 point grid map was used, centered on the ligands’ midpoint.

Carbohydrate ligands, including hydrogen atoms, were prepared in PCModel 9.0 (Serena Software, Bloomington, IN). GAMESS (May 19, 2004 version) calculated ligand Mulliken charges. Ligand nonpolar hydrogen atoms were marked as atom type ‘X’. Rotable bonds were defined using AutoTors, also part of the AutoDock package.

Each ligand was placed in the subsites of interest by overlaying it onto either MCTC crystallized in the H. jecorina CBH and T. fusca EG active sites. MCTC was docked in
subsites −2/−1/+1/+2 of these enzymes starting with its crystal-structure coordinates (Table 4.1). With β-glucose docked in subsites −4 through +4 of H. jecorina CBH or T. fusca EG, ring puckering coordinates were θ = 3.7°, φ = 358.0°, and q = 0.598 Å, those of the optimal \(^4C_1\) conformation of β-glucopyranose\(^{38}\) found by the molecular mechanics program MM3.\(^{39}\) Coordinates of θ = 89.9°, φ = 165.7°, and q = 0.690 Å for the \(^2SO\) conformation of β-glucopyranose in crystal-structure MCTC complexed with H. jecorina CBH\(^4\) were also used for glucose bound in subsite −1. For other ligands docked in subsites −2 to +2, puckering coordinates and glycosidic bond angles of crystal-structure MCTC\(^4\) (Table 4.1) were used. β-Glucosyl residues docked in subsites −4, −3, +3, and +4 had \(^4C_1\) conformations\(^{38}\) and dihedral angles\(^{40}\) of the optimal conformation of β-Cel\(_2\) [φ = −86.0°, ψ = 77.1°, where φ = O5’ – C1’ – O4–C4 and ψ = C1’–O4–C4–C3] found by MM3. Those in subsite −1 as part of longer ligands were in either the \(^4C_1\) or \(^2SO\) conformation. In addition, α-Cel\(_2\) was docked in subsites −2/−1 and then fixed in place while β-Cel\(_4\) was docked in subsites +1 to +4 (Tables 4.2 and 4.3) to examine the effect of a product bound in the active site. Computer scripts to automate overlaying the ligand and renaming the nonpolar hydrogen atoms are available in the Supplementary Material.

**Table 4.1** Ring Properties of MCTC Cocrystallized in the H. jecorina CBH\(^4\) and T. fusca EG Active Sites\(^10\)

<table>
<thead>
<tr>
<th>Subsite</th>
<th>Residue</th>
<th>θ/φ (degrees)</th>
<th>Q(Å)</th>
<th>Conformation</th>
<th>Angle, degrees(^a)</th>
</tr>
</thead>
<tbody>
<tr>
<td>−2</td>
<td>β-D-Glc(^b)</td>
<td>2.7/9.5</td>
<td>0.578/0.538</td>
<td>(^4C_1)</td>
<td>−77.5/−68.3(^c), 142.5/140.9</td>
</tr>
<tr>
<td>−1</td>
<td>β-D-Glc(^b)</td>
<td>89.9/44.5</td>
<td>0.690/0.502</td>
<td>(^2SO^2E)</td>
<td>−125.3/−156.0, 89.6/83.0</td>
</tr>
<tr>
<td>+1</td>
<td>4-Thio-β-D-Glc</td>
<td>4.5/21.3</td>
<td>0.571/0.164</td>
<td>(^4C_1)</td>
<td>−73.7/−75.1, 95.6/100.3</td>
</tr>
<tr>
<td>+2</td>
<td>Methyl β-D-Glc</td>
<td>12.9/21.0</td>
<td>0.562/0.148</td>
<td>(^4C_1)</td>
<td></td>
</tr>
</tbody>
</table>

\(^a\) Angles of glycosidic bonds between residues bound in subsites −2/−1, −1/+1, and +1/+2..

\(^b\) Glucopyranosyl/glucopyranose.

\(^c\) H. jecorina CBH/T. fusca EG.

AutoDock 3.06 used a combined global and local search algorithm with the Lamarkian genetic algorithm (LGA) to determine the docked conformations and total energies (\(E_{\text{Total}}\)) of
the resulting complexes. This algorithm repeatedly perturbs ligands to find their optimal positions and most negative $E_{\text{Total}}$ values, allowing their pendant groups and $\beta$-glycosidic bond angles to rotate but holding their ring conformations and the enzyme rigid. One thousand runs were performed for each docked ligand. The initial population for the LGA was 50 individuals, the maximal number of energy evaluations and generations was $2 \times 10^6$ and 500, respectively, the number of top individuals that survived to the next generation was one, the rate of gene mutation was 0.02, the rate of gene crossover was 0.80, and the worst individual was calculated over a ten-generation window.

For the LGA local search component, $1.5 \times 10^4$ iterations were allowed per local Solis and Wets (SW) search,\textsuperscript{41} the number of consecutive successes or failures before doubling or halving the step size was four, the probability of performing a local search was 0.7, and the lower bound on $\rho$ (the size of local search space to sample) was 0.01. One thousand LGA dockings were performed in a given docking run. Six clusters of docked structures were retained based on these criteria: (1) lowest $E_{\text{Total}}$ values; (2) largest number of docked conformations; (3) clusters within root mean square deviation (RMSD) limits for $H. jecorina$ CBH and $T. fusca$ EG of 1.0 Å for ligands smaller than Cel$_5$ and 2.0 Å for Cel$_5$ and Cel$_6$. The best docked member of each cluster was further minimized 30 times successively\textsuperscript{32} by the SW local search method\textsuperscript{41} with 200 steps per minimization. Local search parameters were identical to LGA docking run parameters except that the maximal number of iterations allowed was 300 and the probability of performing a local search was 1.0. The structure with overall lowest $E_{\text{Total}}$ value along with an RMSD within acceptable limits was chosen as the final global optimal structure. The much greater rigor of the local search compared to that normally used leads to much more negative $E_{\text{Total}}$ values than normally attained by AutoDock\textsuperscript{32}.

Instantaneous forces were calculated by numerical differentiation of the energy landscape in the three Cartesian coordinates.\textsuperscript{30} These forces are represented by the equation $-\nabla E_{\text{Inter}} = F_{\text{Inter}}$, where $F_{\text{Inter}}$ is the force in the three spatial dimensions that the enzyme exerts on the ligand and $E_{\text{Inter}}$ is the intermolecular energy. $F_{\text{Inter}}$ can be nonzero even when the system is at rest. $E_{\text{Total}} = E_{\text{Inter}} + E_{\text{Intra}}$, where $E_{\text{Intra}}$ is the ligand intramolecular energy, and a system at rest lies at a minimum of $E_{\text{Total}}$, although not necessarily at a minimum of either $E_{\text{Inter}}$ or $E_{\text{Intra}}$. When $E_{\text{Total}}$ reaches a minimum, $F_{\text{Total}} = 0$ and $F_{\text{Inter}} = -F_{\text{Intra}}$. The force that the enzyme
exerts on the ligand is exactly countered by the force that the ligand exerts on itself through distortion.

Upon bond cleavage, $F_{\text{Intra}}$ changes and the entire system again moves toward a low-energy state. One can therefore compute the expelling and processive forces on the products by adding the $F_{\text{Inter}}$ values of the products' initial ligand components. We can assume that $F_{\text{Intra}}$ will decrease upon bond cleavage and that $F_{\text{Inter}}$ will direct the system towards global energy minimization. Thus, the $F_{\text{Inter}}$ value for the newly formed reactants gives an idea of which direction the system will move after hydrolysis.

The enzyme-ligand system when optimally docked is at its lowest $E_{\text{Total}}$ value with regard to the ligand’s six transformational degrees of freedom and its dihedral rotational degrees of freedom. Shown instead in Tables 4.2 and 4.3 are values of $E_{\text{Inter}}$, since AutoDock, despite ranking dockings according to $E_{\text{Total}}$, cannot accurately estimate $E_{\text{Intra}}$ values. Inaccuracies arise from AutoDock’s inability to rigorously estimate hydrogen bond geometries and exo-anomeric effects.\textsuperscript{32,42} Despite this, $E_{\text{Inter}}$ values can be compared because they measure the interaction energies of each ligand atom with the enzyme.

Crystallographic studies of GH6 enzymes have shown little change in protein structure upon ligand binding, as established with the first GH6 crystal structure,\textsuperscript{2} which lends validity to the rigid enzyme assumption in AutoDock. Also, pair-fitting of crystal structures for both CBHs and EGs (Figures 4.1a and 4.1b) with both empty active sites and co-complexed ligands shows conservation of the active-site shape, which would indicate that a rigid enzyme is a safe assumption to make for this particular system. Furthermore, docking with and without the water of hydrolysis yielded essentially the same docking results, so the latter option was chosen.

**Results**

**Docking in the H. jecorina CBH Active Site.** MCTC with the initial glycosidic bond angles and ring conformations shown in Table 4.1 was docked in subsites –2/-1/+1/+2 of the *H. jecorina* CBH active site, giving an subsite RMSDs from the crystal-structure MCTC of 0.83, 1.12, 0.48, and 0.92 Å for subsites –2/-1/+1/+2 and an overall RMSD of 0.77 Å. This demonstrates that AutoDock can dock ligands into the *H. jecorina* CBH active site.

β-Glucose molecules were docked individually in subsites –4 through +4. Subsite –2
binds β-glucose most strongly, and this molecule has a much lower RMSD from MCTC
bound in the CBH crystal structure than when docked in subsites –1, +1, or +2 (Table 4.2). In
fact, the majority of β-glucose molecules initially placed in subsites –1, +1, and +2 docked in
subsite –2. Values of $E_{\text{Inter}}$ become progressively less negative in general at greater distances
from subsite –2. Subsite –1 binds $^4C_1$-β-glucose and $^2S_O$-β-glucose equally strongly.

Table 4.2 Intermolecular Energies (kcal/mol) of Ligands Docked in Different Subsites of *H. jecorina* CBH

<table>
<thead>
<tr>
<th>Ligand</th>
<th>Conformation</th>
<th>Subsite –4</th>
<th>–3</th>
<th>–2</th>
<th>–1</th>
<th>+1</th>
<th>+2</th>
<th>+3</th>
<th>+4</th>
<th>Total</th>
<th>RMSD <em>a</em></th>
</tr>
</thead>
<tbody>
<tr>
<td>β-Glucose</td>
<td>$^4C_1$</td>
<td>–53.9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>–53.9</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$^4C_1$</td>
<td>–76.4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>–76.4</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$^4C_1$</td>
<td>–90.7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>–90.7</td>
<td>0.60</td>
</tr>
<tr>
<td></td>
<td>$^4C_1$</td>
<td>–85.9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>–85.9</td>
<td>3.44</td>
</tr>
<tr>
<td></td>
<td>$^2S_0$</td>
<td>–83.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>–83.5</td>
<td>3.06</td>
</tr>
<tr>
<td></td>
<td>$^4C_1$</td>
<td>–82.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>–82.5</td>
<td>3.62</td>
</tr>
<tr>
<td></td>
<td>$^4C_1$</td>
<td>–67.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>–67.0</td>
<td>1.38</td>
</tr>
<tr>
<td></td>
<td>$^4C_1$</td>
<td>–68.7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>–68.7</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$^4C_1$</td>
<td>–56.6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>–56.6</td>
<td></td>
</tr>
<tr>
<td>β-Cel2</td>
<td>$^4C_1$/$^2S_0$</td>
<td>–81.7</td>
<td>–61.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>–143.2</td>
<td>0.58</td>
</tr>
<tr>
<td></td>
<td>$^4C_1$/$^4C_1$</td>
<td>–66.4</td>
<td>–59.9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>–126.3</td>
<td>2.04</td>
</tr>
<tr>
<td></td>
<td>$^2S_0$/$^4C_1$</td>
<td>–74.6</td>
<td>–55.2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>–129.8</td>
<td>1.10</td>
</tr>
<tr>
<td></td>
<td>$^4C_1$/$^4C_1$</td>
<td>–63.3</td>
<td>–65.1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>–128.4</td>
<td>1.23</td>
</tr>
<tr>
<td></td>
<td>$^4C_1$/$^4C_1$</td>
<td>–63.0</td>
<td>–33.8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>–96.8</td>
<td></td>
</tr>
<tr>
<td>β-Cel3</td>
<td>$^4C_1$/$^2S_0$/$^4C_1$</td>
<td>–83.5</td>
<td>–65.6</td>
<td>–50.4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>–199.5</td>
<td>0.87</td>
</tr>
<tr>
<td></td>
<td>$^2S_0$/$^4C_1$/$^4C_1$</td>
<td>–66.5</td>
<td>–53.3</td>
<td>–56.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>–176.3</td>
<td>0.94</td>
</tr>
<tr>
<td></td>
<td>$^4C_1$/$^4C_1$/$^4C_1$</td>
<td>–61.8</td>
<td>–58.6</td>
<td>–36.6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>–157.0</td>
<td>1.02</td>
</tr>
<tr>
<td>β-Cel4</td>
<td>$^4C_1$/$^2S_0$/$^4C_1$/$^4C_1$</td>
<td>–26.0</td>
<td>–38.5</td>
<td>–56.7</td>
<td>–51.1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>–172.4</td>
<td>1.69</td>
</tr>
<tr>
<td></td>
<td>$^4C_1$/$^2S_0$/$^4C_1$/$^4C_1$</td>
<td>–61.7</td>
<td>–53.8</td>
<td>–62.2</td>
<td>–65.1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>–242.8</td>
<td>1.56</td>
</tr>
<tr>
<td></td>
<td>$^4C_1$/$^4C_1$/$^4C_1$/$^4C_1$</td>
<td>–80.5</td>
<td>–47.1</td>
<td>–49.8</td>
<td>–58.3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>–235.7</td>
<td>1.64</td>
</tr>
<tr>
<td></td>
<td>$^4C_1$/$^2S_0$/$^4C_1$/$^4C_1$</td>
<td>–76.1</td>
<td>–62.5</td>
<td>–47.0</td>
<td>–57.1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>–242.7</td>
<td>0.82</td>
</tr>
<tr>
<td></td>
<td>$^2S_0$/$^4C_1$/$^4C_1$/$^4C_1$</td>
<td>–59.9</td>
<td>–44.9</td>
<td>–55.3</td>
<td>–28.8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>–188.9</td>
<td>0.61</td>
</tr>
<tr>
<td></td>
<td>$^4C_1$/$^4C_1$/$^4C_1$/$^4C_1$</td>
<td>–48.3</td>
<td>–56.5</td>
<td>–56.4</td>
<td>–32.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>–193.3</td>
<td>0.89</td>
</tr>
<tr>
<td></td>
<td>$^4C_1$/$^4C_1$/$^4C_1$/$^4C_1$</td>
<td>–56.3</td>
<td>–55.5</td>
<td>–39.0</td>
<td>–27.3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>–178.1</td>
<td>0.63</td>
</tr>
<tr>
<td>β-Cel5</td>
<td>$^4C_1$/$^4C_1$/$^4C_1$/$^4C_1$</td>
<td>–59.1</td>
<td>–45.0</td>
<td>–55.6</td>
<td>–42.8</td>
<td>–26.2</td>
<td></td>
<td></td>
<td></td>
<td>–228.6</td>
<td>2.44</td>
</tr>
<tr>
<td></td>
<td>$^2S_1$/$^4C_1$/$^4C_1$/$^4C_1$</td>
<td>–47.7</td>
<td>–45.3</td>
<td>–53.9</td>
<td>–47.2</td>
<td>–28.7</td>
<td></td>
<td></td>
<td></td>
<td>–222.8</td>
<td>1.85</td>
</tr>
<tr>
<td>β-Cel6</td>
<td>$^4C_1$/$^4C_1$/$^4C_1$/$^4C_1$/$^4C_1$</td>
<td>–57.4</td>
<td>–36.5</td>
<td>–42.4</td>
<td>–32.6</td>
<td>–29.9</td>
<td>–15.1</td>
<td></td>
<td></td>
<td>–213.9</td>
<td>1.93</td>
</tr>
<tr>
<td></td>
<td>$^4C_1$/$^2S_0$/$^4C_1$/$^4C_1$/$^4C_1$</td>
<td>–80.5</td>
<td>–59.0</td>
<td>–45.8</td>
<td>–44.8</td>
<td>–39.6</td>
<td>–22.9</td>
<td></td>
<td></td>
<td>–292.6</td>
<td>1.22</td>
</tr>
</tbody>
</table>

* RMSD measured from corresponding residues of MCTC crystallized in subsites –2/–1/+1/+2.
* a-Cel2 docked in subsites –2/–1.
* β-Cel4 docked in subsites +1/+2/+3/+4 next to a-Cel1 docked in subsites –2/–1.

