Computational methods in the study of carbohydrates and carbohydrate-active enzymes

Anthony David Hill
Iowa State University

Follow this and additional works at: https://lib.dr.iastate.edu/rtd
Part of the Bioinformatics Commons, and the Chemical Engineering Commons

Recommended Citation
https://lib.dr.iastate.edu/rtd/1867

This Dissertation is brought to you for free and open access by the Iowa State University Capstones, Theses and Dissertations at Iowa State University Digital Repository. It has been accepted for inclusion in Retrospective Theses and Dissertations by an authorized administrator of Iowa State University Digital Repository. For more information, please contact digirep@iastate.edu.
Computational methods in the study of carbohydrates and carbohydrate-active enzymes

by

Anthony David Hill

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, Co-major Professor
Zhijun Wu, Co-major Professor
Balaji Narasimhan
James C. Hill
Richard B. Honzatko

Iowa State University
Ames, Iowa
2006

Copyright © Anthony David Hill, 2006. All rights reserved.
# TABLE OF CONTENTS

Chapter 1: General Introduction ................................................................................................1
   Report organization..............................................................................................................2
   References............................................................................................................................4

Chapter 2: Use, Engineering, and Modeling of Enzymes in Hydrolysis of Cellulose and
   Hemicellulose to Monosaccharides .....................................................................................6
   Introduction..........................................................................................................................6
   A concert of enzymes...........................................................................................................7
   Mechanisms .........................................................................................................................9
   Enzyme engineering...........................................................................................................11
   Modeling............................................................................................................................13
   Conclusions........................................................................................................................20
   References..........................................................................................................................21

Chapter 3: Comparing Programs for Rigid-Body Multiple Structural Superposition of
   Proteins ..............................................................................................................................26
   Abstract .............................................................................................................................27
   Introduction........................................................................................................................28
   Computational methods .....................................................................................................31
   Results...............................................................................................................................35
   Discussion..........................................................................................................................42
   Acknowledgments..............................................................................................................44
   References..........................................................................................................................44
Chapter 4: A Gibbs Free Energy Correlation for Automated Docking of Carbohydrates

Abstract
Abbreviations
Introduction and theory
Computational techniques
Results
Discussion
Acknowledgments
References

Chapter 5: Computational Analysis of Glycoside Hydrolase Family 1 Selectivities

Abstract
Introduction
Materials and methods
Results and discussion
Conclusion
References

Chapter 6: Puckering Coordinates of Monocyclic Rings by Triangular Decomposition

Introduction
Calculation of the angles
Conversion to Cartesian coordinates
Implications of puckering coordinates
Compatibility with other reduced coordinate sets
Conclusions
Acknowledgment
References
Appendix E: Parameter Files for AutoDock 3.06.................................................................135

Appendix F: Complete Family 1 Docking Energies...............................................................162

Appendix G: Complete Family 1 ANOVA Analysis ..............................................................166
Chapter 1: General Introduction

Today, more than ever, energy is an increasingly valuable commodity. World energy use is expected to increase by 54% in 25 years, from 425 quadrillion kJ in 2001 to 680 quadrillion kJ in 2025.\textsuperscript{1} With the increased energy demand, there is increasing demand for biorenewable energy sources. Among these is ethanol derived from corn, with production expected to reach 4 billion gallons in 2005.\textsuperscript{2,3}

Most of this production uses cornstarch as the feedstock, with the cellulosic materials sold as the byproduct “distillers grain”. Of the available biomass in corn, approximately 40% is starch, with the remainder being cellulosic. Cellulosic material is currently not used due to process economics. For these reasons, it is imperative to understand the potential to enzymatically catalyze the hydrolysis of cellulosic materials to glucose, a precursor to ethanol and other products.

Corn stalks, cobs, and husks are primarily composed of cell wall material. $\beta$-Linked sugars form the cellulose and hemicellulose in cell walls. \textit{In vivo}, a concert of enzymes break down cellulose and hemicellulose to monosaccharides. Endoglucanases and cellobiohydrolases in fifteen families (based upon amino acid sequences)\textsuperscript{4,5} hydrolyze cellulose into oligosaccharides as short as cellobiose. Endoglucanases attack throughout polysaccharide chains, producing oligosaccharides of various lengths. Cellobiohydrolases, which are exoglucanases, hydrolyze polysaccharide chains from their ends into cellobiose only. Cellobiose is then broken into glucose molecules by $\beta$-glucosidases, which are found in two families with very different structures, but with the same function. In many species, $\beta$-glucosidases are moderately nonspecific, hydrolyzing galactobiose, mannobiose, and xylobiose, along with other disaccharides, at reduced rates. Once glucose is produced, the same fermentation that is used to generate ethanol from starch can be used.
In understanding the enzymatic catalysis of \( \beta \)-linked polysaccharides, it is necessary to realize that in performing the same function, different organisms produce enzymes of the same type with different amino acids in the chain, conferring slightly different properties on each. This diversity can be used in an exploration of the properties that can be engineered into naturally occurring enzymes. Many of the enzymes that have been crystallized and characterized are from thermophilic archaea, as a source of thermostable enzymes that can be used industrially. Among the other species, various substrate specificities and activities are observed. It would be desirable to engineer broad specificity and high activity into thermostable enzymes, to create an economically viable enzymatic cellulose hydrolysis process.

Throughout this dissertation, I will present exercises to identify residues responsible for specificity as well as tools to aid the researcher in comparing multiple enzymes with similar structures. The inspiration for most of this work was understanding which amino acids in Family 1 \( \beta \)-glycosidases (GH1) control specificity. Only two of the chapters actually deal directly with GH1, while the rest of the chapters detail new computational techniques that I developed during my investigation of GH1. These tools are generically applicable. PyMSS, a program to superimpose three-dimensional structures, can be used with any such structure, including any group of protein structures. The method to mathematically describe puckering coordinates is completely generic with any non-aromatic monocyclic ring. The overlap between puckering, structural superposition, and Gibbs free energy can be applied to carbohydrate-active enzymes. I summarize each chapter in the following section.

**Report organization**

Chapter 2 is a literature review for using, engineering, and modelling enzymes that produce ethanol precursors from cellulose and hemicellulose. It emphasizes finding and engineering enzymes for industrial use, particularly methods to design efficient biological catalysts. I wish to show that the approaches used throughout this thesis are both novel and applicable.
Chapter 3 is a manuscript already published in *Proteins: Structure, Function, and Bioinformatics* about software to optimize the structural superposition of enzymes. I wrote the program to more easily accomplish the work in Chapter 4. After presenting my work on β-glycosidases at a student seminar, I received many questions about the software that I used to generate the superposition. Based upon this interest, I conducted a literature search to discover if anyone had solved the problem in the same manner that I had. Finding no similar software, I wrote a manuscript detailing the program.

Chapter 4 is a manuscript to be submitted to *The Journal of Computational Chemistry*. As I worked on the material for the manuscript in Chapter 4, I found that the carbohydrate free energy function published by Laederach and Reilly in 2003[^6] overestimated the binding strength of the enzymes I was studying. Consequently, I set about retraining the Gibbs free energy function in a fashion similar to theirs, only with a larger data set and a more rigorous conformational search. I was able to find nearly ten times as many crystal structures with experimental binding energies as the previous study. Due to the larger sample size, I was able to more realistically model the Gibbs free energy change of binding.

Chapter 5 is a manuscript concerning the use of automated docking to understand the various specificities of Family 1 β-glycosidases. In the manuscript, I detail the experimental procedure and results of computationally docking 28 disaccharides into each of seventeen unique β-glycosidases from fifteen species, producing the binding energies for each disaccharide. Additionally, the Gibbs free energies of binding are predicted. We intend to submit this manuscript to *Biopolymers*.

Chapter 6 was inspired by a collaboration with Dr. Shinya Fushinobu (University of Tokyo) during the summer of 2006, when he wanted to apply a technique developed by Dr. Dr. Chandrika Mulukala to calculate intermolecular forces on docked structures and to predict the transition-state pathway[^7,^8]. Some of those calculations are less than quantitative. In an attempt to help Dr. Fushinobu quantify the forces’ effect on the puckering of the central
ring, I pictured that ring as being composed of flaps that could pucker up and down. From that picture, it was straightforward to find the axes of puckering and the moments about those axes. This picture then turned into Chapter 6, which has already been submitted for review to *The Journal of Chemical Information and Modeling*.

Since most of the chapters include supporting information, I have included them as appendices at the end of this thesis. Additionally, during my time at Iowa State University, I have developed a number of small computer programs to help me in my work. For the sake of these programs living on beyond my stay here, I have included the source code to these programs as additional appendices.

Finally, I was fortunate enough to have the chance to write a grant proposal based on work done in ChE 688. Appendix A contains that grant proposal to the Center for Catalysis (CCAT) at Iowa State University. Since the grant was funded for 2005–2006, I was able to perform some of the research detailed in the proposal. The results of that work are also included in Appendix B.

**References**


Chapter 2: Use, Engineering, and Modeling of Enzymes in Hydrolysis of Cellulose and Hemicellulose to Monosaccharides

Introduction

With the world’s continuing hunger for energy and the deleterious effects being wrought on our environment as a result, we increasingly look toward biorenewable energy sources. Although it is possible to liberate the energy stored in biomass by simply burning it, the environmental impact is not benign, nor is raw biomass a convenient fuel. To remedy these problems, the most commonly used biorenewable energy source is ethanol from starch hydrolysis followed by glucose fermentation. Ethanol is portable, energy-dense, particulate-free and, most important to its early adoption, combustible in unmodified gasoline engines.

Most of the stored energy in biomass is in reduced carbon, typically as carbohydrates. These carbohydrates are complexed mainly in starch or cellulose. Of the two, starch is more readily available, with the $\alpha$-glycosidic bond joining glucose residues being easier to break and $\alpha$-linked polysaccharides being highly soluble. Cellulose, on the other hand, is much more difficult to break into a more easily used energy source. Biologically, cellulose is used primarily for structure, as opposed to starch’s use as energy storage. Cellulose makes a good structural element because its $\beta$-linked glycosidic bonds require more energy to hydrolyze than $\alpha$-glycosidic bonds, but they also allow cellulose chains to pack in highly ordered structures, maximizing inter-cellulose hydrogen bonding and making crystalline cellulose insoluble in water. Therefore it is desirable to access the sugar monomers that compose cellulose, since approximately half of all carbohydrate biomass exists as cellulosic material.

The biological degradation of cellulose is attractive because many organisms metabolize cellulose, yet the chemical process to transform cellulose into a more usable energy source is prohibitively expensive. For this reason, we desire to understand the working of biological
catalysts – in particular, enzymes. The goal of this research is to help to guide research to hydrolyze cellulose to simple sugars that can be fermented to ethanol in an economically viable process. To do so will likely require modification of enzymes for increased activity, stability under process conditions, and tailored specificity.

In deciding where to proceed with enzymatic cellulose hydrolysis, it is necessary to understand the biological process to metabolize cellulose to simple sugars. Fortunately, this area is becoming increasingly well researched. It has also been extensively reviewed.1–4

**A concert of enzymes**

Crystalline cellulose is difficult to break down, with many hydrogen bonds contributing to its stability. It is for this reason that one enzyme alone cannot hydrolyze cellulose completely. Biologically, a concert of enzymes is used.

The first step in degrading cellulose is to weaken the intermolecular interactions that make crystalline cellulose insoluble. Industrially, an acid solution may be used to transform crystalline into amorphous cellulose. At this point, shorter, water-soluble oligosaccharides are needed. Endoglucanases can attack a cellulosic chain to liberate oligosaccharides as short as cellobiose.2 Currently, endoglucanases have been classified into fifteen glycoside hydrolase families (GH5–10, 12, 26, 44, 45, 46, 48, 51, 61, and 74) based on sequence similarity.5,6

The oligosaccharides can be further broken into cellobiose by exoglucanases or cellobiohydrolases.2 Cellobiohydrolases occur in six families (Families 5–7, 9, 10, and 48) based on sequence similarity. As is obvious, all the families that contain cellobiohydrolases also contain endoglucanases. This observation means that cellobiohydrolases and endoglucanases share much of their primary sequences. The difference between endo- and exo-glucanases is purely structural. Both types of enzymes contain loops at the top of the active site (Figure 1). In endoglucanases, these loops are short enough that they do not close over the active-site cleft, while cellobiohydrolases contain a longer loop that closes over it, creating a tunnel.
Although the loop length does not directly cause this, exoglucanases push oligosaccharides along the length of their tunnels, while endoglucanases are non-processive in their action. The closed loops do, however, hinder endoglucanase activity in cellobiohydrolases.