Different ligands from β-Cel2 through β-Cel6 were docked in various combinations of
subsites –4 through +4 of *H. jecorina* CBH. Table 4.2 shows that $E_{\text{Inter}}$ values of residues are
generally most negative at subsite –2 and are somewhat less negative in subsites –1, +1, and
+2, with residues docked in putative subsites –4, -3, +3, and +4 still having significant $E_{\text{Inter}}$ values. This extends the results from β-glucose docking.

As with β-glucose, $E_{\text{Inter}}$ values may be compared for residues with different conformations docked in subsite –1. In two of three cases ligands with $^2S_O$-glucosyl residues in subsite –1 ($^2S_O$-ligands) have more negative overall $E_{\text{Inter}}$ values than the corresponding ligands with $^4C_1$-residues located there (Table 4.2). In the same cases, the $^2S_O$-residues in subsite –1 have significantly more negative $E_{\text{Inter}}$ values than the corresponding $^4C_1$-residues. In all three cases the $^2S_O$-ligands have lower total RMSDs. In general, the most negative overall $E_{\text{Inter}}$ values occur when the nonreducing-end residues of β-Cel$_2$ through β-Cel$_6$ are docked in subsite –2. Results with β-Cel$_6$ docked in subsites –2 through +4, the natural position for hydrolysis to occur, are especially noteworthy, since the $^2S_O$-ligand has an extremely negative total $E_{\text{Inter}}$ value, with more negative values at each subsite compared to those of the $^4C_1$-ligand. The $E_{\text{Inter}}$ value of α-Cel$_2$ is slightly more negative than that of β-Cel$_2$ when each is docked in subsites –2/–1 but the former has a much higher RMSD (Table 4.2). β-Cel$_4$ docked next to α-Cel$_2$ in subsites +1 through +4 has an $E_{\text{Inter}}$ value 15 kcal/mol less negative than when it is docked without α-Cel$_2$, with the $E_{\text{Inter}}$ value of the former’s residue being more negative in subsite +1 but with its residues in subsites +3/+4 having less negative $E_{\text{Inter}}$ values.

Docking in the T. fusca EG Active Site. MCTC with initial glycosidic bond angles and ring conformations shown in Table 4.1 was docked in subsites –2/–1/+1/+2 of the T. fusca EG active site, giving subsite RMSDs from the crystal-structure MCTC of 0.74, 0.37, 0.46, and 0.41 Å for subsites –2/–1/+1/+2 and an overall RMSD of 0.51 Å, respectively. This demonstrates that AutoDock can dock ligands into the T. fusca EG active site.

β-Glucose was docked in T. fusca EG subsites –4 through +4. Unlike H. jecorina CBH, β-glucose docked in subsite +3 has the most negative $E_{\text{Inter}}$ value, followed in order by β-glucose docked in subsites –2, –4, –1, –3, +2, +1, and +4 (Table 4.3). The result for β-glucose docked in subsite +3 can be somewhat discounted because its best docked position is a location deep inside the active site cleft behind subsite +2, which would create physically impossible steric clashes in an actual cellulosic substrate. Eliminating this ligand leaves β-glucose docked in subsite –2 as the docked ligand with the most negative $E_{\text{Inter}}$ value, just as with H. jecorina CBH. Also, the majority of β-glucose molecules initially placed in subsites
–1, +1, and +2 docked in subsite –2. The $E_{\text{inter}}$ value of the $^2S_O$ conformer in subsite –1 is 13 kcal/mol less negative than that of the $^4C_1$ conformer.

Table 4.3 Intermolecular Energies (kcal/mol) of Ligands Docked in Different Subsites of $T. \text{fusca}$ EG

<table>
<thead>
<tr>
<th>Ligand</th>
<th>Conformation</th>
<th>Subsite</th>
<th>Total</th>
<th>RMSD$^a$</th>
</tr>
</thead>
<tbody>
<tr>
<td>β-Glucose</td>
<td>$^4C_1$</td>
<td>–85.4</td>
<td>–85.4</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$^4C_1$</td>
<td>–91.6</td>
<td>–91.6</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$^4C_1$</td>
<td>–81.6</td>
<td>–81.6</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$^2S_0$</td>
<td>–68.6</td>
<td>–68.6</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$^4C_1$</td>
<td>–60.3</td>
<td>–60.3</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$^4C_1$</td>
<td>–73.2</td>
<td>–73.2</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$^4C_1$</td>
<td>–95.9</td>
<td>–95.9</td>
<td></td>
</tr>
<tr>
<td>β-Cel$_2$</td>
<td>$^4C_1/r^2S_0$</td>
<td>–73.5 –70.2</td>
<td>–143.7</td>
<td>0.80</td>
</tr>
<tr>
<td></td>
<td>$^4C_1/r^2C_1$</td>
<td>–76.0 –63.4</td>
<td>–139.4</td>
<td>1.80</td>
</tr>
<tr>
<td></td>
<td>$^2S_0/r^2C_1$</td>
<td>–67.3 –63.7</td>
<td>–131.0</td>
<td>1.02</td>
</tr>
<tr>
<td></td>
<td>$^4C_1/r^2C_1$</td>
<td>–64.0 –51.3</td>
<td>–115.3</td>
<td>1.00</td>
</tr>
<tr>
<td>β-Cel$_3$</td>
<td>$^4C_1/r^2C_1/r^4C_1$</td>
<td>–53.3 –65.3 –62.5</td>
<td>–181.1</td>
<td>0.98</td>
</tr>
<tr>
<td></td>
<td>$^2S_0/r^2C_1/r^4C_1$</td>
<td>–65.9 –50.2 –47.9</td>
<td>–164.0</td>
<td>0.82</td>
</tr>
<tr>
<td></td>
<td>$^4C_1/r^2C_1/r^4C_1$</td>
<td>–49.6 –22.9 –11.0</td>
<td>–83.5</td>
<td>1.17</td>
</tr>
<tr>
<td>β-Cel$_4$</td>
<td>$^4C_1/r^2C_1/r^4C_1/r^2S_0$</td>
<td>–22.3 –33.8 –56.0 –65.0</td>
<td>–177.1</td>
<td>0.79</td>
</tr>
<tr>
<td></td>
<td>$^4C_1/r^2C_1/r^4C_1/r^2C_1$</td>
<td>–43.1 –49.9 –54.8 –39.5</td>
<td>–187.2</td>
<td>0.87</td>
</tr>
<tr>
<td></td>
<td>$^4C_1/r^2C_1/r^4C_1/r^4C_1$</td>
<td>–53.2 –44.2 –53.3 –46.4</td>
<td>–197.1</td>
<td>1.73</td>
</tr>
<tr>
<td></td>
<td>$^2S_0/r^2C_1/r^4C_1/r^4C_1$</td>
<td>–66.1 –61.4 –53.4 –41.8</td>
<td>–222.7</td>
<td>0.81</td>
</tr>
<tr>
<td></td>
<td>$^2S_0/r^2C_1/r^4C_1/r^4C_1$</td>
<td>–65.6 –48.7 –43.6 –19.1</td>
<td>–177.0</td>
<td>0.96</td>
</tr>
<tr>
<td></td>
<td>$^4C_1/r^2C_1/r^4C_1/r^4C_1$</td>
<td>–54.8 –43.7 –28.6 –13.7</td>
<td>–140.8</td>
<td>1.15</td>
</tr>
<tr>
<td></td>
<td>$^4C_1/r^2C_1/r^4C_1/r^4C_1$</td>
<td>–68.5 –45.4 –29.3 –11.0</td>
<td>–154.2</td>
<td>1.15</td>
</tr>
<tr>
<td>β-Cel$_5$</td>
<td>$^4C_1/r^2C_1/r^4C_1/r^4C_1/r^4C_1$</td>
<td>–65.8 –46.3 –46.7 –28.5 –1.3</td>
<td>–188.6</td>
<td>2.57</td>
</tr>
<tr>
<td></td>
<td>$^2S_0/r^2C_1/r^4C_1/r^4C_1$</td>
<td>–63.8 –45.4 –35.4 –19.3 –4.2</td>
<td>–168.1</td>
<td>1.27</td>
</tr>
<tr>
<td>β-Cel$_6$</td>
<td>$^4C_1/r^2C_1/r^4C_1/r^4C_1/r^4C_1$</td>
<td>–72.5 –43.4 –50.9 –47.0 –24.1 –1.4</td>
<td>–239.3</td>
<td>1.72</td>
</tr>
<tr>
<td></td>
<td>$^2S_0/r^2C_1/r^4C_1/r^4C_1/r^4C_1$</td>
<td>–50.4 –47.0 –45.9 –38.4 –33.1 –6.7</td>
<td>–221.5</td>
<td>1.28</td>
</tr>
</tbody>
</table>

$^a$ RMSD measured from corresponding residues of MCTC crystallized in subsites –2/–1/+1/+2.
$^b$ α-Cel$_2$ docked in subsites –2/–1.
$^c$ β-Cel$_4$ docked in subsites +1/+2/+3/+4 next to α-Cel$_2$ docked in subsites –2/–1.

β-Cel$_2$ through β-Cel$_6$ were docked in various combinations of $T. \text{fusca}$ EG subsites –4 through +4. Ligands with nonreducing-end residues docked in subsite –2 tend to have the most negative $E_{\text{inter}}$ values, followed by those docked in subsite –1. $E_{\text{inter}}$ values of individual residues become progressively less negative as their distances from subsites –2 and –1 increase. Residues in subsite +4 have very small $E_{\text{inter}}$ values, caused by their lying at an angle to the plane of crystallized MCTC. All but four ligands from β-Cel$_2$ through β-Cel$_6$...
complexed to *T. fusca* EG have less negative $E_{\text{Inter}}$ values than those of the same ligands complexed to *H. jecorina* CBH.

In two of three cases $^2S_O$-glucosyl residues docked in subsite –1 of *T. fusca* EG have more negative $E_{\text{Inter}}$ values than $^4C_1$-glucosyl residues in the corresponding ligands docked there. However, in only one of three cases are the overall $E_{\text{Inter}}$ values of the $^2S_O$-ligands more negative than those of the $^4C_1$-ligands, although in all three cases their RMSDs are much lower.

Docking of $\alpha$-Cel$_2$ and $\beta$-Cel$_4$ products in *T. fusca* EG subsites –2 to +4 gives results different to those obtained by docking into the same *H. jecorina* CBH subsites. Now the $E_{\text{Inter}}$ value of $\alpha$-Cel$_2$ docked in subsites –2/–1 is less negative than that of $\beta$-Cel$_2$ docked there. In addition, the $E_{\text{Inter}}$ value of $\beta$-Cel4 in subsites +1/+2/+3/+4 is ~13 kcal/mol more negative when subsites –2/–1 are occupied than when they are unoccupied, opposite to what was found with the equivalent ligands in *H. jecorina* CBH.

**Scissile Glycosidic Bond Twisting.** $^2S_O$-$\beta$-Cel$_2$ through $^2S_O$-$\beta$-Cel$_6$ optimally docked in *H. jecorina* CBH and *T. fusca* EG have the glycosidic bond angles linking residues in subsites –1 and +1 shown in Table 4.4. Most are not greatly different from the corresponding angles in the crystal-structure MCTC bound in the *H. jecorina* CBH active site ($\phi = -125.3^\circ$, $\psi = 89.6^\circ$), but they vary substantially from the MM3-calculated optimal angles in non-

<table>
<thead>
<tr>
<th>Ligand</th>
<th>Ring Conformation</th>
<th><em>H. jecorina</em> CBH</th>
<th><em>T. fusca</em> EG</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>$\phi$</td>
<td>$\psi$</td>
</tr>
<tr>
<td>$\beta$-Cel$_2$</td>
<td>$^2S_O\backslash^\phi C_1$</td>
<td>38.3</td>
<td>-31.2</td>
</tr>
<tr>
<td>$\beta$-Cel$_3$</td>
<td>$^4C_i\backslash S_O\backslash^\phi C_1$</td>
<td>-17.2</td>
<td>-27.3</td>
</tr>
<tr>
<td></td>
<td>$^2S_O\backslash^\phi C_i\backslash^\phi C_1$</td>
<td>-39.0</td>
<td>-6.6</td>
</tr>
<tr>
<td>$\beta$-Cel$_4$</td>
<td>$^4C_i\backslash^\phi C_i\backslash S_O\backslash^\phi C_i$</td>
<td>-20.8</td>
<td>-13.5</td>
</tr>
<tr>
<td></td>
<td>$^4C_i\backslash^\phi C_i\backslash^\phi C_i\backslash^\phi C_i$</td>
<td>-3.4</td>
<td>-23.3</td>
</tr>
<tr>
<td></td>
<td>$^2S_O\backslash^\phi C_i\backslash^\phi C_i\backslash^\phi C_i$</td>
<td>-15.9</td>
<td>-22.0</td>
</tr>
<tr>
<td>$\beta$-Cel$_5$</td>
<td>$^2S_O\backslash^\phi C_i\backslash^\phi C_i\backslash^\phi C_i\backslash^\phi C_i$</td>
<td>29.4</td>
<td>-41.4</td>
</tr>
<tr>
<td>$\beta$-Cel$_6$</td>
<td>$^4C_i\backslash^\phi S_O\backslash^\phi C_i\backslash^\phi C_i\backslash^\phi C_i$</td>
<td>-25.9</td>
<td>-18.9</td>
</tr>
</tbody>
</table>

a Glycosidic bond angles defined as $\phi = O5'\backslash -C1'\backslash -O4\backslash -C4$ and $\psi = C1'\backslash -O4\backslash -C4\backslash -C3$.
b Relative to the optimal $E_{\text{Intra}}$ value of $^4C_1$-$\beta$-Cel$_2$ calculated by a hybrid quantum mechanics (B3LYP/6-311++G(d,p))/molecular mechanics (MM3(96)) technique. Glycosidic angles in this reference are defined as $\phi = O5'\backslash -C1'\backslash -O4\backslash -C4$ and $\psi = C1'\backslash -O4\backslash -C4\backslash -C5$ and are converted to those above with PCModel.
c Residue with $^2S_O$ conformation is docked in subsite –1.
bound β-Cel2 ($\phi = -86.0^\circ$, $\psi = 77.1^\circ$). Therefore these scissile β-glycosidic bonds are placed under strain by being bound in the active site, as is MCTC, having $E_{\text{Intra}}$ values in general 3–7 kcal/mol more positive than the steric energy minimum of β-Cel2 calculated by a hybrid molecular mechanics/quantum mechanics method\(^\text{43}\) (Table 4.4). Likewise, the $E_{\text{Intra}}$ value of $^2S_O$-β-glucose calculated by MM3 is ~7 kcal/mol more positive than that of $^4C_1$-β-glucose\(^\text{38}\).