Through the concerted work of endo- and exo-glucanases, cellobiose is produced from cellulose. $\beta$-Glucosidases in GH1 and GH3 can hydrolyze cellobiose into two glucose molecules. Enzymes in both families contain the catalytic residues in a well, with both glucose molecules being expelled after each hydrolytic reaction, and those in both families operate by the retaining mechanism.$^{5,6}$

In vivo, crystalline cellulose is complexed with hemicellulose and lignin to create cell walls. Although cellulose is a purely $\beta$-linked glucosyl polysaccharide, hemicellulose contains xylosidic, galactosidic, glucosidic, and mannosidic $\beta$-linked monomers. In addition, it is highly functionalized with acetyl, methylglucuronyl, and arabinofuranosyl groups.$^7$ Lignin is a highly branched polymer that results from free radical condensation of aromatic alcohols. For cellulose to be hydrolyzed, hemicellulose and lignin must be hydrolyzed first. Enzymatic hydrolysis of lignin is difficult and is only carried out by radical oxidation by peroxidases.$^{8,9}$
There are many enzymes similar to cellulases active on hemicelluloses, with many categorized in the same families as endo- and exo-glucanases and β-glucosidases. In fact, many β-glucosidases demonstrate broad specificity. Depending on the process used to make ethanol, it will be important to control the specificity of cellulolytic enzymes toward products of hemicellulose and lignin hydrolysis.

**Mechanisms**

Most enzymes that hydrolyze β-glycosidic bonds use a pair of carboxylic acids to stabilize the transition state(s) required for hydrolysis, and thus catalyze reaction. These catalytic residues are totally conserved in all species and enzymes that hydrolyze β-glycosidic bonds, one a proton donor and the other a nucleophile. Depending upon the distance between the catalytic residues, one of two mechanisms occur.

If the catalytic residues are approximately 9.5 Å apart, the inverting mechanism operates (Figure 2). In the first step of the mechanism, the nucleophile donates an electron to a hydrogen atom of a catalytic water molecule. The water hydrogen atom then donates an electron to the glycosidic carbon atom on the residue on the nonreducing side of the glycosidic bond to be broken. The glycosidic oxygen atom is then free to donate an electron to the proton donor. Partial positive and negative charges will temporarily exist on the glycosidic carbon atom.

![Figure 2](image)

**Figure 2.** A glycosidic bond is shown being hydrolyzed by the inverting mechanism. Carboxylic acids protrude into the active site. Redrawn from Ly and Withers.
and proton donor, respectively. This transition state proceeds to the hydrolyzed product. The initial state can then be regenerated by an additional water molecule to reshuffle the proton.

If the catalytic residues are closer, around 5.5 Å, hydrolysis is by the double-displacement retaining mechanism (Figure 3). In this mechanism, the nucleophile donates a proton directly to the glycosidic carbon atom, again allowing the glycosidic oxygen atom to donate an electron to the proton donor. This transition state proceeds to a stable intermediate in which the residue to the nonreducing side of the glycosidic bond to be broken is bound to the enzyme. Upon exit of the reducing sugar, water may enter the active site to complete the hydrolysis, as the proton donor and nucleophilic residues reverse roles to shuttle the electron back through the water and nonreducing sugar. The resulting glycon retains the initial stereo configuration of the glycosidic hydroxyl group.
Enzyme engineering

Zhang et al. have mutated six conserved, noncatalytic residues in the endoglucanase of *Thermobifida fusca* Cel6A.\textsuperscript{15} Thirteen single-point and two double-point mutations were constructed, with the majority of the mutations resulting in a smaller side chain. Eight of the single-point mutations changed the side-chain character, i.e. hydrophobic, acidic, basic, and neutral. Eight of the fifteen mutations showed increased activity on carboxymethylcellulose (CMC) and five showed increased activity toward hydroxyethylcellulose (HEC20). All mutations showed decreased activity toward acid-swollen cellulose (SC) and filter paper (FP).

The mutation sites were selected based upon positions of the residues with respect to crystallized and modeled ligands. The two mutations producing the largest positive change in activity are G263D and K259H. These residues are far enough away from the glycosidic bond that their influence is hypothesized to be on ligand binding.

A prior study by Wolfgang and Wilson\textsuperscript{16} mutated all the Asp residues in the active site or *T. fusca* Cel6A endoglucanase to each of Ala, Glu, and Asn, to determine which Asp residues acted as the catalytic acid and base residues. As a result, they identified Asp117 as the catalytic acid, while the catalytic base was less clear. In the process, they found that mutating Asp156 to Glu156 increased activity of *T. fusca* Cel6A toward CMC and SC.

Zhang et al. also investigated fifteen mutations in loop residues and eight mutations in conserved noncatalytic active-site residues in *T. fusca* Cel6B, an exoglucanase.\textsuperscript{17} As in his Cel6A study, mutated residues were selected based upon proximity to both crystallized and modeled ligands. The resulting enzymes were then assayed for activity on a variety of substrates, as well as for thermostability and processivity. In particular, they hoped that engineering of some loop residues would increase thermostability, most notably by introducing a disulfide linkage. The mutations were chosen for three reasons: to correspond to a successful mutation in Cel6A, to test and possibly exploit the hypothesized role of a particular residue,
or to create a disulfide linkage in the loops that close over the active site. Most of the mutations decreased activity and thermostability, though there were a few exceptions. Notably, mutating the active-site His326 to either Ser326 or Ala326 increased activity toward CMC and cellopentaose, while not significantly impacting activity toward SC or cellotetraose. Also, mutating loop residue Leu230 to Cys230 significantly increased activity toward CMC and SC, while decreasing activity on microcrystalline bacterial cellulose (BMCC). Mutating another loop residue from Gly284 to Pro284 increased rates toward CMC, SC, FP, BMCC, cellopentaose, and cellotetraose. The double mutation G234S/G284P increased the processivity as well as hydrolysis of SC, FP, and BMCC, while drastically cutting rates toward cellopentaose and cellotetraose. The disulfide linkages all decreased oligosaccharide hydrolysis rates, and only the N233C/D506C mutation increased the rate toward any polysaccharide.

Protein engineering has also been conducted on *Hypocrea jecorina* (formerly *Trichoderma reesei*) Cel7A cellobiohydrolase. The amino acids and the mutations were chosen by comparing the sequence of *H. jecorina* Cel7A to those of *H. jecorina* Cel7B and *Humicola insolens* Cel7B, both endoglucanases. Five amino acids were mutated to the residue present in *H. insolens* Cel7B, resulting in E233S, A224H, K225V, T226A, and D262G mutations. Only the quintuple mutant was explored. The resulting mutant had decreased activity toward BMCC and a higher optimal pH.

*Pyrococcus furiosus* β-glucosidase has been mutated to increase its activity toward galactobiose by incorporating residues present in *Lactococcus lactis* 6-phospho-β-galactosidase. In addition, *Sulfolobus solfataricus* β-glycosidase was mutated after identifying the residues proximal to docked 4-OH substrate analogs. These mutations resulted in an increase in its β-xylosidase activity and a decrease in its β-fucosidase activity, while a second mutation decreased its activity toward cellobiose and galactobiose. Finally, by using DNA shuffling between *P. furiosus* β-glucosidase and *S. solfataricus* β-glycosidase, Kaper et al. engineered three enzymes with 1.5- to 8.6-fold increased lactose hydrolysis rates.
Modeling

With the importance of carbohydrates in both biology and industry, we need to characterize carbohydrate-active enzymes and to understand how they work. There have been great strides made in modeling carbohydrates and the enzymes that operate on them.

Of course, much of the success in modeling carbohydrates comes from the exponential increase in computing power over time that benefits all computational modelers. Calculations that were once only feasible on supercomputers are now routinely carried out on the desktop. At the same time, many approximations that were used because the full numerical solution was too computationally intensive have been replaced by the full calculations. However, for all the increases in computing resources, we still cannot carry out \textit{ab initio} simulations of biologically relevant systems. To this end, we must choose how best to approximate salient details and still arrive at accurate answers.

Carbohydrates are unique molecules in the simulation world\textsuperscript{22,23}. They are chiral cyclic aldoses or ketoses, with a very high degree of substitution by hydroxyl groups. Carbon’s tetrahedral geometry does not allow all ring atoms of a carbohydrate’s non-aromatic five- or six-member rings to be in the same plane. Because of exocyclic groups, there are significant differences in carbohydrate ring puckering\textsuperscript{22,24,25}. As sugars are primarily used as either energy storage or biological identification, there is much more interest in studying reactions and reaction mechanisms than bulk properties. Following are a few of the unique characteristics of carbohydrates and their implications on the computational modeling of sugars.

\textit{Hydrogen bonding}

One molecule of glucose contains six hydrogen-bond acceptors and six hydrogen-bond donors. Given all the hydroxyl groups of a carbohydrate, with an average contribution of 3–10 kcal/mol per hydrogen bond\textsuperscript{26,27}, hydrogen bonding becomes an important consideration.
in carbohydrate modeling. Although a 36–120 kcal/mol contribution due solely to hydrogen bonds seems high, it helps to explain the extreme solubility of carbohydrates (67.36% w/w for sucrose in water\(^2\)). Such energetics would obviously affect carbohydrate structures in solution or when bound to enzymes.\(^{29,30}\)

Due to the shorter lengths of hydrogen bonds compared to van der Waal’s interactions, they are typically simulated with an equation similar to a Lennard-Jones 12–6 term, a 12–10 potential (eq. 1):

\[
E_{hb} = \sum \left( \frac{C_{ij}}{R_{ij}^{12}} - \frac{D_{ij}}{R_{ij}^{10}} \right)
\]

where \(E_{hb}\) is the energy contribution of the hydrogen bonds, \(i\) and \(j\) are two distinct atoms, \(R_{ij}\) is the distance between those atoms, and \(C_{ij}\) and \(D_{ij}\) are constants for the hydrogen bonding of the atom-type pair of atoms \(i\) and \(j\).

Due to the angular dependence of hydrogen bonds, eq. 1 is often modified:

\[
E_{hb} = \sum \left( \frac{C_{ij}}{R_{ij}^{12}} - \frac{D_{ij}}{R_{ij}^{10}} \right) \cos^m(\theta_{A-H...B})
\]

where \(m = (0, 2, 4)\) and \(\theta_{A-H...B}\) is the angle between the hydrogen bond donor and acceptor. It is important to note that this formulation does not take into account the orientation of the hydrogen bond with respect to the lone pair of electrons that accepts it.

Because of the importance of hydrogen bonds to carbohydrates, it bears noting that many widely used biological force fields, such as the original CHARM\(m\)\(^{31,32}\) and AMBER,\(^{33}\) neglect hydrogen bonding; later versions of the CHARM\(m\) force field include an explicit 12–10 hydrogen bonding term.\(^{34}\) The original version of AMBER had an explicit 12–10 term for hydrogen bonding, but it was removed following a paper discussing its lack of importance in simulating amide crystals.\(^{34,35}\) Although CHARM\(m\) and AMBER are molecular mechanics force fields that are generically applicable to biological molecules, both have been modified to make them more carbohydrate-specific.\(^{23,36,37}\) However, none of these force fields includes
a completely accurate hydrogen bonding term that deals with the position of the lone electron pair in the hydrogen bond acceptor. Rather, they have focused their efforts on modeling the exo-anomeric effect and proper torsions of the exocyclic groups.\textsuperscript{23}

Although existence of the hydrogen bond has been known since the 1920’s\textsuperscript{38,39} and even well understood since the late 1970’s,\textsuperscript{40} whether or not explicit modeling of hydrogen bonds produces a more accurate simulation has been debated. Typically, hydrogen bonds are modeled with electrostatic effects and a slightly exaggerated polarization.\textsuperscript{35,41–48} This method has worked well, enabling researchers to study the effects of hydrogen bonds on carbohydrate structure and solvation.

Recently Allinger et al. updated the MM3\textsuperscript{49} force-field to MM4,\textsuperscript{50} adding several new cross-terms, better polarization, and better hydrogen bonding, including consideration of the angle between the hydrogen bond and the lone pair of electrons accepting it.\textsuperscript{34} The addition of these terms improved the modeling of carbohydrates over MM3.\textsuperscript{51}

Solvation

Since we know that carbohydrates strongly interact with water in solution, we also know that solvation and desolvation strongly affect carbohydrate structure. A great deal of this interaction is via hydrogen bonding; however, that is not the only factor that should be considered in carbohydrate solvation. Since hydrogen bonding is an internal energy term, we must account for the volume change of solvation to obtain an enthalpic term. Finally, if we are interested in solvation thermodynamics, we must take into account entropy changes due to solvation.

Gibbs free energy ($\Delta G$) is the sum of three terms: internal energy ($\Delta U$), a pressure-volume term, and a temperature-entropy term. Since solvation is thermodynamically driven,\textsuperscript{52–56} it is necessary to look examine all three terms in the $\Delta G$ equation (eq. 3):

$$\Delta G_{\text{solv}} = \Delta U_{\text{solv}} + \Delta(PV)_{\text{solv}} - \Delta(TS)_{\text{solv}}$$

(3)
Internal energy is the sum of bonded and nonbonded terms. Of the bonded-atom terms, puckering is the single largest determinant of changes in energy. A more complete discussion of puckering effects follows in a later section. Hydrogen bonding is one of the largest (for carbohydrates) and most poorly modeled component of energy between nonbonded atoms. A more complete discussion of hydrogen bonding proceeded this section. In this section we will discuss the pressure-volume and temperature-entropy terms.