**Net Processive Forces.** Processive forces exerted by the *H. jecorina* CBH and *T. fusca* EG active sites are one-quarter or less those exerted by the corresponding GH7 enzymes *H. jecorina* Cel7A,\(^\text{30}\) a CBH, and *Fusarium oxysporum* Cel7B,\(^\text{31}\) an EG, respectively. These small forces lead to untrustworthy estimates of possible enzymatic processivity and will be discussed only qualitatively.

**Discussion**

The relatively low $E_{\text{Inter}}$ and RMSD values of β-glucose docked in subsite –2 demonstrate its propensity to act as a competitive inhibitor. The tendency for β-glucose to dock there, regardless of initial position, suggests that subsite –2 occupancy provides essential binding interactions to help stabilize the substrate as the subsite –1 glucosyl residue undergoes distortion to the transition-state conformation.

β-Cel2 docks in subsites –2/–1 of *H. jecorina* CBH and *T. fusca* EG significantly more tightly than in subsites –1/+1, supporting experimental results for the latter enzyme\(^\text{44}\) and the nonlikelihood of β-Cel2 as a substrate. The $E_{\text{Inter}}$ values of a-Cel2 product docked in *H. jecorina* CBH and *T. fusca* EG subsites –2/–1 are roughly the same as those for β-Cel2 docked independently in subsites –2/–1, although in each case their RMSDs are much higher. This less favorable binding may foretell the expulsion of the product from the active sites of the two enzymes. Less negative $E_{\text{Inter}}$ results for β-Cel4 ligands docked in subsites +1/+2/+3/+4 next to α-Cel2 product in subsites –2/–1 *H. jecorina* CBH compared to β-Cel4 docked in subsites +1/+2/+3/+4 with subsites –2/–1 unoccupied also suggest product expulsion. The opposite relationship occurs with *T. fusca* EG, with β-Cel4 docked next to an α-Cel2 product having an overall $E_{\text{Inter}}$ value ~14 kcal/mol more negative than when β-Cel4 is docked in subsites +1/+2/+3/+4 with subsites –2/–1 left unoccupied. The majority of this difference in $E_{\text{Inter}}$ is manifested in subsite +1, where β-Cel4 has three potential hydrogen-
bonding interactions of <2.65 Å with the α-Cel₂ product.

β-Cel₃ docks with more negative $E_{\text{Inter}}$ values in H. jecorina CBH and T. fusca EG subsites –2/–1/+1 than in subsites –1/+1/+2, extending the β-glucose and β-Cel₂ docking results. Experimental NMR evidence from H. jecorina CBH agrees well: after 1 h, reducing-end glycosidic bond cleavage to yield α-Cel₂ from the substrate’s nonreducing end and glucose at anomeric equilibrium from its reducing end occurred at an ~2.5:1 ratio to nonreducing-end glycosidic bond cleavage.¹⁹

Docked results for β-Cel₄, β-Cel₅, and β-Cel₆ highlight the essential role binding interactions play in facilitating the conformational shift to the transition state of the subsite –1 glucosyl residue. Ligands with the subsite –1 residue in the $^2S_O$ conformation have much smaller RMSD values than their counterparts with corresponding $^4C_1$ conformations, both for H. jecorina CBH and T. fusca EG. The two out of three cases in H. jecorina CBH where their $^2S_O$-ligands have more negative $E_{\text{Inter}}$ values than their corresponding $^4C_1$-ligands suggest preferential cleavage for a $^2S_O$-like conformation, as previously indicated by kinetics and mutagenesis studies.³ Zou et al. demonstrated that the tunnel-forming loops of H. jecorina CBH formed a more constrictive active site, especially in subsite –1, when the glucosyl residue was in a $^2S_O$ conformation.⁴ Comparison of $E_{\text{Inter}}$ values to H. jecorina CBH kinetic results for $^2S_O$ ligands larger than β-Cel₂ show general agreement: kcat values of 0.061 s⁻¹ for β-Cel₃, 4.1 s⁻¹ for β-Cel₄, 1.1 s⁻¹ for β-Cel₅, and 14 s⁻¹ for β-Cel₆ agree with binding results of –199.5 kcal/mol for β-Cel₃, –242.7 kcal/mol for β-Cel₄, –222.8 kcal/mol for β-Cel₅, and –292.6 kcal/mol for β-Cel₆.²⁰

In contrast, two out of three cases in T. fusca EG with the same ligands as with H. jecorina CBH have stronger overall binding without $^2S_O$ glucosyl residues in subsite –1. The opposite relationship with T. fusca EG may be due to a combination of a crystal structure induced-fit effect and the rigidity of glucosyl ring conformations in AutoDock. The co-complexed MCTC ligand in T. fusca EG has the same glucosyl ring conformation as in H. jecorina CBH in subsite –2, $^4C_1$, but has a $^2E$ conformation in subsite –1. In addition, the glucosyl residues in subsites +1/+2 of T. fusca EG have almost planar conformations, as shown by their much lower puckering amplitudes ($Q = 0.164$ Å and 0.148 Å for the residues in subsites +1 and +2, respectively) than the MM3 amplitude for $^4C_1$ ($Q = 0.598$ Å).³⁸
Support for preferential binding can be seen in the much smaller RMSDs for $^2S_O$ versions of $\beta$-Cel$_4$, $\beta$-Cel$_5$, and $\beta$-Cel$_6$ ligands compared to the $^4C_1$ conformers. These suggest the close proximity of $^2S_O$ to the $T.~fusca$ EG transition state.

The presence of significantly negative $E_{\text{inter}}$ values for residues docked in subsite +4 of $H.~jecorina$ CBH supports the stabilizing stacking action of the Trp272 indole ring (Figure 4.2a).$^{17}$ The lack of significant $E_{\text{inter}}$ values in subsite +4 of $T.~fusca$ EG is clearly due to an

**Figure 4.2.** Active-site structures of a) $H.~jecorina$ CBH and b) $T.~fusca$ EG with crystal-structure MCTC in subsites –2 through +2, the first two nonreducing end residues of $^2S_O$-\$\beta$-Cel$_4$ in subsites –4 and –3, and SB-$\beta$-Cel$_6$ in subsites –2 to +4. Orange: crystal-structure MCTC carbon atoms; blue: docked ligand carbon atoms; red: oxygen atoms; white: hydrogen atoms. The more translucent region between subsites –2 and +2 in Figure 4.2a denotes the $H.~jecorina$ CBH active-site tunnel.
absent Trp272 analog (Figure 4.2b). The contributions of Trp272 in \textit{H. jecorina} CBH can further be explained when examining the differences between the subsite +4 $E_{\text{Inter}}$ values of $\beta$-glucose and $\text{SO}_2\beta$-Cel$_6$ docked in subsites –2 through +4. The $\beta$-glucose ligand presents its apolar $\beta$-face$^{45}$ to Trp272, with four interactions of ~3.5–4.2 Å between the C3, C4, C5, and C6 ligand atoms and the indole ring, along with hydrogen bonding interactions between the ring oxygen and guanidium protons of Arg275 and the C6 hydroxyl group and backbone carbonyl group of Gly268 (Figure 4.3a). The $\beta$-faces of aliphatic rings produce more favorable stacking interaction than $\alpha$-faces with aromatic side chains of amino acids, and the $\beta$-glucose the $\beta$-glucose mirrors the binding pattern seen with galactose in C-type lectins.$^{46}$ In contrast, the glucosyl residue in the $\text{SO}_2\beta$-Cel$_6$ ligand docked in subsite +4 presents an $\alpha$-face with two apolar interactions to Trp272 (Figure 4.3b), due to a 180º rotation from the normal cellulosic planar configuration of the glucosyl residues in subsites +3 and +4. The difference in subsite +4 $E_{\text{Inter}}$ values (–56.6 to –22.9 kcal/mol) between $\beta$-glucose and $\text{SO}_2\beta$-Cel$_6$ can be readily explained because of stacking interactions, and suggests the existence of subsite +4 as a stabilizing area for crystalline cellulose substrates as they progress in the nonreducing direction down the \textit{H. jecorina} CBH active site.$^{18}$ The lack of a more negative $E_{\text{Inter}}$ value for the glucosyl residue in subsite +4 of $\text{SO}_2\beta$-Cel$_6$ is due to the restricted freedom of rotation imposed by active-site binding from the other glucosyl residues.

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure4.3.png}
\caption{(a) Glucose docked into putative subsite +4 and (b) $\text{SO}_2\beta$-Cel$_6$ docked into putative subsites +3 and +4 of \textit{H. jecorina} CBH. Yellow numerals are distances between atoms in Å.}
\end{figure}
in the ligand. Docking β-glucose in subsite +4 was performed to examine binding interactions without the rotational restrictions imposed by a larger ligand, demonstrating the ability for subsite +4 to bind ligands on a comparable basis to subsites –2/1+1/+2.

The more negative $E_{\text{Inter}}$ values exhibited by $H. jecorina$ CBH compared to $T. fusca$ EG, both overall in nearly all corresponding ligands and also individually in subsites –2 to +2, gives evidence of the effect of binding by the extra tunnel-forming loops in the former. Furthermore, $E_{\text{Inter}}$ values of ligands bound in subsites near the cleavage point of this GH6 $H. jecorina$ CBH are slightly less negative than those of the GH7 $H. jecorina$ Cel7A CBH. Total $E_{\text{Inter}}$ values of β-glucose through β-Cel4 docked in subsites near the cleavage point of GH6 $T. fusca$ EG are slightly less negative than total $E_{\text{Inter}}$ values of the analogous ligands docked in the same subsites of GH7 $Fusarium oxysporum$ Cel7B EG.

Forces generated by the GH6 enzymes $H. jecorina$ CBH and $T. fusca$ EG with the corresponding GH7 enzymes $H. jecorina$ Cel7A and $F. oxysporum$ Cel7B shows that the much longer tunnel through which substrates travel toward the reducing end in $H. jecorina$ Cel7A leads to processive forces at least four times the magnitude of those found with GH6 $H. jecorina$ CBH. $F. oxysporum$ Cel7B has four main subsites, the same as in GH6 $T. fusca$ EG, but significantly higher processive forces of the same ligands in analogous subsites. This discrepancy is caused by the catalytic nature of each enzyme. Cel7B has a retaining mechanism that employs a double displacement of the glycosidic oxygen on the anomeric
Catalytic carboxylate groups of GHs with retaining action typically are about 5.5 Å apart, whereas an enzyme employing an inverting, single-displacement mechanism, such as GH6 *T. fusca* Cel6B, will typically have 9.5 Å separating the catalytic carboxylate groups.47,48 *F. oxysporum* Cel7B has a higher density of enzyme-substrate nonbonded interactions of less than 4.0 Å than does *T. fusca* EG (Figure 4.4). This difference in substrate stabilization leads to *F. oxysporum* Cel7B producing higher processive forces upon the ligand than does *T. fusca* EG.31

**Conclusions**

Automated docking of β-glucose in the *H. jecorina* CBH and *T. fusca* EG active sites shows that subsite –2 binds this ligand more strongly than subsites –1 and +1, which flank the cleavage point. β-Cel2 is most strongly bound in subsites –2/–1 than in subsites –1/+1 or elsewhere, suggesting why it is not a substrate. In general, glucosyl residues docked in subsite –2 have more negative E\textsubscript{Inter} values than those docked in other subsites, and ligands with their nonreducing-end glucosyl residues located in subsite –2 have more negative E\textsubscript{Inter} values than corresponding ligands docked elsewhere.

The authors are grateful to the U.S. Department of Agriculture through the Biotechnology Byproducts Consorton and to the National Science Foundation for funding this project.

**References**

T. The active site of Trichoderma reesei cellbiohydrolase II: The role of tyrosine


Drakenberg, T.; Teleman, O. Progress-curve analysis shows that glucose inhibits
the cellotriose hydrolysis catalysed by cellobiohydrolase II from Trichoderma

20. Harjunpää, V.; Teleman, A.; Ruohonen, L.; Teeri, T. T.; Teleman, O.; Drakenberg,
T. Cello-oligosaccharide hydrolysis by cellobiohydrolase II from Trichoderma
reesei. Association and rate constant derived from an analysis of progress curves.

course of the action of the celllobioside hydrolases I and II of Trichoderma reesei. J.


23. Varrot, A.; Macdonald, J. M.; Stick, R. V.; Pell, G.; Gilbert, H. J.; Davies, G. J.
Distortion of a cellbio-derived isofagomine highlights the potential conformational

24. Taylor, J. S.; Teo, B.; Wilson, D. B. Conformational modeling of substrate binding
to endocellulase E2 from Thermomonospora fusca. Protein Eng. (1995) 8, 1145–
1152.

25. André, G.; Kanchanawong, P.; Palma, R.; Cho, H.; Deng, X.; Irwin, D; Himmel, M.
E.; Wilson, D. B.; Brady, J. W. Computational and experimental studies of the
catalytic mechanism of Thermobifida fusca cellulase Cel6A (E2). Protein Eng.


K.; Olson, A. J. Automated docking using a Lamarckian Genetic Algorithm and an


Chapter 5. Analysis of Functional Divergence within Two Structurally Related Glycoside Hydrolase Families

A paper in preparation for submission
Blake Mertz,12 Xun Gu,3,4 Peter J. Reilly2,5♣∇
Departments of Chemical and Biological Engineering and Genetics, Development, and Cell Biology, Iowa State University, Ames, IA 50011

Abstract

Two glycoside hydrolase (GH) families were analyzed to detect the presence of functional divergence using the program DIVERGE. These two families, GH7 and GH16, each contain members related by amino acid sequence similarity, retaining hydrolytic mechanisms, and catalytic residue identity. GH7 and GH16 comprise GH Clan B, with a shared β-jelly roll topology and mechanism. GH7 contains fungal cellobiohydrolases and endoglucanases and is divided into five main subfamilies, four of the former and one of the latter. Cluster comparisons between three of the cellobiohydrolase subfamilies and the endoglucanase subfamily identified specific amino acid residues that play a role in the functional divergence between the two enzyme types. GH16 contains subfamilies of bacterial agarases, xyloglucosyl transferases, 1,3-β-D-glucanases, lichenases, and other enzymes with various substrate specificities and product profiles. Four cluster comparisons between these four main subfamilies again have identified amino acid residues involved in functional divergence between the subfamilies.

Keywords: cellulase; Clan B; DIVERGE; functional divergence; glycoside hydrolase

1 Primary researcher and author.
2 Department of Chemical and Biological Engineering
3 Department of Genetics, Development, and Cell Biology
4 Consulted for use of DIVERGE and to discuss simulation results.
5 Associate professor.
♣ Correspondence to: Peter J. Reilly, Department of Chemical and Biological Engineering, 2114 Sweeney Hall, Ames, IA 50011-2230. Phone: 515-294-5968; fax: 515-294-2689; e-mail: reilly@iastate.edu
Introduction

Carbohydrates are an essential part of many life processes due to their presence in glycoproteins, glycolipids, and oligo- and polysaccharides. Glycoside hydrolases (GHs) are the enzymes that cleave the glycosidic bonds of these materials, and they are classified on the web server CAZy\(^1\) into families based on their primary sequences. The > 100 GH families contain a wide variety of enzymes with different specificities toward substrate glycons, aglycons, and glycosidic bonds. Often members of the same family have different substrate specificities, yielding different products. Despite this, all members of a GH family have a common tertiary fold and hydrolytic mechanism, which is either inversion or retention of the C1 anomeric configuration of the new reducing-end residue after cleavage of the labile glycosidic bond.\(^2\) Many GH families can be aligned into clans, whose primary sequences are no longer conserved but whose common tertiary structures and configurations of both product and substrate\(^3\) suggest that members of these different families are derived from a common ancestor.\(^4\) At present primary sequences on TREMBL\(^5\) outnumber by ~100-fold tertiary structures on the Protein Data Bank (PDB).\(^6\) This difference makes structural biology impractical for elucidating the differences in substrate specificity within many GH families. Even with advanced structural biology tools, an interesting question remains: What evolutionary pathways caused differences in substrate specificity among members of an enzyme family that are otherwise closely related?

Predictive bioinformatics methods have been developed to bridge the primary sequence–tertiary structure gap and to more effectively analyze the structure–function relationships that exist for a related group of genes. Statistical and phylogenetic analysis of a gene family’s primary sequences can potentially answer this question by identifying specific amino acid residues responsible for functional divergence,\(^7\) defined as the occurrence of two closely related gene clusters possessing related but distinct functions. This hypothesis states that a gene duplication event, followed by changes in the evolutionary rate of amino acid mutations, leads to a homologous set of functionally divergent gene clusters.\(^8\) The majority of active-site residues tend to be conserved over time because mutations in those areas normally cause a loss of function. Gu and Vander Velden’s DIVERGE program,\(^9\) based on Gu and Zhang’s method using a maximum likelihood approach,\(^10\) identifies those mutations
occurring over time that allow a shift in enzyme function (e.g., substrate specificity). This method accounts for phylogenetic tree branch lengths and the number of substitutions that have occurred from the ancestral sequence.\textsuperscript{10}

Other functional divergence methodologies have been developed since Gu and Zhang’s work. Knudsen and Miyamoto developed a maximum likelihood method to identify specific amino acid residues that relies on an \textit{a priori} knowledge of functionally divergent subfamilies.\textsuperscript{11} Galtier also developed a maximum likelihood method to analyze covariation rates, similar to Knudsen’s and Miyamoto’s method.\textsuperscript{12}

Gu and Zhang’s method has been used to analyze the relationship among class I $\alpha$-1,2-mannosidases in GH47, finding functional divergence in the surface residues of enzymes residing in different organelles.\textsuperscript{13,14} This method has also been employed in a high-throughput analysis to identify functionally divergent subfamilies,\textsuperscript{15} as well as part of a combinatorial approach to help predict binding sites of hydrolase–inhibitor complexes.\textsuperscript{16} Siltberg and Liberles developed a covarian method to analyze substitution rates in nucleotides based on Gu and Zhang’s method.\textsuperscript{17}

Our objective in this article is to ascertain areas of functional divergence within gene families encoding GHs and, using crystal structures, to delineate their specific structure–function relationships. We have performed phylogenetic analysis on the GH6 gene family,\textsuperscript{18} and in this work we will extend that analysis using Gu and Zhang’s method to include the two GH families, GH7 and GH16, that comprise GH Clan B (GH-B).\textsuperscript{1} GH-B enzymes have $\beta$-jelly roll tertiary structures and use the retaining hydrolytic mechanism. However, members of the two families differ in substrate attacked and products formed. GH7 enzymes, found solely in fungi, hydrolyze cellulose, with cellobiohydrolases (CBHs) cleaving $\beta$-1,4 glycosidic bonds at the reducing end of the chain to produce $\beta$-cellobiose and endoglucanases (EGs) preferentially attacking the same bonds away from chain ends to produce $\beta$-cellooligosaccharides of intermediate lengths. Conversely, GH16 members, which occur rarely in archaea but often in bacteria, fungi, plants, and animals, act on $\beta$-1,3 and $\beta$-1,4 glycosidic bonds of many oligo- and polysaccharides throughout their chains, producing shorter $\beta$-oligosaccharides. In both families the catalytic proton donor/base and catalytic nucleophile are glutamate residues.\textsuperscript{1}
The evolutionary relationship between GH7 and GH16 enzymes in GH-B, emphasizing their differing substrate specificities and the structural topologies of the catalytic amino acid residues, has been addressed. Michel et al. hypothesized that a common ancestor split into GH7 cellulases and GH16 laminarases through gene duplication, with other GH16 enzymes emerging later.

**Computational Methods**

**Sequence analysis**

Amino acid sequences with UniProt and/or Enzyme Commission (EC) numbers were downloaded from CAZy. All sequences were reduced to their catalytic domains, as specified by GenBank or UniProt. Secondary structures (α-helices and β-strands) of those enzymes with known tertiary structures were obtained from UniProt; when not available, they were constructed by hand, using the enzymes’ tertiary PDB files. Multiple sequence alignments (MSAs) of GH7 and GH16 were constructed with ClustalX v1.83 after setting pairwise and multiple alignment gap–opening values to 30 and all other parameters to their default values. The MSAs were then manually adjusted with the Seaview (v2.0) alignment editor.

All phylogenetic work was performed with the PHYLIP (v3.66) program suite. SEQBOOT was used to create 250 bootstrapped data sets from each MSA. The bootstrapped data were then used to create maximum parsimonious (MP) and maximum likelihood (ML) phylogenetic trees in PROTPARS and PROML. Pairwise distances for the bootstrapped alignments were generated using PROTDIST and then inputted into NEIGHBOR to generate a neighbor-joining (NJ) phylogenetic tree. Consensus bootstrap trees were created for the MP, ML, and NJ datasets using CONSENSE and compared for each GH family to identify subfamilies and any discrepancies among the three trees. If discrepancies were found, the violating sequences were removed. Sequence input to PROTPARS, PROML, and NEIGHBOR was randomized. All options were set to default unless otherwise specified.

DIVERGE v1.04 was used to test for evidence of functional divergence within a GH family. The family’s MSA was inputted into DIVERGE, and a rooted NJ tree was generated using the Poisson correction distance measure. Gene clusters were chosen for each subfamily with ≥ 4 sequences (to avoid a large sampling variance) and the same substrate specificity.
and all possible paired cluster comparisons were used to generate the functional divergence coefficient ($\theta_{ml}$) and the posterior probability that a specific amino acid residue site is in a state of functional divergence ($p$) for each completely occupied MSA position. The posterior probability ratio $p^* = p/(1 - p)$ was calculated. Results were ranked with an arbitrary $p$ cutoff value of $\geq 0.5$.

Cluster comparisons displayed one of three characteristic patterns: 1) a minority of sites with significantly higher $p^*$ values (usually $> 4$), indicating that only a few sites are likely to have roles in functional divergence; 2) a majority of sites with $p^* < 1$, indicating that the compared clusters have little functional divergence; 3) a majority of sites with high $p^*$ values (usually $> 25$), indicating that the cluster pair is too dissimilar for functional divergence to have occurred. Cluster pairs displaying the first characteristic pattern were selected for further refinement of their alignment to increase the number of sites used for analysis in DIVERGE and to improve data quality.

Individual MSAs were then generated for each subfamily in a cluster comparison using ClustalX in profile alignment mode. Primary sequences with available secondary structures (see above) were used to align the remaining sequences within the cluster, setting the $\alpha$-helix and $\beta$-strand gap penalties at 5 and all other secondary structure options at their default values, respectively. Cluster pairs were then aligned to each other using ClustalX in profile alignment mode, and the generated MSAs were adjusted by hand using Seaview. The newly constructed MSAs were once again inputted into DIVERGE, using the previously described procedure. All MSAs were rendered using ESPript v2.2 server.

**Structural analysis**

Available crystal structures for each gene cluster pair were visualized using PyMOL. Catalytic amino acid residues and any sites identified as having a high probability of being functionally divergent were highlighted to examine potential structure–function relationships for a given cluster pair. When crystal structures had empty active sites, co-crystallized ligands from other crystal structures within the same subfamily were superimposed if available, using manual pair-fitting of their conserved EXDXE or EXDXXE catalytic residue motifs.
Results and discussion

GH7 functional divergence

The main difference between GH7 CBHs and EGs is that the CBHs have an active-site tunnel\textsuperscript{24} while the EGs have an open cleft caused by their shorter active-site loops\textsuperscript{25} evident in the EG sequence gaps in the GH7 MSA (Figure 5.1). There are five main GH7 subfamilies, and the six GH7 enzymes with crystal structures fall into four of them: ascomycotal sordariomycete CBHs (\textit{Hypocrea jecorina} Cel7A PDB 8CEL), ascomycotal eurotiomycete CBHs (\textit{Talaromyces emeronii} Cel7A 1Q9H), basidiomycotal agaricomycete CBHs (\textit{Phanerochaete chrysosporium} Cel7D 1Z3T), and ascomycotal sordariomycete EGs (\textit{Fusarium oxysporum} Cel7B 1OVW, \textit{Humicola insolens} Cel7B 2A39, and \textit{H. jecorina} Cel7B 1EG1) (Figure 5.2). The fifth subfamily, in which no member has a crystal structure, also contains ascomycotal sordariomycete CBHs. Our strategy with GH7 was to compare gene clusters of known differing function and to uncover less obvious amino acid residue pairs that over time have led to different substrate specificities in GH7 members. Therefore, we tested functional divergence of the three CBH subfamilies having crystal structures against the EG subfamily. We compared the three CBH crystal structures with only that of the \textit{F. oxysporum} EG, since the three EG crystal structures are very similar to each other.

Implementation of Gu and Zhang’s method used in DIVERGE produces two meaningful results: statistical testing for the presence of functional divergence between gene clusters, given by $\theta_{ml}$ (Supplementary Material), and posterior prediction of specific amino acid sites that have experienced a shift in their functional constraints, given by $p^*$.\textsuperscript{7} The term $\theta_{ml}$ is an inverse measure of the evolutionary rate correlation between GH subfamilies. Hence for GH7, when the correlation of evolutionary rates decreases between CBH and EG subfamilies (amino acid variation is different or not proportional among specific sites), $\theta_{ml}$ increases, with $0.15 < \theta_{ml} \leq 1$ being statistically significant. Every cluster comparison presented here produced a statistically significant $\theta_{ml}$ value. High $p^*$ values occur in the specific sites of an MSA where there are pronounced differences in the expected amino acid variation between subfamilies. For example, in Figure 5.1 at position 97, (\textit{H. jecorina} CBH numbering), threonine is completely conserved over the five CBH sequences, while the three EG
sequences have isoleucine, methionine, and leucine residues, indicating the likelihood of a change in function between subfamilies. The secondary structure in the MSA has two $\beta$-strands in the CBHs, absent in the EGs, that affect substrate interaction around CBH subsites.

Figure 5.1. MSA of five GH7 CBHs and three GH7 EGs, created with ClustalX v1.83 and rendered with ESPript v2.2. hjP62694: *Hypocrea jecorina* L27 Cel7A CBH (ascomycotal sordariomycete CBH subfamily 1); teQ8TL9: *Talaromyces emersonii* Cel7A CBH (ascomycotal eurotiomycete CBH subfamily); pcQ09431: *Phanerochaete chrysosporium* Cel7D CBH (basidiomycotal agaricomycete CBH subfamily); anQ9UVS8: *Aspergillus niger* CBS 513.88 CBH B (ascomycotal eurotiomycete CBH subfamily); cpQ00548: *Cryphonectria parasitica* EP155 CBH I (ascomycotal sordariomycete CBH subfamily 1); hjP07981: *H. jecorina* Cel7B EG (ascomycotal EG subfamily); hiP56680 (ascomycotal EG subfamily): *Humicola insolens* Cel7B EG (ascomycotal EG subfamily). X: pyroglutamic acid; Red: functionally divergent residues with overlap between cluster comparisons; black: completely conserved residues; boxed: residues with >0.7 similarity scores; red triangles: catalytic nucleophile ($P. chrysosporium$ numbering). Functional divergence results were obtained from MSAs of individual subfamilies. Enzyme codes are constructed with the initials of the producing organism followed by their six-character UniProt codes.
Figure 5.2. Phylogenetic tree of GH7 CBHs and EGs constructed with PHYLIP v3.66, showing the four main CBH subfamilies and the main EG subfamily. Enzyme codes are constructed with the initials of the producing organism followed by their six-character UniProt or eight-character GenBank codes. Asterisks: enzymes with crystal structures.
−5 and −4. Although higher $p^*$ values mean a higher likelihood of a specific amino acid site being responsible for functional divergence, the choice of a cutoff value is arbitrary (between 1 and 20)$^7$ and depends on several factors, such as MSA quality and the relative sizes of clusters being compared. Secondary structure can also be used as a contributing factor towards accurate $p^*$ values; α-helices and β-strands tend to be more well-conserved than loops, so functionally divergent residue pairs in loops must be considered more carefully before accepting their validity.

**Basidiomycotal CBHs vs. ascomycotal EGs.** The overall distribution of $p^*$ values is similar to the N-myc/C-myc case study conducted by Gu,$^7$ with the majority of residues having $p^* < 1.5$ and with 13 residues having $p^* > 5$ (Table 5.1). One pair, represented by Ser206 in *P. chrysosporium* CBH and Asn196 in *F. oxysporum* EG, with $p^* = 16.8$, is in the active site next to the catalytic nucleophile, Glu207/Glu197. Both residue side-chains face away from the active site, but they are coordinated by a series of hydrophobic and polar interactions that are unique to each (Figures 5.3, 5.4a). Proper nucleophile orientation is necessary for an axial attack on the C1′ carbon of the glucopyranosyl residue in subsite −1. The EG active site binds cellulose chains with less force than does the CBH active site,$^{26,27}$ allowing their freer movement, and requiring more control over the nucleophile orientation toward the scissile glycosidic bond. This is provided in *F. oxysporum* EG by potential hydrogen-bonding interactions of Asn196 with Cys215 and Lys217, which should help Asn196 to act as a lever to orient the adjacent nucleophile.