The pressure-volume term can be explained as follows. First, if we limit ourselves to constant pressure cases, then we only need to consider volume changes. Take, for example, the binding of a carbohydrate ligand to an enzyme. If the ligand is hydrophobic, then it will displace more water than its atomic volume. If the protein is also hydrophobic, then its binding pocket will not be entirely full of water. Upon binding, the volume of water displaced from the binding pocket will not entirely fill the void in the solution, and the system volume will decrease. For hydrophilic/hydrophilic systems, the volume change upon binding would be ~0. Hydrophobic/hydrophilic systems should have a volume change that is slightly negative, but not as negative as hydrophobic/hydrophobic systems.

If we in addition consider only isothermal systems, then the only term left to consider is the entropy change associated with carbohydrate solvation. We need to discuss the entropic changes in both the carbohydrate and the solution, since there are changes in both, and they are of similar magnitude. In considering the solution, we must think about the water molecules associated with solvating the carbohydrate. A solvated carbohydrate will have a tightly associated shell of water molecules. Carbon and nonpolar hydrogen atoms will have a solvation shell about one molecule thick. Due to hydrogen bonding, the hydroxyl groups of the carbohydrate will have a thicker solvation shell. The entropy of the water molecules in the solvation shell is relatively low, since their motion is greatly limited. The entropy change of the solution will be negative, in proportion to the number of water molecules needed to form the solvation shell of the carbohydrate.
Finally, we must consider the effect on the entropy of the sugar molecule by solvation. Again, this effect is not specific to carbohydrates, and so the discussion can be more general. When not solvated, the sugar molecule will have a very low entropy, since many of its degrees of freedom are restricted. When solvated, the molecule will be less restricted. According to the Boltzmann-Gibbs definition of entropy in statistical mechanics, the entropy of a system is related to the energy and probability of its microstates (eq. 4):

\[ S = -k \sum_i p_i \ln p_i \]  

(4)

If we assume that a solvated sugar has high and low energy degrees of freedom, and that it will move primarily in its low-energy ones, then we can reduce eq. 4 to eq. 5:

\[ S = k \ln N \]  

(5)

where \( N \) is the number of low-energy degrees of freedom. Since the lowest-energy degrees of freedom are the dihedral rotations of exocyclic groups, translation, and rotation, we can assume that the change in entropy of the carbohydrate upon solvation is proportional to the logarithm of the number of these degrees of freedom.

**Puckering**

Since carbohydrates are either five- or six-member rings, not all of the atoms can lie in the same plane without significant angle strain. To resolve this angle strain, ring atoms “pucker” out of the plane, minimizing angle strain as well as steric clashes. Six-member carbohydrate rings may be described as moving through 38 canonical states (Figure 4), previously mapped in spherical coordinates. This spherical projection connects these states as continuous movements in each of the three spherical coordinates.\(^{62-64}\)
Often the binding of a carbohydrate to a protein requires both the protein and carbohydrate to change shape. Of course, these changes in shape are caused by favorable intermolecular interactions. The discussion of those energies precedes this section. Balancing those external interactions are the internal bonded energies within the sugar molecule itself. These changes in the ring shape are also necessary for chemical reactions in which the glycosidic bond is formed or broken, primarily due to one of the ring bonds adopting a partial aromatic character during the transition state. The internal energetics of moving a carbohydrate though its puckering states has been studied previously by molecular mechanics. Indeed, advanced force fields allow a carbohydrate ring to pucker and are able to describe the energy of puckering with reasonable accuracy.

Figure 4. The five canonical shapes of a puckered pyranosyl ring: chairs (C, two states), envelopes (E, twelve states), boats (B, six states), skew–boats (S, six states), and half-chairs (H, twelve states). Four atoms are coplanar in four of the conformations, and a fifth is coplanar in envelopes. Superscripts and subscripts denote the atoms not in the plane.
However, advanced force fields are often too computationally intensive for automated docking. To speed calculations, many popular automated docking programs limit the degrees of freedom to the lowest-energy ones.\textsuperscript{68–70} Although ring puckering is a relatively low-energy transformation, it is neglected in several popular automated docking packages\textsuperscript{68,70} because moving a ring through its puckered states requires adjusting bond angles to maintain bond lengths, and this is a very high-energy transformation.

\textit{Forces}

Because many carbohydrate-active enzymes are processive, it is beneficial to consider the intermolecular forces on sugar molecules. With molecular dynamics, these forces are known, and in fact they are a necessary part of the simulation. However, many enzyme–carbohydrate systems are quite large, so molecular dynamics requires considerable computation. Therefore, automated docking is still the simulation of choice if many different enzymes, ligands, or protonation states must be examined. Since forces are not computed throughout the docking simulation, they must be calculated at its end. Of primary importance is the intermolecular force which can be computed from the components of the simulated system’s energy.

Instantaneous forces can calculated by numerical differentiation of the energy landscape in the three Cartesian coordinates.\textsuperscript{71,72} The force exerted on a ligand atom is \(-\nabla E_{\text{Inter}} = F_{\text{Inter}}\), where \(F_{\text{Inter}}\) is the force in the three spatial dimensions that the enzyme exerts on the ligand.

\(F_{\text{Inter}}\) can be nonzero even when the system is at rest. \(E_{\text{Total}} = E_{\text{Inter}} + E_{\text{Intra}}\), 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.

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 all of its dihedral rotational deg-
rees of freedom. Lacking in AutoDock 3.06 are the contributions of ring puckering and enzyme internal movement, along with higher-energy degrees of freedom like bond stretching and angle bending.\textsuperscript{68,69} However, successful catalysis requires that ligand ring puckering change even while the ligand is somewhat stationary in the examined degrees of freedom. This fits with our current understanding of how enzymes work – that a ligand arrives in the active site and remains stationary and in a low-energy state until suddenly it passes through the high-energy transition state, leaving no reactant in its active site, but rather product(s).\textsuperscript{73}

Once bond cleavage occurs, $F_{\text{Intra}}$ changes and the entire system will again move toward a low-energy state. One can thus compute the expelling and processive forces on the products from the $F_{\text{Inter}}$ values of the 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 when the reactant is stationary in the active site gives an idea of which direction the system will move upon bond cleavage.

**Conclusions**

In progressing toward an economically viable cellulose hydrolysis process, it is necessary to engineer not just one enzyme, but three to five enzymes. Obviously, increased activity is desired, though with the knowledge that hemicellulose and lignin will be present in actual feedstocks, specificity control is required. Further, it is necessary to watch for enzyme inhibition by reaction products, particularly as this enzyme system is readied for deployment to industry. The research already accomplished in engineering cellulases shows that there is much promise for increased reaction rates with further research. We also observe that is very difficult to predict the outcome of a mutation.

The methods used by Wilson et al.\textsuperscript{15–17} show that it is possible to screen many mutations \textit{in vitro}. This advance is necessary as we gain more understanding into the relationship be-
tween sequence and function in cellulases. The large number of potential mutations that protein designers will suggest can be effectively tested using this procedure.

Thus far, mutations are made with little idea of outcomes, mostly due to lack of knowledge about the role each amino acid plays in catalysis of cellulolic degradation. High-throughput methods to analyze the role of amino acids are needed to more intelligently design highly active enzymes. Computation is one such method. Zhang et al. have shown that modeling of uncrystallized ligands can be valuable in designing improved enzymes. Computational modeling of site-directed mutagenesis and the resulting substrate binding could prove to be another valuable tool. Gene shuffling has also been valuable in increasing activity. As we sequence and crystallize more cellulolytic enzymes, the effectiveness of these three tools will only increase.

With improved computational techniques and knowledge of fundamental mechanics, we could begin to move protein engineering in silico. Much is known about correctly simulating essential elements of carbohydrate systems, but not all this information has been brought together to accurately direct protein engineering efforts. Research in this area is greatly needed.

References


44. Homans, S. W. Biochemistry 1990, 29, 9110.


71. Mulakala, C.; Reilly, P. J. *Proteins* 2005, 60, 598.
Chapter 3: Comparing Programs for Rigid-Body Multiple Structural Superposition of Proteins


ANTHONY D. HILL AND PETER J. REILLY*

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

Key words: genetic algorithm; MASS; MODELLER; MultiProt; neighbor-joining; protein structural superposition; pseudo-star; PyMSS; STAMP

Short title: SUPERIMPOSING PROTEIN STRUCTURES

Grant sponsor: U.S. National Science Foundation.

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

Different programs and methods were employed to superimpose protein structures, using members of four very different protein families as test subjects, and the results of these efforts were compared. Algorithms based on human identification of key amino acid residues on which to base the superpositions were nearly always more successful than programs that used automated techniques to identify key residues. Among those programs automatically identifying key residues, MASS could not superimpose all members of some families, but was very efficient with other families. MODELLER, MultiProt, and STAMP had varying levels of success. A genetic algorithm program written for this project did not improve superpositions when results from neighbor-joining and pseudo-star algorithms were used as its starting cases, but it always improved superpositions obtained by MODELLER and STAMP. A program entitled PyMSS is presented that includes three superposition algorithms featuring human interaction.
Introduction

Proteins produced by different species but with the same general amino acid sequence adopt roughly the same fold. Their degree of conservation is impossible to quantify without recognition of corresponding amino acid residues between proteins from different sources. Multiple sequence alignment allows comparison of amino acid sequences, while multiple structural alignment allows comparison of folds, where the similarity between two proteins can be compared for each three-dimensional coordinate.

A major use of multiple structural alignment is to align varying results from NMR determination of protein tertiary structures. A second use is for automated docking, where instead of docking many ligands into the active or binding sites of a single protein,\(^1\)\(^2\) a few ligands are docked into proteins of the same family, but from different species. As automated docking is sensitive to the ligand’s initial position, it is necessary that initial and presumptive final positions be near each other. It is therefore important to closely superimpose the docking sites of the many related structures being investigated.

The multiple structural alignment problem can be divided into two subproblems. The first is to identify multiply corresponding structural elements. Traditionally, multiple structural alignment algorithms have focused computational power on this subproblem, that of the NP-hard largest common subsequence.\(^3\) This approach is necessary when superposition of many families is required, and it is used by programs such as COMPOSER,\(^4\) MASS,\(^5\) MODELLER,\(^6\)\(^7\) MultiProt,\(^8\) SCOP,\(^9\) SSAPm,\(^10\) and STAMP.\(^11\) Alternatively, researchers can manually identify which residues are important for structural formation, catalysis, or ligand binding, and then base the superposition on them.

The second subproblem is to calculate the appropriate rigid-body transformation for each structure to create an optimal superposition. With N molecules to be superimposed, the 6N-
dimensional conformational space is huge and highly nonlinear, so heuristics are employed to find the optimal conformation.

An analytical solution using an orthogonal transformation exists to optimally superimpose two structures.\textsuperscript{12} Although many algorithms have been proposed to find an optimal multiple superposition, each of these potentially produces suboptimal results, mostly because they rely solely on pairwise superpositions. In the simplest algorithm, used by both MASS and MultiProt, each structure is superimposed on a chosen reference structure, referred to hereafter as the pseudo-star algorithm (PSA) (Figure 1a). An improvement upon this idea is a neighbor-joining algorithm (NJA) (Figure 1b), used by SSAPm and STAMP, where the structures closest to each other are superimposed upon one another, and then a consensus structure is iteratively used for further superposition with other consensus structures until one superposition is formed. Neighbor-joining can help to reduce bias by superimposing the closest structures first; however, biases will remain, as the initial superpositions will continue to affect the final superposition.

A structural superposition based upon a truly average structure will create a superposition unbiased toward an initial structure (Figure 1c). Examples of this sort of superposition are those used in SCOP, COMPOSER, and MODELLER. Commonly, the average structure is generated iteratively.\textsuperscript{4,6,7} In each iteration, the average structure is computed and all other structures are superimposed in a pairwise fashion to the average structure until the latter changes less than some tolerance. Average structure-based superpositions can still produce suboptimal results due to mirroring. This is an artifact where the distance between the average structure and each of the two structures is optimal, but the distance between both structures can yet be decreased.

Ideally, one would rather minimize the sum of all pairwise root mean squared deviations (sRMSD) (Figure 1d) to obtain an unbiased superposition, free of mirroring or other suboptimalities. A relatively efficient algorithm proposed by Ten Berge\textsuperscript{13} can be used to
arrive at a good approximation of the optimal superposition. In this algorithm, the sRMSD is minimized by iteratively fixing \((N - 1)\) rotational transformations while analytically solving for the remaining transformation that minimizes the sRMSD of \(N\) structures. This procedure is repeated over the \(N\) transformations until the incremental change in sRMSD drops below a specified tolerance. Since an optimal solution cannot be guaranteed, one should calculate its upper and lower bounds to determine if the calculated solution can be globally optimal.14

Unfortunately, the algorithm proposed for this is not available as a program for general use, nor does it always produce a result within the bounds on optimality.14

To compare sRMDSs of multiple structural alignments obtained by PSAs and NJAs, and then of a genetic algorithm (GA) using either of them as a starting value, the program PyMSS (Python\textsuperscript{15} Multiple Sequence Superposition) was written. Corresponding atoms of members of four different protein families were chosen manually, as described below. Mul-

**Figure 1.** Simplified two-dimensional description of group distance metrics with estimated distances listed. Each dot represents a structure, and the optimal superposition collapses all dots to one spot. a) Pseudo-star sum of distances; b) neighbor-joining tree; c) star (or average-structure star) sum of distances; d) sum of all pairwise distances.
Multiple structural alignments of members of the same families were also obtained using MASS, MODELLER, MultiProt, and STAMP, each automatically identifying corresponding atoms. Although MASS and MultiProt use the PSA while SSAPm and STAMP use the NJA to produce their superpositions, they may select different corresponding atoms from each other or from those chosen manually, so their resulting superpositions may be different. To further test the GA in PyMSS, it used the final alignments of either MODELLER or STAMP to initially position structures. At the time of writing, SCOP, SSAPm, and COMPOSER were not available for testing. However, COMPOSER produces results similar to those produced by MODELLER.7

Computational methods

Python is an object-oriented scripting language with interfaces to compiled libraries. This combination allows rapid development while maintaining good application performance. The Genetic-Python library,16 updated in the Debian GNU/Linux distribution,17 is used as the core GA. In addition, the Numerical Python library18 is used for the core matrix operations.