Two amino acid residue pairs with $p^* > 20$ are on the outer surface of the enzyme. One

**Table 5.1.** Functionally Divergent Residue Pairs in Four GH7 Subfamilies.

<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
<th>Secondary structure</th>
<th>$p^*_{14}$</th>
<th>$p^*_{24}$</th>
<th>$p^*_{34}$</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>P. chrysosporium</em> CBH</td>
<td>Thr20</td>
<td></td>
<td></td>
<td></td>
<td>Β-Strand</td>
<td>7.87</td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>T. emersonii</em> CBH</td>
<td>Thr19$^a$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>H. jecorina</em> CBH</td>
<td>Gly36</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>F. oxysporum</em> EG</td>
<td></td>
<td>Trp40$^a$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Ile71</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(continued)
Table 5.1. (continued)

<table>
<thead>
<tr>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
<th>Secondary structure</th>
<th>( p^*_{14} )</th>
<th>( p^*_{24} )</th>
<th>( p^*_{34} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Thr97</td>
<td>Thr97(^a)</td>
<td>Ile98</td>
<td>( \beta)-Strand/loop</td>
<td>7.21</td>
<td>15.4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Leu116</td>
<td>Ile116</td>
<td>Met119</td>
<td>Loop</td>
<td>6.19</td>
<td>5.57</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ala145</td>
<td>Glu148</td>
<td>Loop</td>
<td>8.64</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Asp147(^a)</td>
<td>Asp152</td>
<td>Loop</td>
<td>7.84</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ser206</td>
<td>Asn196</td>
<td>Loop</td>
<td>16.8</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pro226(^a)</td>
<td>Pro214</td>
<td>Loop</td>
<td>8.95</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Asp228(^a)</td>
<td>Ser216</td>
<td>Loop</td>
<td>5.43</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Met257(^a)</td>
<td>Ile246</td>
<td>Loop</td>
<td>12.8</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Glu294(^a)</td>
<td>Glu302(^a)</td>
<td>Glu282</td>
<td>( \beta)-Strand</td>
<td>66.7</td>
<td>19.6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ser321(^a)</td>
<td>Phe309</td>
<td>Loop/( \alpha)-Helix</td>
<td>12.9</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Phe326(^a)</td>
<td>Tyr314</td>
<td>( \alpha)-Helix</td>
<td>13.9</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Phe352</td>
<td>Met332(^a)</td>
<td>( \alpha)-Helix</td>
<td>2.19</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lys353</td>
<td>Gly333(^a)</td>
<td>( \alpha)-Helix</td>
<td>1.64</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Asn355</td>
<td>Gln362</td>
<td>Arg338</td>
<td>( \alpha)-Helix</td>
<td>9.58</td>
<td>24.9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Asp377</td>
<td>Gly360(^a)</td>
<td>Loop</td>
<td>24.9</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Val393</td>
<td>Ala362</td>
<td>Loop</td>
<td>1.63</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ser396</td>
<td>Pro364(^a)</td>
<td>Loop</td>
<td>2.40</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\(^a\) Residue conserved throughout its subfamily.

pair, represented by Asp377 of *P. chrysosporium* CBH/Gly360 of *F. oxysporum* EG, is in a flexible loop region between \( \beta \)-sheets of the concave face of the jelly roll, with no apparent role. The other pair, Glu294 in *P. chrysosporium* CBH, which is completely conserved in this CBH subfamily, and Glu282 in *F. oxysporum* EG, which is highly variable (Glu, Asp, Lys, Ser, Ala, Ser, or Gly) in other EGs, could directly affect the shape of the reducing end of the active-site cleft (Figures 5.3, 5.4b). The Glu294 carboxyl group in *P. chrysosporium* CBH is 2.57 Å from the Ser321 hydroxyl group, another functionally divergent residue with \( p^* > 10 \)
and a potential partner for hydrogen bond formation. This interaction may stabilize \( \alpha \)-helix 6 (1Z3T secondary structure numbering), which forms part of the CBH subsite +3. The corresponding EG glutamate residue, Glu282, is too distant from Phe309 (3.86 Å) to provide stabilizing interactions. Conversely, the same but much smaller EG \( \alpha \)-helix is held in place by a disulfide bond (Cys239/Cys315), while the CBH \( \alpha \)-helix lacks this disulfide bond stabilization despite the presence of two cysteine residues in close proximity, requiring a different stabilizing interaction to maintain its position.

Several functionally divergent residues with \( 5 < p^* < 20 \) are on the surface near the non-reducing ends of both the CBH and EG active sites. Trp40 and Tyr82 of *P. chrysosporium* CBH form stabilizing interactions in subsites –7 and –6, as noted earlier\(^\text{28}\) and when

---

**Figure 5.3.** a) *P. chrysosporium* CBH with functionally divergent residues in orange (paired with *F. oxysporum* EG). b) *T. emersonii* CBH with functionally divergent residues in blue (paired with *F. oxysporum* EG). c) *H. jecorina* CBH with functionally divergent residues in violet (paired with *F. oxysporum* EG). d) *F. oxysporum* EG with functionally divergent residues in orange (paired with *P. chrysosporium* CBH), in blue (paired with *T. emersonii* CBH), in violet (paired with *H. jecorina*), and in grey (paired with more than one CBH).
compared to *H. jecorina* CBH. The corresponding functionally divergent residues in *F. oxysporum* EG are Gly36 and Ala83, respectively, which reside in a shallower trough than the active-site cleft and lack the aromatic side-chain associated with stacking interactions (Figure 5.3). α-Helix 3 is structurally conserved in both enzymes, but half of the residues in this α-helix have $5 < p^* < 20$, which would suggest that a point mutation did not cause a loss of function, but rather contributed to the substrate stabilization (or lack thereof) needed for activity.

Ascomycotal eurotiomycete CBHs vs. ascomycotal EGs. *T. emersonii* CBH is very similar in overall structure to the *H. jecorina* and *P. chrysosporium* CBHs, but some of its loops are shorter, leading to a more open active site and to greater activity on substrates, traits commonly associated with EGs. Due to these similarities, overlap of functionally divergent residues is expected, and four *T. emersonii* CBH/*F. oxysporum* EG residue pairs with $p^* > 5$ were also found in the *P. chrysosporium* CBH/*F. oxysporum* EG cluster comparison (Table 5.1): 1) Glu302 (*T. emersonii* CBH residue numbers, unless otherwise noted) has a potential hydrogen-bonding interaction with Ser328 to stabilize α-helix 6 (1Q9H secondary structure numbering) in the reducing end of the active site; 2) both Ile116 and Gln362 lie between the β-jelly roll proximal to the catalytic nucleophile, and the role of these residues is unclear; 3) Thr97 is near subsites –5 and –6, and it interacts with one end of α-helix 3, unlike the corresponding Ile98 in the *F. oxysporum* EG, which lies in a loop peeled away from the active site (Figure 5.3). Several additional functionally divergent residues lie in the same region as Ile116 and Gln362 and are mainly polar. Although it is unclear what role these specific residues in the *T. emersonii* CBH and *F. oxysporum* EG play, their considerable surface area and polar nature may contribute to surface interactions with crystalline cellulose and the overlying water layer, which are essential to cellulose degradation.
Ascomycotal sordariomycete CBHs vs. ascomycotal EGs. *H. jecorina* CBH has longer tunnel-forming loops than the other two GH7 CBHs with crystal structures, leading to superior substrate binding.\(^{33}\) Despite these differences, only two *H. jecorina* CBH/F. *oxysporum* EG amino acid residues, Phe352/Met332 and Ser396/Pro364, have \(p^* > 2\), a much lower value than pairs between the other two CBH subfamilies and the EG subfamily (Figure 5.3). Both *H. jecorina* CBH Phe352 and *F. oxysporum* EG Met332 are hydrophobic residues near the nucleophile (4.78 Å and 3.72 Å distant, respectively), with the main difference between the two being that Phe352 can rotate and form a hydrophobic platform that interacts with the backbone of the nucleophile, providing more stability to the catalytic machinery hydrolyzing the scissile glycosidic bond (Figure 5.4c). Ser396 and Pro364 are both exposed as surface residues in their respective enzymes, but they occur in loop regions that lie distal from the active site. Although their role may be to stabilize loops that interact more directly with the substrate, their overall significance remains unclear.
Three more *H. jecorina* CBH/*F. oxysporum* EG residue pairs, Trp40/Gly36, Lys353/Gly333, and Val393/Ala362, have $1.5 < p^* < 2.0$. The first pair is also functionally divergent in the *P. chrysosporium* CBH/*F. oxysporum* cluster comparison; Trp40 is important for stacking interactions with the cellulosic substrate in subsite $-7$, and is one of the key aromatic residues that is absent from the majority of GH7 EGs (Figures 5.1, 5.3). The Lys353/Gly333 pair is adjacent to the nucleophile-interacting residues mentioned in the previous paragraph, and the two residues have widely different physicochemical properties. Lys353 in *H. jecorina* CBH is a polar residue with a large surface area exposed to solvent; the aliphatic portion of the side-chain packs $\alpha$-helix 5 due to hydrophobic interactions, while the terminal amide group is 2.86 Å distant from Thr296, setting up potential hydrogen bonding (Figure 5.4c). These combined characteristics could also help potential interactions between Phe352 and the catalytic nucleophile, Glu212. In contrast, Gly333 of *F. oxysporum* EG lies buried within the enzyme and serves only to facilitate the turn of the $\alpha$-helix. Finally, the Val393/Ala362 pair lies in a loop region distal from the reducing end of the active site, and has no apparent function.

Since the three CBHs have similar structures, the low number of residues with high $p^*$ values in the *H. jecorina* CBH/*F. oxysporum* cluster comparison seems unusual. This may be caused by structural differences of the GH7 enzymes; their overall tertiary structure is conserved, but loop lengths and conformations vary within the CBHs and between CBHs and EGs. Only three residue pairs with higher $p^*$ values in the *H. jecorina* CBH/*F. oxysporum* EG cluster comparison occur in loop regions (Table 5.1), indicating that they are so dissimilar that it is almost certain that point mutations there could not have caused enzyme function to diverge over time.

**GH16 functional divergence**

The GH16 family contains xyloglucan:xyloglucosyl transferases (EC 2.4.1.207), endo-1,3(4)-β-glucanases (EC 3.2.1.6), glucan endo-1,3-β-D-glucanases (EC 3.2.1.39), lichenases (1,3-1,4-β-D-glucan 4-glucanohydrolases) (EC 3.2.1.73), β-agarases (EC 3.2.1.81), κ-carrageenases (EC 3.2.1.83), keratan sulfate endo-1,4-β-galactosidases (EC 3.2.1.103), and xyloglucan endo-β-1,4-glucanases (EC 3.2.1.151). A GH16 MSA and phylogenetic tree fol-
low the general features delineated by Barbeyron et al., with four main subfamilies: 1) bacterial β-agarases; 2) plant xyloglucosyl transferases; 3) bacterial glucan endo-1,3-β-D-glucanases; and 4) bacterial lichenases (Figures 5.5, 5.6).

Despite low sequence similarity within GH16, many active-site residues are highly conserved, making identification of functionally divergent active site residues possible. There are two crystal structures each in the first two subfamilies, one in the third, and three in the fourth. The bacterial β-agarase subfamily has two Zobellia galactanivorans Dsij β-agarase (bacterial β-agarase subfamily); paQ59078: Pseudoalteromonas atlantica ATCC 19262 β-agarase (β-agarase subfamily); ptQ8GD5: Populus tremula x Populus tremuloides xyloglucan endo-transglycosylase/hydrolase (plant xyloglucosyl transferase subfamily); tmQ07524: Tropaeolum majus xyloglucan endotransferase 1/xyloglucanase (plant xyloglucosyl transferase subfamily); nsBAE54302: Nocardioopsis sp. F96 β-1,3-glucanase (bacterial glucan endo-1,3-β-D-glucanase subfamily); pfO73951: archaeal Pyrococcus furiosus DSM 3638 β-1,3-glucanase (bacterial glucan endo-1,3-β-D-glucanase subfamily). Red: functionally divergent residues with overlap between cluster comparisons; black: conserved residues; boxed: residues with >0.7 similarity scores; red triangles: catalytic nucleophile (Glu105) and proton donor/base (Glu109).
Table 5.2 Functionally Divergent Residue Pairs in Four GH16 Subfamilies

<table>
<thead>
<tr>
<th></th>
<th>Z. galacta nivorans β-agarase</th>
<th>P. tremula XTH</th>
<th>Nocardiosis sp. glucan-1,3-β-glucanase</th>
<th>B. licheniformis lichenase</th>
<th>Secondary structure</th>
<th>p*&lt;sub&gt;14&lt;/sub&gt;</th>
<th>p*&lt;sub&gt;23&lt;/sub&gt;</th>
<th>p*&lt;sub&gt;24&lt;/sub&gt;</th>
<th>p*&lt;sub&gt;34&lt;/sub&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ser45</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Asp46</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Asn56</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Leu3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Val7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lys66</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pro76&lt;sup&gt;a&lt;/sup&gt;</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ile109</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gln117</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Thr139</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Thr173</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gly109&lt;sup&gt;a&lt;/sup&gt;</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gln115</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Arg116</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Leu119&lt;sup&gt;a&lt;/sup&gt;</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Trp120&lt;sup&gt;a&lt;/sup&gt;</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Asp122&lt;sup&gt;a&lt;/sup&gt;</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(continued)
subfamily used the *Bacillus licheniformis* BgI crystal structure (1GBG), since the *Paenibacillus macerans* Bgi lichenase structure is a hybrid and *Fibrobacter succinogenes* lichenase has a catalytic domain reversed from all other GH16 1,3-1,4-β-D-glucanases.

There are six possible cluster comparisons of the four main GH16 subfamilies. However, two of them, β-agarases vs. bacterial glucan endo-1,3-1,4-β-D-glucanases and plant xyloglucosyl transferases vs. bacterial lichenases, yielded statistically poor data and will not be further considered.

**Bacterial β-agarases vs. bacterial lichenases.** The cluster comparison between the β-agarase and lichenase subfamilies, with comparisons of the *Z. galactanivorans* β-agarase and *B. licheniformis* lichenase crystal structures, respectively, yields a normal distribution of $p^*$ values, twelve residue pairs having $p^* > 3.5$ (Table 5.2). Functional divergence data must be interpreted carefully, since there are only five proteins in the β-agarase subfamily, two of

<table>
<thead>
<tr>
<th>(1) Z. galactanivorans β-agarase</th>
<th>(2) P. tremula XTH</th>
<th>(3) <em>P. aeruginosa</em> sp. glucan-1,3-β-glucanase</th>
<th>(4) <em>B. licheniformis</em> lichenase</th>
<th>Secondary structure</th>
<th>$p^*_{14}$</th>
<th>$p^*_{23}$</th>
<th>$p^*_{24}$</th>
<th>$p^*_{34}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pro123&quot;</td>
<td>Gly161</td>
<td>Asp139</td>
<td>Loop</td>
<td>5.17</td>
<td>4.06</td>
<td>10.6</td>
<td>699</td>
<td>16.1</td>
</tr>
<tr>
<td>Leu142</td>
<td>Ala165</td>
<td>Ala140</td>
<td>α-Helix (1)</td>
<td>63.7</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pro147&quot;</td>
<td>Ala188</td>
<td>Tyr159</td>
<td>α-Helix (3)</td>
<td>63.7</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Asn153</td>
<td>Ser194</td>
<td>Leu164</td>
<td>β-Strand</td>
<td>63.7</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ile169</td>
<td>Met210</td>
<td>Asp205</td>
<td>Loop (4)</td>
<td>63.7</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Glu267&quot;</td>
<td>Ala180&quot;</td>
<td>Gln174</td>
<td>Loop</td>
<td>63.7</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ala180&quot;</td>
<td>Glu191</td>
<td>Asp192</td>
<td>Loop (2)</td>
<td>63.7</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Thr181&quot;</td>
<td>Thr170</td>
<td>Leu202</td>
<td>Loop (2)</td>
<td>63.7</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gly184&quot;</td>
<td>Asp205</td>
<td>Gln174</td>
<td>Loop (2)</td>
<td>63.7</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Arg282&quot;</td>
<td>Arg210</td>
<td>Asp237&quot;</td>
<td>Loop (2)</td>
<td>63.7</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*a* Residue conserved throughout its subfamily.

...
them from the same organism. Better alignment of $\alpha$-helix and $\beta$-strand regions occurred in the C-terminal half of the primary sequences, and this is reflected in the higher $p^*$ values found in the N-terminal region (Figure 5.5). Three pairs (Thr139/Thr93, Glu267/Glu191, and Arg282/Arg210) have identical residues, but in each case, the residue is conserved within only one subfamily (Table 5.2). Such evolutionary variation at a specific site following gene duplication could result from either a relaxation of functional constraint (random mutations

![Figure 5.7.](image)

**Figure 5.7.** a) *Z. galactanivorans* $\beta$-agarase with functionally divergent residues in green (paired with *B. licheniformis* lichenase). b) *P. tremula* XTH with functionally divergent residues in orange (paired with *Nocardiosis* sp. glucan-1,3-$\beta$-glucosidase), in magenta (paired with *B. licheniformis* lichenase), and in grey (paired with each). c) *Nocardiosis* sp. glucan-1,3-$\beta$-glucosidase with functionally divergent residues in orange (paired with *Z. galactanivorans* $\beta$-agarase), in blue (paired with *B. licheniformis*), and in grey (paired with each). d) *B. licheniformis* lichenase with functionally divergent residues in green (paired with *Z. galactanivorans* $\beta$-agarase), in magenta (paired with *P. tremula* XTH), in blue (paired with *Nocardiosis* sp.), and in grey (paired with at least two). Partially obscured red sphere in a), c), and d): $Ca^{2+}$ ion.
Figure 5.8. Close-up views of GH16 enzymes. a) *B. licheniformis* lichenase, showing hydrogen bond between functionally divergent residue Glu63 and the O6 hydroxyl group of the middle glucopyranosyl residue of the β-1,3-1,4-glucotriose ligand. b) *Nocardiopsis* sp. glucan-1,3-β-glucanase, showing six functionally divergent residues and the coordinating interactions between Asp237 and the Ca\(^{2+}\) ion of the enzyme. c) *P. tremula* XTH, showing functionally divergent and fully conserved Leu119–Pro123 loop, hydrogen bonding between Thr181 and Gly92, and potential stabilizing interactions between Pro147 and the XTH α-helix. d) *P. tremula* XTH, showing hydrogen bonds between functionally divergent residue Arg116 and the co-crystallized ligand XLLG. e) *Nocardiopsis* sp. glucan-1,3-β-glucanase, showing residue Asn39 hydrogen-bonding the glycosidic oxygen connecting the glucopyranosyl residues of the superimposed ligand in subsites −2/−1. Distances in Å.
with no effect on the original enzyme function) or positive selection (gain of a functional role in a previously unimportant residue). Past research led to the hypothesis that diverse substrate specificity in GH16 enzymes was caused by a gene duplication event, and these three functionally divergent pairs lend positive support to this.

Three residue pairs, Ile109/Glu63, Pro76/Asn37, and Thr173/Asn121 in the β-agarase and lichenase subfamilies, respectively, the latter two pairs being aligned in cluster comparisons (Supplementary Material) but not in the overall MSA (Figure 5.5), are near the active site (Figure 5.7a,d). Although both enzymes have open-cleft active sites, β-agarases have a more extensive binding network than lichenases, with eight subsites rather than six. The most noticeable difference between the two subfamilies is the nonbonded interaction with the ligand in subsite –2. The lichenase Glu63 carboxyl group lies ~2.6 Å from the C6 hydroxyl group of the middle glucopyranosyl residue of the β-1,3-1,4-glucotriose ligand in the B. licheniformis lichenase active site (Figure 5.8a), whereas the β-agarase Ile109 is > 9 Å from the co-crystallized oligoagarose ligand in the Z. galactanovorans β-agarase active site.

Both the β-agarase Pro76 and the lichenase Asn37 lie too far from the active site to appreciably interact with the substrate, but each α-helix in which they reside forms an integral part of the nonreducing end of their respective active sites (Figure 5.7). Asn37 is the middle residue in an α-helix that lies at the beginning of a loop that interacts with the nonreducing end of the oligoagarose ligand, and it may have an effect on the loop conformation. Pro76, on the other hand, produces unique conformations within protein structures due to the nature of its backbone, and it could influence the larger nonreducing-end loop, which appears to pinch off that end of the active site by folding over the cleft region.

The final active-site pair, Thr173/Asn121, is at the reducing end of the active site, and it may indicate how lichenases compensate for a shorter active site with more advantageous ligand binding. Even though there is no ligand present, the lichenase Asn121 is placed for favorable hydrogen-bonding interactions with a potential substrate; similar modeling studies on the circularly permuted GH16 F. succinogenes lichenase demonstrated that its corresponding residue, Gln81, forms hydrogen bonding contacts with the C3 hydroxyl group of the glucopyranosyl residue in the subsite +2 (Figure 5.7). The other residue in this functionally divergent pair is the β-agarase Thr173, and despite being in the same location as
Asn121, its side-chain faces away from the active site. This may have been an adaptation towards the bulkier, hydrophobic nature of the 3,6-anhydrogalactopyranosyl units present in every second residue in agarose, as well as the result of having a longer active site with more stacking interactions, leading to less selective pressure for hydrogen bonding around subsite +2. One interesting point to note is the conservation of specific surface binding sites in the agarase; Allouch et al. demonstrated in a second crystal structure of β-agarase A that a surface-binding site exists for an agarose substrate. None of the residues in this site are functionally divergent, indicating that this particular surface binding is a function complementary to β-1,4-bond hydrolysis.

**Plant xyloglucosyl transferases vs. bacterial glucan endo-1,3-β-D-glucanases.**

Xyloglucosyl transferases modify plant cell-wall xyloglucans by either hydrolyzing or transferring single xylopyranosyl residues to another polymer chain. The most common substrate of endo-1,3-β-glucanases is the β-1,3-glucan laminarin, in which they hydrolyze only internal β-1,3 glycosidic linkages. However, they also hydrolyze mixed linked β-1,3-1,4-glucans such as lichenan. The only available endo-1,3-β-glucanase tertiary structure is from *Nocardiopsis sp. F96* (2HYK), which exhibits an uncharacteristic preference for mixed linked β-1,3-1,4-glucans (typically lichenin).

Eighteen amino acid residue pairs with $p^* > 20$ were identified (Table 5.2). As with other GH16 cluster comparisons, the N-terminal regions in the primary sequences of each subfamily are poorly aligned; in this case, consistent secondary structure agreement between *P. tremula* XTH and *Nocardiopsis* glucanase did not occur until the second major β-strand of each protein (XTH residues 51–61 and glucanase residues 88–95), suggesting that $p^* = 2070$ for the Gly44/Gly71 pair is a false positive. Another factor to consider with interpretation of the $p^*$ data is the difference in size between the two subfamilies; the xyloglucosyl transferase subfamily contained 25 members, whereas the glucanase subfamily has nine members, which could lead to a slight bias favoring the larger subfamily. The evolutionary rates being used to represent each cluster are based on the average rate in each subfamily; when these two values are combined to calculate $p$ at a specific residue site, more weight will be given to the larger cluster.

Most functionally divergent residues are located outside the active site on both ends of
the binding cleft (Figure 5.7). One noticeable characteristic is the prevalence of divergent residues near the *Nocardiopsis* glucanase Ca\(^{2+}\) ion, which increases structural stability of other GH16 glucanases.\(^{44,45}\) One of the three residues interacting with Ca\(^{2+}\) in a pentagonal-bipyramidal geometrical arrangement, Asp237, is functionally divergent from the XTH Arg200 (Figure 5.8b). It was thought that Ca\(^{2+}\) coordination was completely conserved in GH16;\(^4\) however, the two xyloglucosyl transferase tertiary structures do not have Ca\(^{2+}\) ions.\(^{36,46}\) The large number of functionally divergent residues (Tyr45, Thr46, Leu60, Ile62, Asp237, Tyr238, Arg240, Val241, and Tyr242) proximal to the *Nocardiopsis* glucanase Ca\(^{2+}\) ion and the nonreducing end of the active site indicates that the absence of Ca\(^{2+}\) ions to maintain structural stability in xyloglucosyl transferases has significantly contributed to different substrate specificities in the two subfamilies.

The Leu119/Met157 pair has a very high \(p^*\) value (Table 5.2). It occurs at the flexible loop connecting \(\beta\)-strands at the reducing end of the *P. tremula* XTH active site and at the C-terminal end of the last *Nocardiopsis* glucanase \(\beta\)-strand (Figure 5.8c). Three of the next four residues are also functionally divergent (28 < \(p^*\) < 64), with the Leu119–Trp120–Phe121–Asp122–Pro123 loop in the xyloglucosyl transferase subfamily being completely conserved. Pro159 is the only residue conserved in the corresponding glucanase loop, indicating that this loop is essential in the xyloglucosyl transferase subfamily but not in the glucanase subfamily. Since the xyloglucosyl transferase loop is outside the active site, it is unclear what role it plays in enzyme function, but it appears to have much more flexibility than in *Nocardiopsis* glucanase, with a shorter preceding \(\beta\)-strand and no following \(\alpha\)-helix as in the glucanase (Figure 5.7). One possible role for this loop could be to help maintain proper orientation between the \(\beta\)-strands of each face at the reducing end of the active site.

Two other residue pairs have \(p^* > 100\), and both involve completely conserved residues in the xyloglucosyl transferase subfamily and highly variable residues in the glucanase subfamily, suggesting a conserved functional role for the xyloglucosyl transferase residues. The first pair, Pro147/Ala188, is located in the convex \(\beta\)-sheet of the \(\beta\)-jelly roll, underneath the reducing end of the active site. The most noticeable structural difference is that Pro147 lies near a large XTH \(\alpha\)-helix, present only in xyloglucosyl transferase (Figure 5.8c). Pro147 could aid \(\alpha\)-helix orientation, although the unique physicochemical properties of proline
residues could lead to a different function of this residue.

The second high-$p^*$ pair, Thr181/Asp219 again aligned in the cluster comparison (Supplementary Material) but not in Figure 5.5, is in a flexible loop that extends over the non-reducing end of the active site. Despite being too distant from the substrate for significant non-bonded interactions, Thr181 interacts with several residues important to substrate binding, especially Trp179, which provides a hydrophobic platform to stabilize the ring faces of the substrate glucopyranosyl backbone through stacking interactions. Through potential hydrogen bonding with Gly92, which is completely conserved in the xyloglucosyl transferase subfamily, Thr181 helps to maintain Trp179 in the proper orientation for stacking to occur (Figure 5.8c). This stabilizing mechanism may have developed differently from that of the glucanases; most of the residues surrounding Asp219 are well-conserved, with Trp220 providing stacking interactions in the Nocardiosis glucanase. Three to four glycine residues are also present in the glucanase loop, lending added conformational flexibility, and this may help explain why endo-1,3-β-D-glucanases hydrolyze a smaller substrate than do xyloglucosyl transferases.

*Plant xyloglucosyl transferases vs. bacterial lichenases.* Eighteen xyloglucosyl transferase/lichenase residue pairs with $p^* > 10$ occur, of which half are found in β-strands (Table 5.2). As with the other cluster comparison, the N-terminal end of the MSA has poorly aligned secondary structures, with high variability in the xyloglucosyl transferase subfamily until the Glu32-Leu37 β-strand (Supplementary Data). The N-terminal region consists mainly of α-helices, short β-strands, and flexible loops forming the lip in the non-reducing end of the active site and around the lichenase Ca$^{2+}$. Therefore our discussion of the results begins from the Asp38/Thr52 pair, proceeding in the C-terminal direction.

The most notable active-site divergent residue pair is Arg116/Ile133, with Arg116 potentially forming three hydrogen bonds < 4.4 Å from the O5 atom and the C1 and C4 hydroxyl groups of the xylopyranosyl residue α-(1,6)-linked to the glucopyranosyl residue in subsite +2. Conversely, the corresponding Ile133 lies along the rim of the active-site cleft, a sufficient distance to prevent meaningful interactions with a lichenin substrate (Figure 5.8d). A likely role for lichenase Ile133 is to maintain structure at the active-site reducing end; a significant portion of its hydrophobic side-chain is exposed to solution (Figure 5.7). Only one
other active-site interaction, Gln79/Gly96, is present, but its high $p^*$ value indicates an important difference between the two residues. Loop conformations will often dictate the quality of substrate binding in GH families, such as in GH6$^{18}$ and GH7.$^{24,25}$ The XTH Gln79 loop is much longer than the lichenase Gly96 loop and is slightly folded back, presumably to accommodate the bulkier xyloglucan substrate. A wider active site cleft would also require more residues to provide binding interactions with the substrate, and Gln79 could potentially provide those interactions in subsites $-2/-3$ (Figure 5.7).

Previous crystallographic studies of the $P.\ tremula$ and $Trapuleum\ majus$ xyloglucosyl transferases outlined how the differing active-site widths could lead to their unusual substrate specificities.$^{36,46}$ However, most of the functionally divergent residues in the xyloglucan transferase/lichenase cluster comparison lie outside the active site, with those in lichenases having much more of their side-chains exposed to solution than those in xyloglucan transferases. Several bacterial species produce cellulosomes, assemblies of various enzymes that act synergistically to degrade biomass, and several GH16 glucanases/lichenases are part of these complexes, specifically in $Ruminococcus$ bacteria.$^{47}$ The larger number of solution-exposed lichenase residues than xyloglucan transferase residues could indicate that this function developed in the lichenases to improve interactions with other cellulosome components and to increase efficiency.