Before running PyMSS with the PSA, NJA, or GA, a multiple sequence alignment and preliminary study of the tertiary structure of the protein family should be performed if necessary to identify either a few key catalytic and/or binding residues or conserved residues well distant from each other, depending upon the user’s requirements.

The selection of key atoms is non-trivial. Without knowledge of the system being superimposed, it can be quite time-consuming to find key residues. To validate PyMSS, we tested it on four protein families, only one of which we knew well. The time required for subjecting the remaining three families to literature searches, multiple sequence alignments, and BLASTing their sequences against each other was far greater than program runtimes. However, if the user knows which atoms are important to protein function or binding, then selection of a few key atoms is quick and yields very good results.
PyMSS can apply three different algorithms to superimpose proteins. Of the three, the GA offers the most theoretical promise, since it can simultaneously optimize all transformation coordinates of all the proteins. In contrast, the PSA and NJA use heuristics and pairwise transformations\textsuperscript{12} to solve the problem three transformation coordinates at a time. Simultaneous optimization results in a highly dimensional problem space that is not necessarily easily minimized. The GA can provide an efficient means of searching that space, yet certain requirements should be met if that efficient search is to be done in a timely fashion: 1) Since the dimensionality of the problem space grows proportionally to the number of structures, the GA should not be used for large numbers of structures (in practice, this means > 50); 2) The GA should start with a solution somewhat close to the optimal solution, which can be achieved by starting with either the PSA or NJA and then following with the GA.

In the PSA, one structure of the group is selected as a reference structure. All the remaining structures are optimally superimposed upon that structure. Since it cannot be guaranteed that any chosen structure will produce the best result, this process is repeated, iteratively selecting each structure to be the reference structure. The superposition with the best sRMSD is used.

The NJA is an iteratively solved process, where each iteration pairs and superimposes the most similar proteins (those having the lowest residual RMSD). From each pairwise superposition, a pairwise average structure is computed that replaces the pair in the set of possible structures. After all possible pairs have been replaced by an average structure, the algorithm moves onto the next iteration with approximately half the original structures to be superimposed. This process is iterated until only one structure remains. The iterations leave a transformation history that can then be applied in series to each original structure to produce a superposition.
PyMSS can also optionally use a GA to further refine the superposition. GAs draw upon an analogy to evolution to “breed” an optimal solution using analogous methods to mutation, reproduction, and natural selection. In PyMSS, the superposition is defined by the translational coordinates applied to each structure. This description is then transformed into GA terms with each translational coordinate becoming a “gene” in a “chromosome”, and the “chromosome” completely defines the state of the superposition. Several possible solutions are randomly generated from the initial state, with each solution represented by a distinct “chromosome”. These “chromosomes” are contained within “individuals”. A “population” of “individuals” then represent a host of different possible solutions, each with a different resulting sRMSD – a quantity inversely proportional to the “individual’s” “fitness”.

“Individuals” are allowed to “breed”, crossing their “genes” into new “individuals”, with mutation occurring as well. After the “population” reaches a certain size, it is cut, with only the most “fit” surviving to the next “generation”. This process is repeated for several “generations” until the most “fit” “individual” is chosen – with this “individual” corresponding to the lowest sRMSD solution.

All the configurable options and complete algorithmic details are found in the PyMSS User Manual in Appendix C.

Other Programs

MASS and MultiProt were obtained from the Nussinov-Wolfson research group at Tel-Aviv University via http://bioinfo3d.cs.tau.ac.il/. MODELLER was from the Sali research group at the University of California, San Francisco via http://salilab.org/modeller/modeller.html. STAMP was obtained from the Barton research group at the University of Dundee according to the instructions in http://www.compbio.dundee.ac.uk/downloads/README.

DSSP specified secondary structures for MASS to align. This was configured by editing mass.config so that “sse-file-type = DSSP”. No other edits to mass.config were made. MASS
was then run from the command line by typing “mass `ls *.pdb`”. The resulting superpositions were constructed from the mass output using the mass2pdb command to extract all alignments.

MultiProt was run with the default configuration, with the structures to be aligned specified on the command line by typing “multiprot.Linux `ls *.pdb`”. The distributed corresp_pdb.pl script converted the resulting conformations to pdb files. All the computed conformations were outputted into separate directories from the solution set containing all structures. Since both MASS and MultiProt report the results of various alignments, the lowest sRMSD of the alignments is reported.

MODELLER uses a short script to read the structure and sequence of each pdb file. The sequences were aligned using the alignment.malign() function, with default parameters, including local_alignment, set to false and the end gap penalty set to 0. This function progressively aligned each sequence in the order read to the consensus sequence of all previous sequences. The alignment.malign3d() function of MODELLER was run using the default parameters, except that write_fit was set to true. The resulting superposition is contained in the files named *_fit.pdb. The MODELLER script used is included in Appendix D.

STAMP uses a domain file consisting of each protein structure listed, with the appropriate Protein Data Bank (PDB) filename and all chains selected. Next, the program PDBSEQ produced a PIR sequence file by specifying “pdbseq –f pir –tl 500 –f family.domain”. This file was used as input for the AMPS program suite. Specifically, the MULTALIGN program was run in pairwise mode, with the blosum matrix file used and the constant set to zero. The program ORDER produced the treewise grouping, again using the blosum matrix and constant set to zero. MULTALIGN was run once again, this time in treewise mode, using the output from ORDER, the blosum matrix, and the constant set to 0 to produce a treewise alignment. This alignment was inputted to ALIGNFIT to produce the superposition. The
command-line command was “alignfit –f family.align –d family.domain –s family_stamp.trans”. This transformation file was finally inputted to the program TRANSFORM by typing “transform –f family_stamp.trans –het –hoh”. All the models produced in all.pdb were then split into individual pdb files. Appendix D contains the command files used for MULTALIGN and ORDER.

Performance of all four programs was judged by values of sRMSD obtained with the key atoms chosen for PyMSS runs.

Results

Glycoside Hydrolase Family 1

Glycoside hydrolase Family 1 (GH1) encompasses many β-glycosidases with (β/α)₈ barrel structures. One chain of each of the fifteen GH1 structures (PDB ID’s 1BGG, 1CBG, 1E1F, 1E6X, 1GNX, 1GOW, 1HXJ, 1NP2, 1OD0, 1PBG, 1QOX, 1QVB, 1UG6, 1V02, and 1VFF) was extracted for alignment with each program. The atom names are those listed in the PDB file, as no IUPAC violations were found. The sRMSD of the CA, CB, and terminal hydrogen bond acceptor atoms of the three amino acid residues (Q20, E166, and E352, the numbering based on Paenibacillus polymyxa β-glucosidase) proximal to the scissile glycosidic bond of a natural substrate was calculated for each alignment (Table 1). Also calculated was the sRMSD of the Cα atoms for G72, T189, S330, and Y343 (four helix-cap residues), in addition to the Cα atom of Q20 and the Cα and Cδ atoms of E166 and E352 of P. polymyxa β-glucosidase, a more separated grouping (Table 2).

Tables 1 and 2 demonstrate that the PSA and NJA within PyMSS obtained the lowest sRMSD values with both closely- and widely-spaced key atoms. However, the results from all programs are not significantly different. The GA did not improve results from the PSA
Table 1. Comparison of sRMSDs (Å), Based on Closely-Spaced Key Residues, of Superpositions Obtained by Multiple Structural Alignment Methods

<table>
<thead>
<tr>
<th>Programs</th>
<th>GH1</th>
<th>Myoglobin</th>
<th>Cytochrome B</th>
</tr>
</thead>
<tbody>
<tr>
<td>MASS</td>
<td>0.27</td>
<td>0.37</td>
<td>N/A&lt;sup&gt;a&lt;/sup&gt;</td>
</tr>
<tr>
<td>MODELLER</td>
<td>0.27</td>
<td>0.37</td>
<td>2.87</td>
</tr>
<tr>
<td>MultiProt</td>
<td>0.27</td>
<td>0.37</td>
<td>0.95</td>
</tr>
<tr>
<td>STAMP</td>
<td>0.28</td>
<td>0.37</td>
<td>1.47</td>
</tr>
<tr>
<td>Pseudo-star algorithm</td>
<td>0.25</td>
<td>0.051</td>
<td>0.60</td>
</tr>
<tr>
<td>Neighbor-joining algorithm</td>
<td>0.25</td>
<td>0.051</td>
<td>0.60</td>
</tr>
<tr>
<td>Genetic algorithm&lt;sup&gt;b&lt;/sup&gt;</td>
<td>0.25</td>
<td>0.22</td>
<td>0.66</td>
</tr>
<tr>
<td>Genetic algorithm&lt;sup&gt;c&lt;/sup&gt;</td>
<td>0.25</td>
<td>0.28</td>
<td>0.59</td>
</tr>
<tr>
<td>Genetic algorithm&lt;sup&gt;d&lt;/sup&gt;</td>
<td>0.25</td>
<td>0.061</td>
<td>0.60</td>
</tr>
<tr>
<td>Genetic algorithm&lt;sup&gt;e&lt;/sup&gt;</td>
<td>0.25</td>
<td>0.067</td>
<td>0.60</td>
</tr>
</tbody>
</table>

<sup>a</sup>Not all structures could be superpositioned.

<sup>b</sup>Starting with final alignment from MODELLER.

<sup>c</sup>Starting with final alignment from STAMP.

<sup>d</sup>Starting with final alignment from the pseudo-star algorithm.

<sup>e</sup>Starting with final alignment from the neighbor-joining algorithm.

and NJA, and reduced sRMSD values of MODELLER and STAMP only to those attained by the PSA and NJA. Figures 2a and 2b show the multiple superpositions computed by the GA when using key atoms near the scissile bond, while Figure 2c shows the superposition when the key atoms are distant from each other. Use of widely-spaced key atoms yields structures that are much better aligned overall, as expected.
Table 2. Comparison of sRMSDs (Å), Based on Widely-Spaced Key Residues, of Superpositions Obtained by Multiple Structural Alignment Methods

<table>
<thead>
<tr>
<th>Programs</th>
<th>GH1</th>
<th>Rhodopsin</th>
<th>Myoglobin</th>
<th>Cytochrome B</th>
</tr>
</thead>
<tbody>
<tr>
<td>MASS</td>
<td>0.28</td>
<td>N/A</td>
<td>0.30</td>
<td>N/A</td>
</tr>
<tr>
<td>MODELLER</td>
<td>0.30</td>
<td>2.11</td>
<td>0.30</td>
<td>2.50</td>
</tr>
<tr>
<td>MultiProt</td>
<td>0.28</td>
<td>3.10</td>
<td>0.30</td>
<td>0.81</td>
</tr>
<tr>
<td>STAMP</td>
<td>0.30</td>
<td>1.50</td>
<td>0.31</td>
<td>1.33</td>
</tr>
<tr>
<td>Pseudo-star algorithm</td>
<td>0.24</td>
<td>1.64</td>
<td>0.23</td>
<td>0.60</td>
</tr>
<tr>
<td>Neighbor-joining algorithm</td>
<td>0.24</td>
<td>1.64</td>
<td>0.23</td>
<td>0.60</td>
</tr>
<tr>
<td>Genetic algorithm\textsuperscript{b}</td>
<td>0.25</td>
<td>1.32</td>
<td>0.26</td>
<td>0.69</td>
</tr>
<tr>
<td>Genetic algorithm\textsuperscript{c}</td>
<td>0.25</td>
<td>1.33</td>
<td>0.25</td>
<td>0.62</td>
</tr>
<tr>
<td>Genetic algorithm\textsuperscript{d}</td>
<td>0.24</td>
<td>1.64</td>
<td>0.24</td>
<td>0.60</td>
</tr>
<tr>
<td>Genetic algorithm\textsuperscript{e}</td>
<td>0.24</td>
<td>1.64</td>
<td>0.24</td>
<td>0.60</td>
</tr>
</tbody>
</table>

\textsuperscript{a}Not all structures could be superpositioned.

\textsuperscript{b}Starting with final alignment from MODELLER.

\textsuperscript{c}Starting with final alignment from STAMP.

\textsuperscript{d}Starting with final alignment from the pseudo-star algorithm.

\textsuperscript{e}Starting with final alignment from the neighbor-joining algorithm.