**Bacterial glucan endo-$1,3$-$\beta$-$D$-glucanases vs. bacterial lichenases.** The final cluster comparison for GH16 enzymes involved the bacterial endo-$1,3$-$\beta$-glucanase and bacterial lichenase subfamilies. The main difference between glucanases and lichenases is that glucanases cleave $1,3$-$\beta$-D-glucosidic linkages in $1,3$-$\beta$-glucans, whereas lichenases cleave $1,4$-$\beta$-glucosidic bonds in mixed $1,3$-$1,4$-$\beta$-glucans. Structural mapping should delineate several similarities between the two subfamilies, since $Nocardiopsis$ glucanase has significant lichenase activity.$^{43}$

Eleven residue pairs were identified with $p^* > 4$, and the distribution of $p^*$ values is normal, with the vast majority in the lowest grouping. Two of the eleven pairs are located in the active site, with only one of these pairs having significant substrate interactions. Both $Nocardiopsis$ sp. glucanase Met210 and $B.\ licheniformis$ lichenase Ile179 are in the nonreducing end of the active site with their side-chains facing away from the substrate, and
would presumably have no effect on enzyme function. However, both Asn39 of the glucanase and Asn28 in the lichenase (Supplementary Material) are located in the long loop forming one side of the active-site cleft, and with the flexibility typical of loop dynamics, they could conceivably form a hydrogen bond with the substrate (Figure 5.7c,d). The more plausible candidate for this is *Nocardiopsis* Asn39, which resides 2.92 Å from the glycosidic oxygen connecting the glucopyranosyl residues of the superimposed ligand in subsites −2/−1 (Figure 5.8e); this asparagine residue is completely conserved in its subfamily except for one member, indicating an essential role for glucanase function (Figure 5.5 and Supplemental material). The lichenase Asn28 has a less favorable interaction, lying ~5 Å from the substrate. Ironically, the lichenase subfamily has another asparagine residue (Asn26) that is completely conserved and can potentially hydrogen-bond with the C6 hydroxyl group of the ligand in subsite −2. The corresponding residue in the glucanases is also conserved, but it is a hydrophobic substrate-stabilizing tryptophan. This difference in active site arrangement supports the final conclusion in Fibriansah et al.\textsuperscript{42} about endo-1,3-β-glucanases: substrate binding in subsites −2/−1/+1 is much tighter than in lichenases, and it is required to successfully hydrolyze a nonlinear 1,3-β-glucan such as laminarin compared to the more linear lichenan. Asn39 could be one of the sites that was conserved in 1,3-β-glucanases as the lichenase subfamily evolved to hydrolyze lichenan, grain endosperm cell walls, and similar substrates.\textsuperscript{4}

Each enzyme’s N-terminal functionally divergent residues have three factors in common: 1) significant surface exposure; 2) proximity to the Ca\textsuperscript{2+} ion; and 3) conformational effects on substrate binding loops in the active-site nonreducing end. As mentioned earlier, surface interactions in GHs can aid either in stabilizing substrates, as with β-agarases,\textsuperscript{40} or in interacting with cellulosomal components,\textsuperscript{47} and both glucanase and lichenase surfaces could have evolved to adopt these functions. Ca\textsuperscript{2+} coordination is necessary for protein stability in the majority of GH16 subfamilies, but the high degree of amino acid variability in these regions across all subfamilies suggests that several combinations exist to accommodate Ca\textsuperscript{2+} interaction. The Arg35/Thr46 pair is part of an extensive loop that helps form subsite −2/−1 in the glucanases and lichenases. As stated in the previous paragraph, the *Nocardiopsis* sp. glucanase loop extends further into the active site than in the *B. licheniformis* lichenase, and
Thr46 may help maintain this loop conformation due its backbone inflexibility compared to that of arginine.

Conclusions
This article describes computational analysis of the structurally similar glycoside hydrolase families 7 and 16 to identify functionally divergent amino acid residues by analysis of amino acid sequences. These two families share similar tertiary structures and mechanisms, but they are active on different substrates and produce different products. Furthermore, GH7 subfamilies produce different products from the same substrate, while GH16 subfamilies attack different substrates. Here the program DIVERGE was successfully used to compare three CBH subfamilies individually with an EG subfamily in GH7, while in GH16 four comparisons of different subfamily pairs were conducted. Corresponding residues that vary between subfamilies through mutation were identified, and in many cases their roles in functional divergence were specified.

This project was funded by the National Research Initiative of the U.S. Department of Agriculture Cooperative State Research, Education and Extension Service through contract 2007-35504-18252. The authors are grateful for this assistance.

References

4. Michel, G., Chantalat, L., Duee, E., Barbeyron, T., Henrissat, B., Kloareg, B.,


Chapter 6. General Conclusions

General discussion

Future work will have a two-fold focus: 1) using molecular dynamics (MD) and quantum mechanics/molecular mechanics (QM/MM) as tools to investigate the reaction mechanisms of various glycoside hydrolase (GH) families; and 2) using MD simulations to test the validity of functional divergence results. Chapters 3 and 5 demonstrated that phylogeny, and specifically functional divergence, can be used to help understand the fundamental relationships between related enzyme families; namely, that the differences between CBHs, EGs, and GH6 enzymes with both CBH and EG activities are readily seen in both their primary and tertiary structures, and that residues outside the active site may play an important role in the functional evolution of different cellulases. I was able to prove in Chapter 4 that docking could be used to verify the presence of putative subsites in GH6 enzymes and to explain how the presence, or lack of it, of stabilization through aliphatic non-bonded stacking interactions between the substrate and enzyme helps to facilitate the point of attack on cellulose for CBHs and EGs.

Recommendations for future research

Transition-state pathways of glycoside hydrolases

Although GH mechanisms are characterized as producing inversion or retention of the anomeric configuration, many remain ambiguous, either because 1) catalytic residues have been difficult to identify, or 2) wide substrate specificity introduces slight mechanistic differences. Hence, several GH families are promising targets for mechanism elucidation through the use of computational simulation. All of the proposed docking studies have followed or will follow the protocols previously developed in our research group. MD simulations will use the program NAMD and QM/MM simulations will use the program CPMD to model the glycosylation and deglycosylation steps in the retaining mechanism, events that require the formation and cleavage of bonds and which are unattainable with MD or docking.

The GH1 family is one of the largest among GHs, with over 1300 entries in CAZy, as well as 22 published crystal structures. GH1 enzymes use a retaining mechanism to cleave a
wide variety of disaccharides into monosaccharides, and the ability of GH1 enzymes to 
hydrolyze multiple O-glycosides (and in the case of myrosinases, S-glycosides) makes them 
excellent candidates for functional divergence and transition-state analysis. Our group has 
already investigated the role of active-site residues in substrate specificity for sixteen GH1 
enzymes using docking simulations, and another docking study was undertaken by Mitchell 
Anderson (a Clarkson University student participating in the Research Experiences for 
Undergraduates (REU) program in this department during Summer 2007) under my guidance 
to investigate the GH1 transition-state pathway.

In this study, the crystal structure from the archaeabacterium Sulfolobus solfataricus P2 
(PDB 1UWT and 1UWU) was chosen because of its increased activity under extreme condit-
ions, similar kinetics and structural recognition of gluco- and galacto-configured substrates, 
and oxocarbenium ion-like conformation of the co-crystallized ligands. Cellobiose and 
glucopyranose ligands, with the nonreducing-end residue in each of the 38 ring conforma-
tions on the pseudo-rotational itinerary, were constructed for docking into the 1UWT active 
site. Approximately half of the cellobiose ligands were correctly put together by Anderson, 
with the rest constructed by me. Currently, docking simulations have been completed for 
cellobiose only, and the results remain to be analyzed. Preliminary data shows that and 
conformers, similar to the conformation of the crystallized ligand, bind most tightly. To 
complete this project, docking studies with the glucopyranosyl ligands as well as QM/MM 
simulations of both glucopyranosyl and cellobiose ligands need to be performed. Since GH1 
enzymes hydrolyze glycosidic bonds with a retaining mechanism, two steps, glycosylation 
and deglycosylation, occur; molecular mechanics programs like AutoDock and MD 
programs like NAMD are incapable of modeling bond cleavage and formation, so CPMD 
will be used to identify the important active-site motions that occur during the glycosylation 
and deglycosylation reaction steps.

GH7 enzymes are widely used for industrial purposes, and previous docking studies in 
our group demonstrated that GH7 CBHs bind cellulose more tightly than EGs, with force 
analysis revealing that the long active-site tunnel of CBHs acts as a molecular machine that 
forces the cellulose chain in a processive manner, increasing enzymatic efficiency. In an 
extension of this work and our previous work on the GH47 transition-state pathway,
undergraduate Scott Munhall constructed glucopyranose and cellobiose ligands with the nonreducing-end residue in all possible ring conformations and docked them into the active site of *H. jecorina* CBH (PDB 8CEL), using the local search algorithm of AutoDock. Preliminary results indicate that the conformational pathway to the transition state passes through $^{1}S_{3}$ on the way to $^{4}E$, but because GH7 enzymes have a retaining mechanism, QM/MM must also be used to identify the glycosylation and deglycosylation transition states.

GH43 is part of clan GH-F, whose members share a unique five-bladed propeller tertiary fold. Among the enzymes in GH43 are β-xylosidases, and these enzymes are distinct from other xylosidases because they alone use an inverting hydrolytic mechanism. Until recently, *Geobacillus stearothermophilus* T-6 XynB3 had the only published crystal structure of a GH43 xylosidase, and with a xylobiose (β-D-xylopyranosyl-1,4-β-D-xylopyranoside) ligand co-crystallized into the active site of a XynB3 mutant, it provided an excellent template for docking studies. Dylan Murray, another ISU undergraduate in our group, spent the 2007–08 academic year constructing the 38 nonreducing-end xylobiose conformers and docking them into the xylosidase (PDB 2EXH) active site. Since the xylobiose dockings will only provide the transition-state pathway before hydrolysis, a separate set of dockings must be performed with xylopyranose. MD simulations will also be required to verify the putative transition state, since Murray’s preliminary data indicated a $^{2,5}B$ transition state, which does not lie in the pathway of the $^{0}E$ conformation of the co-crystallized ligand. Previous work on GH94 enzymes will serve as the model for applying MD techniques to the GH43 xylosidases.

**Extension of functional divergence applications to cellulases**

Functional divergence analysis has been applied to clan GH-A cellulases to generate preliminary results. Further analysis will entail multiple sequence alignment (MSA) construction of specific cluster comparisons, and implementation of site-directed mutagenesis into MD simulations to examine the effects on enzyme stability and substrate specificity. Procedures and data will be presented in Appendix A.
References


Appendix A. Functional divergence in glycoside hydrolase Clan A enzymes that attack cellulosic complexes (a partially completed manuscript)

Blake Mertz,¹ Xun Gu,² Peter J. Reilly¹
Departments of ¹Chemical and Biological Engineering and ²Genetics, Development, and Cell Biology, Iowa State University, Ames, IA 50011

ABSTRACT:

Keywords: cellulase; Clan A; DIVERGE; functional divergence; glycoside hydrolase

Correspondence to: Peter J. Reilly, Department of Chemical and Biological Engineering, 2114 Sweeney Hall, Iowa State University, Ames, IA 50011-2230. Phone: 515-294-5968; fax: 515-294-2689; e-mail: reilly@iastate.edu

Contract grant sponsor: National Research Initiative of the U.S. Department of Agriculture Cooperative State Research, Education and Extension Service
Contract grant number: 2007-35504-18252

Introduction

Glycoside hydrolases (GHs) are a major family of enzymes that cleave glycosidic bonds between a carbohydrate glycon and an aglycon that usually is also a carbohydrate. There are > 100 GH families, each of which contains enzymes with significant amino acid sequence similarity.¹ Family members have similar tertiary structures and identical hydrolytic mechanisms, either retaining or inverting configuration of the newly produced anomeric hydroxyl group from the configuration of the scissile glycosidic bond, depending on the family. However, family members often exhibit functional divergence, differing in the substrates that they preferentially attack and in the products that they form.

Some GH families are linked to others in clans by the similar mechanisms and tertiary structures of their members,¹ indicating that they are descended from a common protein ancestor and therefore a single gene. However, that ancestor is so distant that little or no primary sequence similarity remains. At present, 48 GH families are part of fourteen clans, of which the largest is Clan A (GH-A), containing seventeen families. Members of all these
families in GH-A cleave glycosidic bonds by the retaining mechanism and have similar
(\(\beta,\alpha\))₈ tertiary structures.

Six GH-A families, GH1, GH5, GH10, GH26, GH39, and GH51, contain enzymes that
primarily attack carbohydrate components of the cellulosic complex, not only cellulose and
its hydrolytic products but also different forms of xylan and other hemicelluloses. These en-
zymes are of high interest not only for their potential in producing glucose and other fermente-
table monosaccharides, but also because they are prime examples of divergent evolution. In
response to the development by plants of cellulosic complexes highly resistant to enzymatic
and physical degradation, GHs active on the cellulosic complex exhibit high functional diver-
gence, not only within GH-A but over many GH clans.

We and others have used computational means to study functional divergence within en-
zyme families. Because far more primary sequences than tertiary structures of these enzymes
have been determined, it is necessary that techniques in this area be concentrated on the for-
mer, although it is desirable to have at least one tertiary structure of a family member. Spec-
ifically, some of those amino acid residues that vary among family members are linked by
statistical and phylogenetic analysis to responsibility for functional divergence,²,³ where
closely related gene clusters give rise to different functions. Gu and Zhang have used a maxi-
mum likelihood approach, considering phylogenetic tree branch lengths and the number of
substitutions that have occurred from the ancestral sequence, to identify specific amino acid
residues responsible for functional divergence after gene duplication.⁴ Other maximum
likelihood methods have been developed by Knudsen and Miyamoto⁵ and by Galtier.⁶ A
fundamentally different approach was advanced by Hill and Reilly, who identified active-site
residues in GH1 members contributing to functional divergence by automated docking and
statistical methods.⁷

Gu and Zhang’s method has been used to analyze the relationship between class I \(\alpha\)-1,2-
mannosidases in GH47, whose members found in the endoplasmic reticulum are more spec-
ific in the glycosidic bonds that they cleave than those found in the Golgi apparatus.⁸,⁹
Abhiman et al. employed this method in a high-throughput analysis to identify functionally
divergent subfamilies,¹⁰ and Sen et al. used it as part of a combinatorial approach to help
predict hydrolase–inhibitor binding sites.¹¹ Also, a covarian method used it to analyze in
nucleotide substitution rates, specifically using the program DIVERGE. Finally, we have employed DIVERGE and other computational tools to study functional divergence in the GH7 and GH16 families, both part of Clan GH-B.

We present in this article our effort to understand functional divergence in the six families in Clan GH-A that attack components of the cellulosic complex. This new information will also contribute to our understanding of the cellulolytic degradation process and will provide new avenues to make this degradation more commercially viable.

**Computational methods**

We used computational methods identical to those that we used earlier. In summary, after downloading primary sequences and tertiary structures, we constructed multiple sequence alignments (MSAs) of the former. We constructed 250 bootstrapped data sets from each MSA. This led to maximum parsimonious, maximum likelihood, and neighbor-joining phylogenetic trees. We produced consensus bootstrap trees from these to identify subfamilies. DIVERGE was employed to determine whether functional divergence existed in individual families, and a rooted neighbor-joining tree was generated. This allowed gene clusters coding for enzymes with the same substrate specificity to be chosen, and this yielded posterior probability ratios for each amino acid residue site. Gene clusters with moderate values of posterior probability ratios were chosen, and the catalytic residues of cluster pairs and other sites likely to be functionally divergent were inspected, using available tertiary structures. Molecular dynamics was employed to determine putative effects on enzyme–substrate binding and large-scale movements of virtual mutations at identified sites.

**Results and Discussion**

**GH1 results.** Retaining mechanism. β-glucosidases, β-galactosidases, β-mannosidases, β-glucuronidases, β-D-fucosidases, β-primerverosidases, hydroxyisourate hydrolases. 89 sequences.
**GH5 results.** Retaining mechanism. Chitosanases, \( \beta \)-mannosidases, cellulases, glucan-1,3-\( \beta \)-glucosidases, lichenases, endo-1,4-\( \beta \)-xylanases, cellulose-1,4-\( \beta \)-cellobiosidases, endo-1,6-\( \beta \)-galactanases. 179 sequences.