Rhodopsin

Rhodopsin is a trans-membrane protein with retinal bound to K296, the numbering based on the Bos taurus form (1L9H). Twenty crystal structures of rhodopsin (1CF1, 1DZE, 1F88, 1FQJ, 1GU8, 1GUE, 1GZM, 1H2S, 1H68, 1HZX, 1JFP, 1JGJ, 1L9H, 1LN6, 1QKO, 1QKP, 1QM8, 1QYR, 1U19, and 1XIO) are available from the PDB. Four of these structures, 1AYR, 1CF1, 1FQJ, and 1LN6, are nonhomologous, so they were not further tested. One
chain from each structure was selected. Since retinal binding greatly changes the rhodopsin structure, the residues proximal to retinal were not used as the sole basis for superpositioning by PyMSS. Rather, the CA atoms of four residues, Y43, Q64, N200, and Q225, on the ends of four outer helices were used in addition to two atoms each of W265 (CA and CZ3) and K296 (CA and NZ).

Values of sRMSD of all alignments are presented in Table II. MASS was unable to align all structures. MODELLER and MultiProt gave high sRMSD values, while lower and equal values were obtained by the PSA and NJA, and STAMP gave the lowest values. The GA did not further reduce values obtained by the PSA and NJA, but reduced sRMDS values from MODELLER and STAMP to levels lower than those reached by the PSA and NJA. The superpositioning of the rhodopsin structures by the GA using STAMP-derived initial values is shown on Figure 3.

Myoglobin

Myoglobin is a globular protein used for shuttling oxygen via a heme group. Crystal structures of 224 myoglobins from animals, both vertebrate and invertebrate, are found in the PDB. Of these, all but 5CYT show significant sequence homology. 1WVP lacks the histidine to which the heme is bonded. Twelve of the remaining structures (1BVC, 1BVD, 1IOP, 1J3F, 1UFJ, 1UFP, 1V9Q, 1YMC, 1YOG, 1YOH, 1YOI, and 2CMM) do not include the heme group. A pseudo-random sampling of 114 structures out of the remaining 212 were selected. One chain from each structure was used. The heme iron atom as well as four heme carbon atoms (C3A, C3B, C3C, and C3D) were the basis for superpositioning the closely-spaced case. For more widely-spaced key atoms, the CA atoms of two helix caps, G80 and H119, the CA atoms of two residues around the heme, H64 and H93, and the heme iron atom (numbering based on *Physeter catodon* myoglobin, 1A6G) were used as the basis.
Figure 2. Stereographs of alignments of 15 GH1 β-glycosidases by the GA: a) the active sites, based on active-site key atoms; b) the overall structures, based on active-site key atoms; c) the overall structures, based on widely-spaced key atoms.
Figure 3. Stereograph of alignment of 16 rhodopsins by the GA using MODELLER-derived structures as initial values, based on widely-spaced key atoms.

The sRMSDs of all myoglobin alignments are listed in Tables 1 and 2. MASS, MODELLER, MultiProt, and STAMP yielded equal but different sRMSD values using closely- and widely-spaced key atoms, in both cases higher than those attained by the PSA and NJA. The GA reduced values from MODELLER and STAMP, but not to the levels of the PSA and NJA. It slightly increased sRMSD values from the PSA and NJA. The alignment of the rhodopsin structures by the NJA appears in Figure 4.

Cytochrome B

The PDB lists 86 crystal structures of cytochrome b from bacteria, yeast, vertebrates, and plants. Forty-nine found solely in bacteria, yeast, and vertebrates contain homologous structures around the heme. Of these 49, the 46 (1AQA, 1AW3, 1AWP, 1AXX, 1B5A, 1B5B, 1B5M, 1BFX, 1BLV, 1CXY, 1CYO, 1DO9, 1EHB, 1ES1, 1EUE, 1F03, 1F04, 1FCB, 1HKO, 1I5U, 1I87, 1I8C, 1IB7, 1ICC, 1IET, 1IEU, 1J0Q, 1JEX, 1KBI, 1LCO, 1LDC, 1LJ0, 1LQX, 1LR6, 1LTD, 1M20, 1M2I, 1M2M, 1M59, 1MJ4, 1MNY, 1NX7, 1SH4, 1U9M, 1U9U, and 2AXX) containing the heme group were superimposed. One chain from each structure was selected. The closely-spaced key atoms were the CA and ND atoms of
histidines coordinated to the heme (H39 and H63, numbering based upon 1AQA from Escherichia coli), as well as the CA atoms of G51 and Q53, two residues proximal to the heme. The widely-spaced atoms were the ones listed above plus the CA atoms of H26 and G77 (Figure 5).

The results from all alignments are shown in Tables 1 and 2. The PSA and NJA aligned the structures best, followed by MultiProt, STAMP, and MODELLER. MASS was unable to align all cytochrome B structures. The GA did not improve PSA and NJA results, but brought those from MODELLER and STAMP close to the levels attained by the PSA and NJA.

**PyMSS Timing on Different Platforms**

PyMSS is compatible with any platform that supports Python, including at present Windows, Macintosh OS X, GNU/Linux, DOS, BSD, and all other POSIX-compatible platforms. The GA was timed in triplicate on Windows, Macintosh OS X, and GNU/Linux for cytochrome b
superpositions, while varying the number of generations from 20 to 600 and the number of generations per iteration from 10 to 50. Runtime increases linearly with increasing total number of generations on all platforms, as expected. Also, sRMSD values decrease asymptotically toward the presumptive optimal sRMSD. Choosing the appropriate number of total generations is an exercise in balancing optimality against runtime.

Running the GA on the four protein families with different values of \( M \) (the number of key atoms per structure times the number of structures) but with a constant number of generations and of generations per iteration confirms that runtime increases linearly with \( M^2 \).

**Discussion**

As Tables 1 and 2 indicate, PyMSS, using key atoms chosen manually, can be an efficient means to achieve superpositions better than those achieved by programs using automatically determined key atoms. Furthermore, the GA option within PyMSS often can improve superpositions attained by the latter programs.

The first conclusion is that key atom selection matters. Even the very simple PSA performs well when key atoms are selected by the user, rather than by using an automated

![Figure 5. Stereograph of alignment of 46 cytochrome b structures by the NJA followed by the GA, based on widely-spaced key atoms.](image)
approach. Of the automatic key atom selection routines, none stands out as clearly better than another; rather it depends on the routine needed. MultiProt uses secondary structure information to align structural elements; thus it selected well for cytochrome b, where all structures included one highly conserved fold, but different subdomains and sequences. STAMP bases its atom selection on primary sequence alignment; so it performed well on rhodopsin, which has high sequence homology, but very different folds, depending on whether or not retinal was bound to the active site.

The second conclusion is that some problems are too large for the GA of PyMSS. The superposition of myoglobin by MODELLER and STAMP was largely unchanged by the GA, due to the large number of structures aligned. Since moving most rather than all of the structures in a superposition is sufficient to dramatically reduce the sRMSD, the GA regularly produced a myoglobin superposition with a few structures that were not well superimposed. In effect, the GA would reach a fairly flat region of the problem space and not be able to move the solution in a better direction quickly.

The reason that GA is able to improve upon the superpositions of MODELLER and STAMP for rhodopsin to values lower than those attained by the PSA and NJA (Table 2) but is unable to further reduce the latter values is unclear. The most likely cause is that the superpositions of the former are in different parts of the solution space than those produced by the latter.

Although PyMSS is an effective program for protein rigid-body superposition, it does not even attempt the more challenging problem of detecting structural motifs among proteins with similar structures but different sequences. In cases where key atoms cannot be determined manually, programs that spend more computational time finding them, such as COMPOSER, MASS, MODELLER, MultiProt, SCOP, SSAPm, and STAMP, will work better than PyMSS.
PyMSS, on the other hand, excels at superimposing protein structures based upon the user’s prior or newly acquired knowledge. This allows the user to superimpose highly divergent structures based upon the position of the bound ligand, or to bias the superposition toward a specified region. In these cases, PyMSS will usually produce the best superposition of available programs, attaining the lowest sRMSDs in reasonable computational times.

The rapid prototyping nature of the Python language and the open nature of the GPL community made it possible to write and optimize PyMSS in approximately four weeks. By releasing PyMSS as open-source software, we hope to further enable rapid development of even more useful tools. To this end, PyMSS is available under terms of the GNU General Public License version 2.0 (GPLv2) at http://reillygroup.cheme.iastate.edu/pymss.

Acknowledgments

We are grateful to the National Science Foundation for financially supporting this project, the Nussinov-Wolfson research group for supplying MASS and MultiProt, Andrej Sali for providing MODELLER, and Geoffrey Barton for supplying STAMP.

References


Chapter 4: A Gibbs Free Energy Correlation for Automated Docking of Carbohydrates

A manuscript to be submitted for publication to *The Journal of Computational Chemistry*

ANTHONY D. HILL AND PETER J. REILLY

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

Abstract

Thermodynamic information can be inferred from static atomic configurations. To accurately model the thermodynamics of carbohydrate binding to proteins, a large binding data set has been assembled from the literature. The data set contains information from 262 unique protein-carbohydrate crystal structures for which experimental binding information is known. Hydrogen atoms were added to the structures and training conformations were generated with the automated docking program AutoDock 3.06, resulting in a training set of 225,920 all-atom conformations. In all, 288 formulations of the AutoDock 3.0 free energy model were trained against the data set, testing each of four alternate methods of computing the van der Waals, solvation, and hydrogen-bonding energetic components. The van der Waals parameters from AutoDock 1 produced the lowest errors, and an entropic model derived from statistical mechanics produced the only models with five physically and statistically significant coefficients. Eight models predict the Gibbs free energy of binding with an error of less than 40% the error of any similar models previously published.

**Key words:** AutoDock; carbohydrates; Gibbs free energy
Correspondence to: Peter J. Reilly, Department of Chemical and Biological Engineering, 2114 Sweeney Hall, Iowa State University, Ames, IA 50011-2230. Telephone: 515-294-5968; fax: 515-294-2689; e-mail: reilly@iastate.edu

Contract/grant sponsor: U. S. National Science Foundation; grant number: BES0313878

Abbreviations

Ab., antibody; ABP, arabinose-binding protein; Acr, acarbose; AD, AutoDock; Ara, arabinose; BGal, β-galactosidase; BGlc1, GH1 cellulase; bzl, benzyl-; CBM, carbohydrate binding module; CDA, cytidine deaminase; Cel2, cellobiose; Cel3, cellotriose; ConA, concanavalin A; DANA, 2,3-didehydro-2-dO-NeuNAc acid; deO, deoxy-; DHZ, 3,4-dihydrozebularin; DMDHBAdOMan, 1,2-OMe2-4-(2,4-dihydroxy-butyramido)-4,6-dO2-α-Man; DNJ, 1-deOnojirimycin; EG, endoglucanase; frag., fragment; Fuc, fucose; GA, glucoamylase; Gal, galactose; Gal-sp., galactose-specific; GCP, galactose chemoreceptor protein; Glc, glucose; Gls, β-Glc spirohydantoin; Gox, gluconohydrox-imino-1,5-lactam; GPB, glycogen phosphorylase b; G20, 5-NAc-4-guanidino-6-Me(propyl) carboxamide-4,5-dihydro-2H-pyran-2-carboxylic acid; G28, 5-NAc-4-amino-6-diethyl carboxamide-4,5-dihydro-2H-pyran-2-carboxylic acid; HAR, hepatic asialoglycoprotein receptor; haRMSD, RMSD based on heavy atoms; hb, hydrogen bonding; HGPRT, hypoxanthine-guanine phosphoribosyltransferase; Htp, 4,5,6-trihydroxy-7-hydroxyMe-4,5,6,7-tetrahydro-1H-[1,2,3]triazolo[1,5-α]pyridin-8-ylium; Ifg, isofagomine; Ifgl, Ifg lactam; Iso., isoform; Lac, lactose; Lec., lectin; LNNFP-V, Lac-N-neoFuc5-V; Lox, lactam-oxime; L–R, Laederach–Reilly; LT, heat-labile enterotoxin; Mal, maltose; Man, mannose; Man2, mannobiose; MBP, mannose-binding protein; Me, methyl-; MOAD, Mother of All Databases; mut., mutant; NAc, N-acetyl-; NDase, neuraminidase; Ntz, nojirimycin tetrazole; Oxz, tetrahydrooxazine; PDB, Protein Data Bank; PEG, 2-phenethyl-7-(2,3-dihydrophthal–azine-1,4-dione)-α-Glc; -
R, R-state; RBP, ribose-binding protein; RBSD, high affinity ribose transport protein; Rib, ribose; RMSD, Root mean squared deviation; RMSE, Root mean squared error; st., strain; spHEX, \textit{S. plicatus} $\beta$-N-acetylhexosaminidase; Suc, sucrose; -T, T-state; vDW, van der Waals; WGA, wheat germ agglutinin; Xyl, xylose isomerase; Xyl, xylose; XylL, $\beta$(1,4)-xylanase; Xyl$_2$, xylobiose; Xyn10A, endo-$\beta$(1,4)-xylanase A; Zeb-H$_2$O, 3,4-hydrated pyrimidin-2-one riboside; 4$\alpha$GT, 4-$\alpha$-glucanotransferase; 5D5TGlc, 5-dO-5-thio-$\alpha$-Glc

**Introduction and theory**

Automated docking is among the simplest of the different types of molecular simulations. Molecular dynamics (MD) can explore a particular simulation more thoroughly than automated docking but at the cost of much longer simulation times. In the same time, many more automated dockings than MD simulations can be performed, a feature that aids drug screening, the primary application of automated docking.

Automated docking primarily yields information about the energetics of protein-ligand interactions. Final atomic conformations are the result of minimizing the energy in a force field by one of many possible schemes. AutoDock\textsuperscript{1} (AD) is a software suite for performing automated docking by simulated annealing, local gradient search, and genetic algorithm. The combination of a genetic algorithm with inheritance of local optimizations, yielding a Lamarckian genetic algorithm,\textsuperscript{1} endows AD with very good search performance. In addition to an efficient search algorithm, recent versions of AD include a Gibbs free energy (\(\Delta G_{\text{bind}}\)) correlation derived from molecular conformations:

\[
\Delta G_{\text{bind}} = \Delta G_{\text{vdW}} + \Delta G_{\text{hb}} + \Delta G_{\text{elec}} + \Delta G_{\text{solv}} + \Delta G_{\text{tor}} \tag{1}
\]

where

\[
\Delta G_{\text{vdW}} = \sum_{ij} \left( \frac{A_{pq}}{r_{ij}^{12}} - \frac{B_{pq}}{r_{ij}^6} \right)
\]
\[
\Delta G_{hh} = f_{hh} \sum_{ij} E(\theta_{ij}) \left( \frac{C_{pq}}{r_{ij}^{12}} - \frac{D_{pq}}{r_{ij}^{10}} \right) - \sum_{i} (\Delta G_{p,\text{water}})
\]

\[
\Delta G_{\text{elec}} = f_{\text{elec}} \sum_{ij} \frac{q_i q_j}{\varepsilon(r_{ij}) r_{ij}}
\]

\[
\Delta G_{\text{solv}} = -f_{\text{solv}} \sum_{ij} S_p V_q e^{-r_{ij}^2/2\sigma^2}
\]

\[
\Delta G_{\text{tor}} = f_{\text{tor}} N_{\text{tor}}
\]

where

- \( A_{pq}, B_{pq} \) = Lennard-Jones 12–6 terms for nonbonded contacts between atom types \( p \) and \( q \)
- \( C_{pq}, D_{pq} \) = Lennard-Jones 12–10 terms for hydrogen bonds between atom types \( p \) and \( q \)
- \( E(\theta_{ij}) \) = modifier for the angle between \( i \) and \( j \), with Columbic electrostatic shielding
- \( f_k \) = linear coefficients associated with the free energy change of van der Waals (vdW), hydrogen bond, electrostatic, solvation, and torsional terms, respectively
- \( \Delta G_{p,\text{water}} \) = free energy change of hydrogen bonding between atom type \( p \) and water
- \( N_{\text{tor}} \) = number of rotatable bonds
- \( p, q \) = atom types of atoms \( i \) and \( j \), respectively
- \( q_i, q_j \) = charges of atoms \( i \) and \( j \)
- \( r_{ij} \) = distance between atoms \( i \) and \( j \)
- \( S_p \) = solvation parameter for atom type \( p \), defined as the volume change of solvating atom type \( p \)
- \( V_q \) = atomic volume of atom type \( q \)

\[
\varepsilon(r_{ij}) = F + \frac{H}{l + ke^{-\Delta H_e}}
\]

\( \sigma = 3.5 \text{ Å} \)

where

\( H = e_0 - F \)

\( e_0 = \text{dielectric constant of water at 25°C} = 78.4 \)
\( F, l, k = \text{sigmoidal parameters}, \) –8.5525, 0.003627, and 7.7839, respectively

All the terms in this free energy function have been described previously.\(^1\)–\(^4\) Generically, the terms that are not \( f \)-coefficients can be grouped into a force field. AD makes it easy to change many of the force-field parameters in text files. Therefore it is feasible to use any of several different force fields with AD. The linear nature of the \( f \)-coefficients means that the training of a free energy correlation is a linear regression of docking energies against experimentally determined \( \Delta G_{\text{bind}} \) values.

The AD free energy correlation was earlier trained for use on carbohydrate ligands by Laederach and Reilly.\(^5\) In this study, a training set of 30 carbohydrate–protein pairs was constructed from previous efforts\(^1\)–\(^6\) and a manual search of the literature. The only criteria applied to the protein–ligand pairs in the previous studies were knowledge of their binding energy and a structure in the Protein Data Bank (PDB).

Since 2003 we have made extensive use of the Laederach–Reilly (L–R) free energy model. Proper use of this model requires evaluation of two different energy functions: one a force field based upon AMBER\(^7\) for the actual conformational search, and the second a free energy function to correlate the conformation with its \( \Delta G_{\text{bind}} \) value. The reason for doing this is that the low-energy points in the AMBER-derived model are spatially closer to crystal conformations than low-energy points in the L–R model.

Recently we have begun to perform more rigorous dockings to find more negative energy minima than those used in training the L–R model.\(^8\) In so doing, we have found many conformations that produce lower energies than do the analogous crystal structure conformations when the L–R model is applied. As a result, the L–R model underpredicts \( \Delta G_{\text{bind}} \) (Figure 1).

Closer examination reveals a problem with eq. 1, in that there is no proper entropic term in the free energy function. In eq. 1, \( \Delta G_{\text{tor}} \) is a proxy for the entropy change upon binding. However, the linear model is an improper calculation of the change in entropy upon
ligand binding. To see this, we will decompose the Gibbs free energy function in basic thermodynamic terms:

\[ \Delta G = \Delta H - T \Delta S \]  
\[ \Delta H = \Delta U + P \Delta V \]

Most of the terms in the AD1 free energy function address purely enthalpic terms, and so we can write the change of enthalpy due to binding as eq. 4, where the listed enthalpic terms are the same as the corresponding free energy terms in eq. 1:

\[ \Delta H = f_{vdW} \Delta H_{vdW} + f_{hb} \Delta H_{hb} + f_{elec} \Delta H_{elec} + f_{solv} \Delta H_{solv} \]

The change in entropy follows the Shannon definition used in statistical mechanics:

---

**Figure 1.** L–R Model A plotted with conformations optimized by the Lamarckian genetic algorithm in AD 3.06.
\[ S = -k \sum_l p_l \ln p_l \]  

(5)

where \( k \) is the Boltzmann constant and \( p_l \) is the probability of microstate \( l \) occurring in the given ensemble. If there are \( N \) possible microstates, each with equal probability of occurring, then eq. 5 can be rewritten:

\[ S = k \ln N \]  

(6)

As an approximation of the entropy associated with bound and unbound states, we can observe the following: When bound, the ligand is held tightly, approximating an ensemble with one possible microstate, and therefore \( S_{\text{bound}} = 0 \). When free, the ligand has a number of microstates that are related to its degrees of freedom (DoF), and \( S_{\text{free}} = f_{\text{entropy}} \cdot \ln(N_{\text{DoF}}) \). Thus \( \Delta S_{\text{bind}} = -S_{\text{free}} \). Since we are not guaranteed a constant volume upon ligand binding, we need to also consider one more term, accounted for in the free energy change of solvation:

\[ \Delta H_{\text{solv}} = P \Delta V_{\text{bind}} \]  

(7)

This volume change can be explained as follows. If the ligand is hydrophobic, then it will displace more water than its atomic volume. If the protein is also hydrophobic, then its binding pocket will not be entirely full of water. Upon binding, the volume of water displaced from the binding pocket will not entirely fill the void in the solution, and the system volume will decrease. For hydrophilic/hydrophilic systems, the volume change upon binding would be \( \sim 0 \). Hydrophobic/hydrophilic systems should have a volume change that is slightly negative, but not as negative as hydrophobic/hydrophobic systems. Thus we can rewrite eq. 1:

\[ \Delta G_{\text{bind}} = f_{\text{vdW}} \Delta H_{\text{vdW}} + f_{\text{hb}} \Delta H_{\text{hb}} + f_{\text{elec}} \Delta H_{\text{elec}} + f_{\text{solv}} \Delta V_{\text{solv}} + f_{\text{entropy}} \ln(N_{\text{DoF}}) \]  

(8)

The bound ligand has an entropy very close to, but not equal to, zero. Additionally, not all the DoF’s have the same energetics, and thus they do not have perfectly equal probabil-
ities. For this reason, five different models for the entropic term can be considered. These models consider DoF’s as either the heavy atom torsions or all torsions, with or without the three translational and three rotational DoF’s, as well as a mixed model where the torsions have a different probability than the six transformational DoF’s. The derivation of the mixed model follows.

The relative probability of a microstate occurring due to movement in a transformational DoF is given by $p_{\text{trans}}$, while the probability of due to movement in a torsional DoF is given by $p_{\text{tors}}$. If we set $p_{\text{tors}} = \xi \cdot p_{\text{trans}}$, then

$$p_{\text{trans}} = \frac{1}{6 + \xi N_{\text{tors}}}, \quad p_{\text{tors}} = \frac{\xi}{6 + \xi N_{\text{tors}}}$$

Placing the definitions in eq. 9 into eq. 5 yields

$$S = -k \sum_{i=1}^{6} p_{\text{trans}} \ln p_{\text{trans}} + \sum_{i=1}^{N_{\text{tors}}} p_{\text{tors}} \ln p_{\text{tors}}$$

Rearranging results in the mixed model (eq. 11). As $\xi \to 1$, $\Delta S_{\text{bind}}$ becomes the same as defining $N_{\text{DoF}}$ as the number of torsions plus the six translational DoF’s:

$$\Delta S_{\text{bind}} = -k \left[ \ln(6 + \xi N_{\text{tors}}) - \frac{\xi N_{\text{tors}}}{6 + \xi N_{\text{tors}}} \ln \xi \right]$$

In this paper, we will use a large set of protein/carbohydrate data for which crystal structures have been solved and thermodynamic data are available to find the terms in this model by linear regression. Since the hydrogen-bonding and hybrid entropic terms contain variables that cannot be solved by linear regression, we solve models with rational values of $\Delta G_{\text{p,water}}$ and $\xi$, in a fashion similar to Laederach and Reilly. Additionally, there are several pre-existing formulations of vdW and hydrogen-bonding parameters besides those used in AD, and these will be tested as well. Finally, the solvation model will be parameterized, and different formulations of these parameters will be tested.
The Binding Mother of All Databases (MOAD) is a human-curated database of PDB structures with known binding information and resolutions of <2.5 Å. This database provides more than enough protein-ligand pairs to train a correlation, and it will be used here.

There are two problems in constructing an accurate $\Delta G_{\text{bind}}$ correlation. The first is to accurately predict the conformation of atoms corresponding to the lowest energy, while the second is to correctly predict $\Delta G_{\text{bind}}$ from that conformation. Neither of these problems is trivial. The first problem stems from a lack of data — not the lack of enough crystal structures, but rather the lack of energetics data for atomic configurations that deviate from optimality. In other words, it is impossible to know how the energetic landscape actually changes as atoms move away from their crystallized locations. Unfortunately, no additional number of crystal structures can address this problem, as their atomic configurations are found only in optimal states. Furthermore, even if we were to examine nonideal atomic configurations, we would not have any measured energies to associate with the deviations.

The second problem is simply one of empiricism. Thirty structures, the number used by Laederach and Reilly, are insufficient for the number of coefficients in the regression. A further complication is the various types of proteins used in the regression. For enzymes that catalyze hydrolysis, specific carbohydrate subsites nearly always contain an amino acid with an aromatic side chain to coordinate binding by hydrophobic packing. The subsites in such enzymes are very sharply defined. This is not the case for other carbohydrate binding proteins like lectins, whose binding sites are poorly defined. Spreading only 30 training structures over a diverse set of proteins and ligands with a wide range of binding energies makes it difficult to correctly estimate parameters for all conformations.

To solve the first problem, both optimal and nonoptimal atomic configurations are needed. Because atoms can deviate many ways from their optimal configurations, it actually is necessary to have more nonoptimal structures than optimal ones in a training set. Furthermore, energies must be associated with both optimal and nonoptimal structures. As there is
no method of experimentally measuring the energies of nonoptimal configurations, some estimation method is needed. It can be assumed that there is some positive relationship between molecular deviation, measured in root mean squared deviation (RMSD) from the crystal structure, and the binding enthalpy. However, since hydrogen atoms are not found in crystal structures, their deviations cannot be measured. Therefore, one can either exclude the contributions of hydrogen atoms from the energy estimation of nonoptimal structures, or one can apply a traditional force field to them to measure their contributions. The latter option is circular, since it results in a force field that is being trained by another force field. However, it is viable if iteration results in stability.

**Computational techniques**

The training set was created by starting with the carbohydrate structures used in previous studies.\textsuperscript{1,5,6} In addition to these structures, the Binding MOAD was mined for carbohydrate ligands. The final training set contains 267 protein-ligand pairs with 104 unique carbohydrates from 237 unique PDB entries. Each chain in the multi-chain PDB files was examined, keeping only unique chain–ligand pairs. However, many PDB files contained chains or ligands somewhat distant from each other. In these cases, multiple chains were used in the training set. Non-carbohydrate ligands within 15 Å of the substrate, along with all metal atoms, were included in the protein structures.