**GH10 results.** Retaining mechanism. Endo-1,3-\( \beta \)-xylanases. 97 sequences.

**GH26.** GH26 enzymes are over 90% bacterial, the rest being fungal. They consist mainly of \( \beta \)-1,4-mannanases with some \( \beta \)-1,3-xylanases. It has been hypothesized that substrate glycons of GH26 enzymes have either \( B_2 \) or \( ^4H_3 \) transition-state conformations, depending on whether the glycon is a \( \beta \)-mannosyl or \( \beta \)-xylosyl residue, respectively, to help facilitate the GH double-displacement retaining mechanism.\(^{15,16} \) The subfamilies identified for cluster comparison of functional divergence were either too small (four very similar sequences) or too diverse to provide analytical results. For example, a comparison of proteobacterial \( \beta \)-1,3-xylanases and bacterial \( \beta \)-1,4-mannanases, both containing four sequences, reveals that despite high primary sequence homology in the vicinity of the catalytic acid/base (Glu116 in *Vibrio sp. AX-4* \( \beta \)-1,3-xylanase (PDB 2DDX) and Glu212 in *Cellivibrio japonicus* \( \beta \)-1,4-mannanase (PDB 1J9Y) and the catalytic nucleophile (Glu212 in *Vibrio sp.* \( \beta \)-1,3-xylanase and Glu320 in *C. japonicus*) \( \beta \)-1,4-mannanase, large discrepancies exist in secondary structure homology at both the N- and C-terminal ends of the MSA (see Supplementary Material).

**GH39 results.** GH39 \( \alpha \)-L-iduronidases are found in a few fungi and animals. They cleave the nonreducing-end \( \alpha \)-linked iduronic acid portions of glycosaminoglycans, heparin sulfate, and dermatan sulfate. The \( \beta \)-D-xylosidases found in GH39 occur in bacteria, and they catalyze the exo-hydrolysis of 1,4-\( \beta \)-D-linkages from the non-reducing termini of xylans, providing a marked contrast in function from \( \alpha \)-L-iduronidases. The presently available GH39 tertiary structures consist of \( \beta \)-D-xylosidases from *Geobacillus stearothermophilus* and *Thermoaerobacterium saccharolyticum*, along with an unposted structure of a *Homo sapiens* \( \alpha \)-L-iduronidase.

Results from DIVERGE show that only two active-site amino acid residues, Arg52 and Trp316 (*G. stearothermophilus* \( \beta \)-D-xylosidase numbering), with \( p^* \) values > 0.95 are present
in GH39 enzymes: \textbf{(Figure GH39.1)}. Arg52 is necessary for maintaining deprotonation of the catalytic nucleophile, Glu278,\textsuperscript{17,18} and is conserved within the two clusters but not in other GH39 sequences. Trp316 is conserved only within the β-D-xylosidase cluster; the sequence from \textit{H. sapiens} has a Phe residue one position past Trp316, but this may point to the nature of differing substrates between β-D-xylosidases and α-L-iduronidases. β-D-Xylosyl subunits possess a markedly different electron density at C5 than that of α-L-iduronic acid; β-xylose lacks a pendant group, which makes it more likely to have favorable hydrophobic stacking interactions with the aliphatic carbons of the Trp indole ring, whereas the carboxyl group present on the C5 of α-L-iduronic acid make it a candidate for favorable interactions with a basic amino acid.

One other note of interest between the β-D-xylosidase and α-L-iduronidase clusters is that all the divergent residues with $p^* > 0.95$ reside within the (β/α)$_8$ barrel (\textbf{Figure GH39.2}). This would indicate and verify that the N-terminal β-sandwich domain and helix insertion have no functional significance.\textsuperscript{18}

\textbf{GH51 results.} Retaining mechanism. α-L-arabinofuranosidases, endoglucanases. 40 sequences.

\section*{Conclusions}

\section*{Acknowledgments}

The authors are grateful to the U.S. Department of Agriculture for the funding that made this work possible.

\section*{References}


Figure GH39.1. Arg52 and Trp316 are important residues to the functional divergence between β-xylosidases and α-iduronidases in GH39. Crystallized xylose ligands are shown, along with nucleophile Glu278 and catalytic acid/base Glu160. Numbering is according to PDB structure 2BS9 from *G. stearothermophilus*. 
Figure GH39.2. Divergent residues are absent from the helix insertion and β-sandwich regions of *T. saccharolyticum* 1PX8. Divergent residues shown as spheres, and xylosyl ligands, catalytic residues, and Arg52 and Trp316 shown in stick representation.
Figure GH39.3. Maximum likelihood phylogenetic tree for GH39 enzymes. Subfamilies used for the cluster comparison are bracketed and denoted.
Appendix B. Molecular Dynamics Simulations of Dicer Ribonuclease

The following material is from a NASA Graduate Student Researchers Program proposal submitted in January 2007:

Specific Aims

RNA interference (RNAi) is a biological mechanism finding increasing use for treating disorders and viruses such as cancer and hepatitis. RNAi is based on microRNA (miRNA) and small interfering RNA (siRNA) syntheses, small molecules consisting of < 30 nucleotides (nt) that create base pairs with messenger RNA (mRNA) and trigger the silencing, translational repression, or expression of genes. Dicer ribonuclease (Figure 6.1) is the enzyme used to create the final double-stranded (dsRNA) segment before being used by the RNA-induced silencing complex (RISC), the protein assembly responsible for pairing miRNA or siRNA to mRNA. Dicer falls into the ribonuclease class III (RNase III) enzyme family, and although several crystal structures have been elucidated that help to demonstrate the divalent metal ion dependence of catalysis, the Dicer mechanism has been somewhat elusive. Understanding this mechanism and the role that enzyme subunits play would increase the effectiveness of scientists’ abilities to develop pathways for using RNAi, since the Dicer reaction represents the gateway for using RISC. Increased control of the RNAi process would allow design of...
suitable dsRNA Dicer substrates to reduce the mammalian protein kinase R (PKR) response toward longer dsRNA molecules, elucidation of gene expression relationships through identification of expressed or silenced genes, and design of specific siRNAs to be used for gene knockdown of viral processes and oncogenes.

Researchers have yet to develop a viable system to use insights gained from one class of RNase III enzymes and apply them to another. Our long-term goal is to contribute to the development of a common reaction mechanism for RNase III enzymes through computational methods, and to determine the Dicer reaction mechanism in the immediate future. Steps have been made toward understanding Dicer catalysis since the enzyme’s discovery, but three differing mechanistic models have been proposed. These models have been based exclusively upon experimental results. The central hypothesis behind the proposed research is that computational simulation, specifically MD, can be used to elucidate the Dicer reaction mechanism and thus begin development of a system to analyze other RNase III enzymes.

Two specific aims are designed to provide greater understanding into the various steps of the Dicer reaction mechanism:

1. **Provide support for the Dicer model of two-metal-ion phosphodiester hydrolysis.**
   The *Giardia intestinalis* Dicer crystal structure (2FFL) has two Mn$^{2+}$ ions at each catalytic active site, but no dsRNA substrate; an MD simulation will provide a quantitative comparison between a two-metal-ion system versus a one-metal-ion system when bound to dsRNA.

2. **Demonstrate the role of PAZ in Dicer.** 2FFL contains a PAZ domain with an extra loop not seen in Argonaute PAZ domains, creating several different interactions with dsRNA. MD will examine the effect PAZ has on dsRNA binding in Dicer.

Dicer lies at the RNAi pathway entrance, and understanding its reaction mechanism will help to provide a systematic approach to clinical applications of RNAi. Because there are fundamental differences in the type and number of domains in prokaryotic and eukaryotic Dicer structures, advances in using RNAi treatments on humans have been extremely limited. Achieving the specific aims will help create new therapeutic methods for using RNAi.
Background and Significance

RNAi was first acterized in the 1990s as a process that used dsRNA to silence genes.\(^1\) Precursor dsRNA strands, hundreds of nt in length, are processed by the enzyme Drosha\(^2\) into shorter (~70 nt) dsRNA units with a 2-nt overhang at the 3’ end that allows specific recognition by Dicer (Figure 6.2).\(^3\) Dicer then hydrolyzes the double strand to create a ~22-nt siRNA or miRNA, which serves as the Watson–Crick base pairing factor that RISC uses to locate mRNA substrates. RISC is a protein assembly\(^4\) consisting of several dsRNA recognition domains such as PAZ and PIWI, along with the mRNA cleavage enzyme Argonaute.\(^5\) SiRNA substrates have perfect base-pair complementation with mRNA substrates and lead to degradation, while miRNA substrates cause ‘gene knockdown’, a non-hereditary gene silencing due to approximate complementarity with mRNA.\(^6\)

RNAi is a process that could find widespread medical use,\(^7\) and much progress has been made towards understanding the RNAi mechanism. Several applications have been developed: siRNA has been intravenously injected to suppress influenza A virus expression,\(^8\) bound ionically to polymers used in gene therapy to silence influenza strains,\(^9\) and placed in plasmids for treatment of Huntington’s disease.\(^10\) Despite siRNA sequences being conserved across many types of organisms, the proteins carrying out RNAi vary to some degree. \textit{Drosophila melanogaster} has two Dicer-like proteins\(^3\) with only one performing dsRNA hydrolysis, whereas mammalian species possess a single Dicer enzyme. These RNAi framework differences, along with the lack of characterization in higher eukaryotes relative to the
amount of research performed on *D. melanogaster* and *Caenorhabditis elegans*, have led to delays in anticipated therapeutic breakthroughs.

Dicer’s relationship to other proteins in the RNAi pathway underscores the significance in gaining a complete understanding of this enzyme’s reaction mechanism. Other RNase III enzymes have been mechanistically characterized, including *E. coli* RNase III,\(^{11}\) *H. sapiens* Dicer,\(^{12}\) and *G. intestinalis* Dicer.\(^{13}\) Distinct differences in secondary structure occur between all three classes of RNase III enzymes. For example, *E. coli* RNase III acts as a homodimer, with each molecule containing a catalytic domain (RIII) that utilizes a two-metal-ion mechanism to hydrolyze both cleavage points of the dsRNA substrate, creating an 11-nt cleavage product (Figure 6.3).\(^{13}\) *HsDicer*, in contrast, acts as a pseudodimer, with two RIII domains (RIII\(_a\) and RIII\(_b\)) that hydrolyze only one dsRNA cleavage point, due to the fact that RIII\(_b\) has one non-functioning active site (Figures 6.2, 6.3).\(^{12}\) A 22-nt product is the result of this condition, which helped researchers elucidate the role of the 2-nt overhang in Drosha products. Initial research on *Aquifex aeolicus* RNase III led scientists to hypothesize a single-metal-ion hydrolysis mechanism (preferably Mg\(^{2+}\)) leading to 11-nt products,\(^{14}\) which was inconsistent with the 22-nt length of Dicer products. Later research using mutagenic and kinetic studies provided evidence for the presence of two Mg\(^{2+}\) ions per strand in dsRNA substrate,\(^{12,13}\) and the crystal structure 2FFL for giDicer clearly supports this hypothesis, showing the presence of two Mn\(^{2+}\) ions at the site of phosphodiester cleavage.\(^{15}\) The main drawback to the 2FFL structure is the lack of a dsRNA substrate.

![Figure B.3](image-url) **Figure B.3.** Secondary structure comparison of RNase III enzymes. Ec: *E. coli*; Dm: *D. melanogaster*; Hs: *H. sapiens*; DUF283: domain of unknown function 283.15
HsDicer also has different domains (Figures 6.2 and 6.3) from lower eukaryotic RNase III enzymes. PAZ and dsRBD domains are common to other proteins in the RNAi pathway, most notably Argonaute, which helps to explain the role that they play with Dicer, but differences also occur, such as the longer loop length in the PAZ domain of giDicer from that seen in Argonaute. Dicer is also involved in the beginning stages of RISC assembly. Elucidating the hsDicer reaction mechanism will provide a focus point for understanding many essential RNAi mechanisms due to the similarities between Drosha, Dicer, Argonaute, and other RISC proteins.

Current therapeutic applications of RNAi consist of: 1) shotgun methods using recombinant Dicer to create siRNA libraries to locate the complementary match to mRNA genes responsible for specific disorders; 2) injection of long (> 100-nt) dsRNA strands that contain known siRNA sequences for gene knockdown; and 3) gene therapy through conjugation with polymeric carriers for cell absorption. Several disadvantages exist for each of these options, the main ones being activating the inflammatory response of PKR, lack of physiological targeting, and lack of concrete identification of the mRNA genes that control disorders. Understanding the hsDicer reaction mechanism will give researchers the flexibility to customize miRNA and siRNA products and give them greater insight into the necessary requirements for using RISC toward specific mRNA genes.

The benefits of performing this research will be to:
1. Confirm or disprove previous research into the double-metal-ion hydrolysis mechanism of RNase III enzymes, specifically Dicer.
2. More easily deduce the similarities and differences of RNase III enzymes.
3. Guide mutagenesis studies of Dicer to enable efficient analysis of the reaction mechanism.

Rationale

Currently there is no published computational work on Dicer. Experimental research on Dicer has increased dramatically since its discovery by Hannon’s group in 2001. Hannon has continued Dicer research, while Filipowicz’s group performed mutagenesis studies on other RNase III enzymes along with Dicer, marshaling evidence for the double-metal-ion
reaction mechanism that supports the metal ion configuration in 2FFL. Filipowicz has continued studies on Dicer, focusing on hsDicer. Nicholson’s group performed kinetics studies on Dicer, also concluding that it functioned by a two-metal-ion mechanism, and has continued examining E. coli RNase III. We believe that computational work using 2FFL will help to support the current work being done on Dicer, and that it will also help complete the picture of the catalytic mechanism, allowing for synthesis of more efficient pathways to use Dicer’s role in RNAi.

**Computational Research.**

Specific Aim #1: Provide support for Dicer model of two-metal-ion phosphodiester hydrolysis. The two Dicer mechanisms presently in question are those of single-metal-ion catalysis and double-metal-ion catalysis. Mg\(^{2+}\) appears to be the preferred metal ion, but Mn\(^{2+}\) also facilitates activity in Dicer, as well as being the ion crystallized in 2FFL. Crystallization possesses two inherent disadvantages: inadequate enzyme-substrate solvation and a rigid structure that cannot portray small- and large-body motions. NAMD will overcome these shortcomings by completely solvating giDicer and allowing small-body protein flexibility. PCModel (Serena Software, Bloomington, IN) will be used to construct dsRNA substrates of varying lengths (30–50 nt) and sequences and modeled into the active site. Two scenarios for each metal ion (Mg\(^{2+}\) and Mn\(^{2+}\)) will be simulated: Scenario 1 will place one metal ion within the active site, while Scenario 2 will place two metal ions in the active site. All simulations will be run for at least 2 ns on the Iowa State BlueGene-L supercomputer. Our expected outcome is to find a more thermodynamically favorable result with the two-Mg\(^{2+}\) scenario, in accordance with previous experimental results.

Specific Aim #2: Demonstrate the role of PAZ in Dicer. MacRae et al. found an extended surface loop in giDicer PAZ domain compared to H. sapiens Argonaute PAZ, suggesting the loop difference could lead to specific recognition of the 3’ 2-nt extension found in all dsRNA Dicer substrates. MD simulations using larger timescales will be run to compare the differences between giDicer PAZ and hsArgonaute PAZ when interacting with the 3’ end of a dsRNA substrate. Data from the simulations in Specific Aim #1 (SA1) will also be analyzed for PAZ-dsRNA interactions. Our expected outcome is to see different interactions in the giDicer PAZ due to the high frequency of basic amino acid residues in the extended loop,
along with enzyme-substrate interactions 25 nt in length, consistent with the product length of Dicer catalysis.

References