Each chain was prepared by the previously used method.\textsuperscript{5} Carbohydrates were prepared by first adding hydrogen atoms using OpenBabel\textsuperscript{13,14} at pH 7.0, and then charges were computed using the restricted Hartree-Fock calculation of GAMESS.\textsuperscript{15} Torsions were defined according to the AD Users Manual.\textsuperscript{16} All nonpolar hydrogen atoms in the protein and ligand were changed from the atomic symbol of “H” to “X”, and the nonorganic atoms were changed to “M”.
All protein–ligand pairs were then docked using an AMBER-derived force field and the Lamarckian genetic algorithm of AD. A total of 1000 unique conformations were generated for each protein–ligand pair. The resulting conformations were clustered with a 1.0-Å RMSD cutoff. The four conformations with lowest heavy-atom RMSDs were sampled, along with the four conformations having the lowest binding energy, the latter defined by evaluating a conformation’s energy with the unweighted AMBER-derived force field. Along with the eight samples, up to eight more samples were taken, excluding the clusters that were already represented by three or more samples. As a result, each protein–ligand pair was represented by conformations from at least four clusters, including the conformations with the lowest binding energies as well as those with the lowest RMSDs from the crystallized ligand.

The contributions of all terms in eq. 1 were evaluated for each sampled conformation by using the epdb command of AD. The Lennard–Jones and hydrogen bonding terms were evaluated using the AMBER99, CHARMM22, MM3PRO, and AD 1.0 (AD1) force fields. In addition, the solvation term was evaluated four ways: either all heavy atoms or only carbon atoms contributed toward the solvation term, as well as two sets of solvation parameters that include interactions for hydrogen atoms and little entropic contribution from the solvating water molecules being freed to the bulk solution. Values of $\Delta G_{p,\text{water}}$ were evaluated as either 0.0, 1.0, 2.5, or 5.0 kcal/mol for $p =$ oxygen and polar hydrogen. Finally, seven entropic models were tested: four as defined previously, and three mixed models with $\xi = \{0.1, 0.33, 0.67, \text{and } 1\}$. As before, multiple linear regression was used to find weights for the linear energy correlation. Linear regression was solved by JMP. Both the undocked and docked training sets are available at http://reillygroup.cheme.iastate.edu/tonyhill/.

Four vDW/hb models, four solvation models, four values of $\Delta G_{p,\text{water}}$, and four values of $\xi$ result in $4^4$ or 256 models. Additionally, to test the dependency of the entropic change on only torsions, the vDW/hb model was fixed to the AD1 model, and the entropy change due to limiting movement is defined as $\ln(N_{\text{Torr}})$. Four solvation models, four values of $\Delta G_{p,\text{water}}$, and
two ways of counting the number of torsions (heavy-atom or all-atom), yields an additional \(4^2 \times 2\) or 32 additional models. Therefore 288 models were evaluated.

To improve the accuracy of the training conformations, only those that were within 1.0 Å haRMSD (RMSD of heavy atoms) of the crystallized structure were considered. In addition, both the over-representation of large clusters and more importantly, the optimization of hydrogen atom placement were considered. Both of these selection criteria were addressed by restricting the training structures to those with the lowest total binding energy conformation in a cluster, using the AD1 force field. As a result, only the most highly optimized member of each cluster within 1.0 Å was used to train the models (Table 1).

**Table 1.** Protein–carbohydrate pairs used in the training of the free energy function. Binding constants were first gathered from the Binding MOAD\(^{12}\) and then double-checked in original publications. \(\Delta G_{\text{bind}}\) was calculated from \(\Delta G = -RT\ln K\). Crystal structures with multiple qualitatively different chains are indicated with superscripts in the PDB ID field.

<table>
<thead>
<tr>
<th>Organism</th>
<th>Protein</th>
<th>Ligand</th>
<th>Formula</th>
<th>PDB ID</th>
<th>(\Delta G_{\text{bind}}) kcal/mol</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Amaranthus caudatus</em></td>
<td>Agglutinin</td>
<td>6-bzl-T-antigen</td>
<td>C(<em>{21})N(</em>{12})O(<em>{13})H(</em>{31})</td>
<td>1JLX</td>
<td>7.90(^{17})</td>
</tr>
<tr>
<td><em>Arachis hypogaea</em></td>
<td>Peanut lec.</td>
<td>Lac</td>
<td>C(<em>{12})O(</em>{12})H(_{22})</td>
<td>2PEL</td>
<td>3.94(^{18,19})</td>
</tr>
<tr>
<td><em>Arthobacter st. B3728</em></td>
<td>Xyl</td>
<td>5D5TGlce</td>
<td>C(<em>{6})O(</em>{9})S(<em>{3})H(</em>{12})</td>
<td>1XLI</td>
<td>2.02(^{20})</td>
</tr>
<tr>
<td><em>Aspergillus awamori X100</em></td>
<td>GA</td>
<td>Acr</td>
<td>C(<em>{25})N(</em>{18})O(<em>{13})H(</em>{33})</td>
<td>1AGM</td>
<td>16.36(^{21,22})</td>
</tr>
<tr>
<td><em>A. awamori X100</em></td>
<td>GA</td>
<td>1-DNJ</td>
<td>C(<em>{6})N(</em>{10})O(<em>{13})H(</em>{13})</td>
<td>1DOG</td>
<td>5.48(^{23,24})</td>
</tr>
<tr>
<td><em>Bacillus agaradhaerens</em></td>
<td>EG 5A</td>
<td>Glc-β(1,4)-Ifg</td>
<td>C(<em>{12})N(</em>{8})O(<em>{16})H(</em>{31})</td>
<td>1OCQ</td>
<td>7.07(^{25})</td>
</tr>
<tr>
<td><em>B. agaradhaerens</em></td>
<td>EG 5A</td>
<td>Cel-β(1,4)-Oxz</td>
<td>C(<em>{17})N(</em>{14})O(<em>{13})H(</em>{31})</td>
<td>1W3L</td>
<td>8.57(^{26})</td>
</tr>
<tr>
<td><em>Bacillus halodurans</em></td>
<td>Bh0236 protein</td>
<td>β(1,4)-Xyl(_2)</td>
<td>C(<em>{10})O(</em>{8})H(_{16})</td>
<td>1W9T(^{(1)})</td>
<td>3.38(^{27})</td>
</tr>
<tr>
<td><em>Bacillus subtilis</em></td>
<td>RBSD</td>
<td>Rib</td>
<td>C(<em>{10})O(</em>{10})H(_{18})</td>
<td>1OGD</td>
<td>4.13(^{28})</td>
</tr>
<tr>
<td><em>Canavalia ensiformis</em></td>
<td>ConA</td>
<td>α(1,2)-Man(_2)</td>
<td>C(<em>{12})O(</em>{11})H(_{22})</td>
<td>1I3H</td>
<td>6.30(^{29,30})</td>
</tr>
<tr>
<td><em>Cellulomonas fimi</em></td>
<td>Xyl(_I)</td>
<td>Xyl-β(1,4)-DNJ</td>
<td>C(<em>{10})N(</em>{9})O(<em>{19})H(</em>{19})</td>
<td>1FH7</td>
<td>7.14(^{31,32})</td>
</tr>
<tr>
<td><em>C. fimi</em></td>
<td>Xyl(_I)</td>
<td>Xyl-β(1,4)-Ifg</td>
<td>C(<em>{10})N(</em>{9})O(<em>{19})H(</em>{19})</td>
<td>1FH8</td>
<td>9.39(^{31,32})</td>
</tr>
<tr>
<td><em>C. fimi</em></td>
<td>Xyl(_I)</td>
<td>Xyl-β(1,4)-Lox</td>
<td>C(<em>{10})N(</em>{9})O(<em>{19})H(</em>{19})</td>
<td>1FH9</td>
<td>8.77(^{31,32})</td>
</tr>
<tr>
<td><em>C. fimi</em></td>
<td>Xyl(_I)</td>
<td>1(N)-imino-Xyl(_2)</td>
<td>C(<em>{10})N(</em>{9})O(<em>{19})H(</em>{19})</td>
<td>1J01</td>
<td>8.82(^{33})</td>
</tr>
<tr>
<td><em>Coprinopsis cinerea</em></td>
<td>Galectin-2</td>
<td>T-antigen</td>
<td>C(<em>{14})N(</em>{10})O(<em>{15})H(</em>{25})</td>
<td>1ULG(^{(3)})</td>
<td>5.73(^{34})</td>
</tr>
<tr>
<td><em>Crotalus atrox</em></td>
<td>Gal-sp. Lec.</td>
<td>Lac</td>
<td>C(<em>{12})O(</em>{11})H(_{22})</td>
<td>1JZN(^{(4)})</td>
<td>5.59(^{35})</td>
</tr>
<tr>
<td><em>Erythrina</em></td>
<td>Lec.</td>
<td>GalNAc</td>
<td>C(<em>{6})O(</em>{12})H(_{15})</td>
<td>1AX0</td>
<td>4.26(^{36,37})</td>
</tr>
<tr>
<td><em>Erythrina coralloendron</em></td>
<td>Lec.</td>
<td>Lac</td>
<td>C(<em>{12})O(</em>{12})H(_{22})</td>
<td>1AX1</td>
<td>4.48(^{36,37})</td>
</tr>
<tr>
<td><em>E. coralloendron</em></td>
<td>Lec.</td>
<td>LacNAc</td>
<td>C(<em>{14})N(</em>{12})O(<em>{15})H(</em>{25})</td>
<td>1AX2</td>
<td>5.44(^{36,37})</td>
</tr>
<tr>
<td><em>E. coralloendron</em></td>
<td>Lec.</td>
<td>Gal</td>
<td>C(<em>{6})O(</em>{12})H(_{12})</td>
<td>1AXZ</td>
<td>4.36(^{36,37})</td>
</tr>
<tr>
<td>Organism</td>
<td>Protein</td>
<td>Ligand</td>
<td>Formula</td>
<td>PDB ID</td>
<td>$\Delta G_{\text{bind}}$ kcal/mol</td>
</tr>
<tr>
<td>------------------------</td>
<td>-------------</td>
<td>-----------</td>
<td>---------------</td>
<td>-----------</td>
<td>----------------------------------</td>
</tr>
<tr>
<td>E. crista-galli</td>
<td>Lec.</td>
<td>Lac</td>
<td>C$<em>{12}$O$</em>{11}$H$_{22}$</td>
<td>1GZC</td>
<td>4.76$^{68}$</td>
</tr>
<tr>
<td>Escherichia coli</td>
<td>Maltoporin</td>
<td>Suc</td>
<td>C$<em>{12}$O$</em>{11}$H$_{22}$</td>
<td>1AF6</td>
<td>2.49$^{39,40}$</td>
</tr>
<tr>
<td>E. coli</td>
<td>ABP mut. P254G</td>
<td>Fuc (α &amp; β)</td>
<td>C$<em>{6}$O$</em>{5}$H$_{12}$</td>
<td>1APB$^{(1)}$</td>
<td>7.94$^{41}$</td>
</tr>
<tr>
<td>E. coli</td>
<td>ABP mut. P254G</td>
<td>Ara (α &amp; β)</td>
<td>C$<em>{6}$O$</em>{5}$H$_{12}$</td>
<td>1BAP$^{(1)}$</td>
<td>9.34$^{41}$</td>
</tr>
<tr>
<td>E. coli</td>
<td>MBP</td>
<td>Mal</td>
<td>C$<em>{12}$O$</em>{11}$H$_{22}$</td>
<td>1ANF</td>
<td>7.44$^{42}$</td>
</tr>
<tr>
<td>E. coli</td>
<td>RBP</td>
<td>Rib</td>
<td>C$<em>{4}$O$</em>{4}$H$_{10}$</td>
<td>2DRI</td>
<td>9.39$^{43}$</td>
</tr>
<tr>
<td>E. coli</td>
<td>RBP mut. G134R</td>
<td>Rib</td>
<td>C$<em>{4}$O$</em>{4}$H$_{10}$</td>
<td>1DRJ</td>
<td>10.09$^{43}$</td>
</tr>
<tr>
<td>E. coli</td>
<td>RBP mut. G143A</td>
<td>Rib</td>
<td>C$<em>{4}$O$</em>{4}$H$_{10}$</td>
<td>1DRK</td>
<td>10.24$^{43}$</td>
</tr>
<tr>
<td>E. coli</td>
<td>CDA</td>
<td>DHZ</td>
<td>C$<em>{4}$O$</em>{2}$H$_{14}$</td>
<td>1CTT</td>
<td>6.17$^{44,45}$</td>
</tr>
<tr>
<td>E. coli</td>
<td>CDA</td>
<td>Zeb</td>
<td>C$<em>{4}$O$</em>{2}$H$_{14}$</td>
<td>1CTU</td>
<td>16.25$^{44,45}$</td>
</tr>
<tr>
<td>E. coli</td>
<td>LTβ</td>
<td>PEPG</td>
<td>C$<em>{22}$N$</em>{2}$O$_{24}$</td>
<td>1EEF$^{(3)}$</td>
<td>3.98$^{46}$</td>
</tr>
<tr>
<td>E. coli</td>
<td>F17-ag lec.</td>
<td>GalNAc</td>
<td>C$<em>{4}$O$</em>{2}$H$_{15}$</td>
<td>1O9W</td>
<td>3.99$^{47}$</td>
</tr>
<tr>
<td>E. coli</td>
<td>AraC</td>
<td>Fuc</td>
<td>C$<em>{4}$O$</em>{2}$H$_{12}$</td>
<td>2AAC</td>
<td>3.05$^{48,49}$</td>
</tr>
<tr>
<td>E. coli</td>
<td>AraC</td>
<td>L-Ara</td>
<td>C$<em>{4}$O$</em>{2}$H$_{12}$</td>
<td>2ARC</td>
<td>3.44$^{49,50}$</td>
</tr>
<tr>
<td>E. coli</td>
<td>GCP</td>
<td>Glc</td>
<td>C$<em>{4}$O$</em>{2}$H$_{12}$</td>
<td>2GBP</td>
<td>9.41$^{51,54}$</td>
</tr>
<tr>
<td>Homo sapiens</td>
<td>Galectin-1 mut. C2S</td>
<td>Lac</td>
<td>C$<em>{12}$O$</em>{11}$H$_{22}$</td>
<td>1W6O$^{(1)}$</td>
<td>4.56$^{55}$</td>
</tr>
<tr>
<td>H. sapiens</td>
<td>Cdc42 and Par6</td>
<td>GTP</td>
<td>C$<em>{10}$N$</em>{4}$O$<em>{3}$P$</em>{3}$H$_{17}$</td>
<td>1NF3</td>
<td>9.95$^{56}$</td>
</tr>
<tr>
<td>Influenza A virus</td>
<td>NDase</td>
<td>DANA</td>
<td>C$<em>{11}$N$</em>{4}$O$<em>{7}$H$</em>{17}$</td>
<td>1F8B</td>
<td>7.36$^{57,58}$</td>
</tr>
<tr>
<td>Influenza A virus</td>
<td>NDase</td>
<td>4-amino-DANA</td>
<td>C$<em>{11}$N$</em>{4}$O$<em>{7}$H$</em>{18}$</td>
<td>1F8C</td>
<td>10.09$^{57,58}$</td>
</tr>
<tr>
<td>Influenza A virus</td>
<td>NDase</td>
<td>9-amino-DANA</td>
<td>C$<em>{11}$N$</em>{4}$O$<em>{7}$H$</em>{18}$</td>
<td>1F8D</td>
<td>4.63$^{57,58}$</td>
</tr>
<tr>
<td>Influenza A virus</td>
<td>NDase</td>
<td>4,9-diamino-DANA</td>
<td>C$<em>{11}$N$</em>{4}$O$<em>{7}$H$</em>{19}$</td>
<td>1F8E</td>
<td>6.58$^{57,58}$</td>
</tr>
<tr>
<td>M. musculus</td>
<td>Adenosine deaminase</td>
<td>1-DzA</td>
<td>C$<em>{11}$N$</em>{4}$O$<em>{4}$H$</em>{14}$</td>
<td>1ADD</td>
<td>9.20$^{60,61}$</td>
</tr>
<tr>
<td>Narcissus pseudonarcissus</td>
<td>Ab. S-20-4 Fab frag.</td>
<td>DMDHBAAdOMan</td>
<td>C$<em>{12}$N$</em>{4}$O$<em>{3}$H$</em>{3}$</td>
<td>1F4X</td>
<td>7.62$^{62,63}$</td>
</tr>
<tr>
<td>Oryctolagus cuniculus</td>
<td>GPB</td>
<td>Glc</td>
<td>C$<em>{6}$O$</em>{6}$H$_{12}$</td>
<td>2GPB</td>
<td>3.68$^{64}$</td>
</tr>
<tr>
<td>O. cuniculus</td>
<td>GPB-T</td>
<td>Htp</td>
<td>C$<em>{6}$O$</em>{5}$H$_{12}$</td>
<td>1AXR</td>
<td>4.46$^{67}$</td>
</tr>
<tr>
<td>O. cuniculus</td>
<td>GPB-T</td>
<td>aminocarabamate</td>
<td>C$<em>{6}$O$</em>{5}$H$_{16}$</td>
<td>1B4D</td>
<td>6.55$^{68}$</td>
</tr>
<tr>
<td>O. cuniculus</td>
<td>GPB-T</td>
<td>Glc</td>
<td>C$<em>{6}$O$</em>{5}$H$_{12}$</td>
<td>1HLF</td>
<td>7.57$^{69}$</td>
</tr>
<tr>
<td>O. cuniculus</td>
<td>GPB-T</td>
<td>Bzurea</td>
<td>C$<em>{14}$N$</em>{5}$O$<em>{5}$H$</em>{18}$</td>
<td>1K06</td>
<td>7.28$^{70}$</td>
</tr>
<tr>
<td>O. cuniculus</td>
<td>GPB-T</td>
<td>Bzurea</td>
<td>C$<em>{14}$N$</em>{5}$O$<em>{5}$H$</em>{18}$</td>
<td>1K08</td>
<td>7.28$^{70}$</td>
</tr>
<tr>
<td>O. cuniculus</td>
<td>GPB-R + PO$_{4}$</td>
<td>Ntz</td>
<td>C$<em>{6}$O$</em>{5}$H$_{10}$</td>
<td>1NOI</td>
<td>5.83$^{71}$</td>
</tr>
<tr>
<td>O. cuniculus</td>
<td>GPB-T + PO$_{4}$</td>
<td>Ntz</td>
<td>C$<em>{6}$O$</em>{5}$H$_{10}$</td>
<td>1NOJ</td>
<td>5.83$^{71}$</td>
</tr>
<tr>
<td>O. cuniculus</td>
<td>GPB-T</td>
<td>Ntz</td>
<td>C$<em>{6}$O$</em>{5}$H$_{10}$</td>
<td>1NOK</td>
<td>4.30$^{71}$</td>
</tr>
<tr>
<td>O. cuniculus</td>
<td>GPB-T</td>
<td>Glc</td>
<td>C$<em>{6}$O$</em>{5}$H$_{12}$</td>
<td>2GPB</td>
<td>3.68$^{72,73}$</td>
</tr>
<tr>
<td>Pseudomonas</td>
<td>Lec. PAI-L</td>
<td>Gal (α &amp; β)</td>
<td>C$<em>{6}$O$</em>{5}$H$_{12}$</td>
<td>1OKO$^{(1)}$</td>
<td>6.18$^{75,76}$</td>
</tr>
<tr>
<td>Organism</td>
<td>Protein</td>
<td>Ligand</td>
<td>Formula</td>
<td>PDB ID</td>
<td>$\Delta G_{\text{bind}}$ kcal/mol</td>
</tr>
<tr>
<td>--------------------------</td>
<td>---------------</td>
<td>--------------</td>
<td>-------------</td>
<td>------------</td>
<td>----------------------------------</td>
</tr>
<tr>
<td><em>P. aeruginosa</em></td>
<td>Lec. PAII-L</td>
<td>Fuc</td>
<td>C$_6$O$_3$H$_12$</td>
<td>1GZT</td>
<td>7.03$^{77,78}$</td>
</tr>
<tr>
<td><em>P. aeruginosa</em></td>
<td>Lec. PAII-L</td>
<td>Fuc</td>
<td>C$_6$O$_3$H$_12$</td>
<td>1UZV</td>
<td>7.03$^7$</td>
</tr>
<tr>
<td><em>P. aeruginosa</em></td>
<td>Lec. PAII-L</td>
<td>LNNFP-V</td>
<td>C$<em>{32}$N$</em>{23}$O$<em>{15}$H$</em>{55}$</td>
<td>1W8F</td>
<td>8.44$^7$</td>
</tr>
<tr>
<td><em>Ralstonia solanacearum</em></td>
<td>Lec. RSL</td>
<td>(\alpha)-L-MeFuc</td>
<td>C$_7$O$_3$H$_14$</td>
<td>2BT9(1)</td>
<td>8.44$^{80}$</td>
</tr>
<tr>
<td><em>Rattus norvegicus</em></td>
<td>Lec. HAR</td>
<td>GalNAc</td>
<td>C$_8$N$_1$O$_3$H$_5$</td>
<td>1BCP(2)</td>
<td>4.65$^{81}$</td>
</tr>
<tr>
<td><em>Salmonella</em></td>
<td>LT2 NDase</td>
<td>DANA</td>
<td>C$_{11}$N$_2$O$<em>3$H$</em>{17}$</td>
<td>2SIM</td>
<td>4.66$^{82,83}$</td>
</tr>
<tr>
<td><em>Sinapis alba</em></td>
<td>Myrosinase</td>
<td>Ntz</td>
<td>C$_{6}$O$<em>3$H$</em>{10}$</td>
<td>1E6Q</td>
<td>4.30$^{84}$</td>
</tr>
<tr>
<td><em>S. alba</em></td>
<td>Myrosinase</td>
<td>Gox</td>
<td>C$_{6}$N$_2$O$<em>3$H$</em>{12}$</td>
<td>1E6S</td>
<td>4.39$^{84}$</td>
</tr>
<tr>
<td><em>S. alba</em></td>
<td>Myrosinase</td>
<td>Gox</td>
<td>C$_{6}$N$_2$O$<em>3$H$</em>{12}$</td>
<td>1E72</td>
<td>4.39$^{84}$</td>
</tr>
<tr>
<td><em>Streptomyces lividans</em></td>
<td>Xyn10A</td>
<td>Xyl-Ifgl</td>
<td>C$_{10}$N$<em>1$O$</em>{17}$</td>
<td>1OD8</td>
<td>6.95$^{85}$</td>
</tr>
<tr>
<td><em>S. lividans</em></td>
<td>Xyn10A</td>
<td>Xyl-Ifg</td>
<td>C$_{10}$N$<em>1$O$</em>{19}$</td>
<td>1V0K</td>
<td>6.00$^{86}$</td>
</tr>
<tr>
<td><em>S. lividans</em></td>
<td>Xyn10A</td>
<td>Xyl-Dnj</td>
<td>C$_{10}$N$<em>1$O$</em>{19}$</td>
<td>1V0L</td>
<td>8.62$^{86}$</td>
</tr>
<tr>
<td><em>Streptomyces plicatus</em></td>
<td>spHEX</td>
<td>GalNAc-Ifg</td>
<td>C$_{8}$N$_2$O$<em>3$H$</em>{16}$</td>
<td>1JAK</td>
<td>7.59$^{87}$</td>
</tr>
<tr>
<td><em>Sulfolobus solfataricus</em></td>
<td>BGal</td>
<td>Gox</td>
<td>C$_{6}$N$_2$O$<em>3$H$</em>{12}$</td>
<td>1UWU</td>
<td>8.16$^{88}$</td>
</tr>
<tr>
<td><em>Sus scrofa</em></td>
<td>Peptide YY</td>
<td>Fuc ((\alpha) &amp; (\beta))</td>
<td>C$_{7}$O$<em>3$H$</em>{12}$</td>
<td>1ABF(1)</td>
<td>7.39$^{89-93}$</td>
</tr>
<tr>
<td><em>Thermoanaerobacter</em></td>
<td>HGPRT</td>
<td>IMP</td>
<td>C$_{10}$N$_2$O$_3$P$<em>3$H$</em>{13}$</td>
<td>1YFZ(1)</td>
<td>5.93$^{94}$</td>
</tr>
<tr>
<td><em>T. maritima</em></td>
<td>Cel5A</td>
<td>Oxz</td>
<td>C$<em>{5}$N$<em>1$O$</em>{11}$H$</em>{11}$</td>
<td>1W3J</td>
<td>8.61$^{26}$</td>
</tr>
<tr>
<td><em>T. maritima</em></td>
<td>CBM9</td>
<td>Cel$_5$</td>
<td>C$<em>{12}$O$</em>{1}$H$_{22}$</td>
<td>1I82</td>
<td>8.66$^{96,97}$</td>
</tr>
<tr>
<td><em>T. maritima</em></td>
<td>CBM9</td>
<td>(\beta)Glc</td>
<td>C$_{6}$N$_2$O$<em>3$H$</em>{12}$</td>
<td>1I8A</td>
<td>5.63$^{96,97}$</td>
</tr>
<tr>
<td><em>T. maritima</em></td>
<td>BGlc1</td>
<td>Ifg</td>
<td>C$_{6}$N$_2$O$<em>3$H$</em>{13}$</td>
<td>1OIFF</td>
<td>6.95$^{98}$</td>
</tr>
<tr>
<td><em>T. maritima</em></td>
<td>BGlc1</td>
<td>1-DNJ</td>
<td>C$_{6}$N$_2$O$<em>3$H$</em>{13}$</td>
<td>1OIM</td>
<td>9.77$^{98}$</td>
</tr>
<tr>
<td><em>Triticum aestivum</em></td>
<td>WGA3</td>
<td>(\beta)1,4(GlcNAc)$_2$</td>
<td>C$_{16}$N$_2$O$<em>3$H$</em>{18}$</td>
<td>1K7U(1)</td>
<td>5.05$^{99}$</td>
</tr>
<tr>
<td><em>Vibrio cholerae</em></td>
<td>Sialidase</td>
<td>Sialic acid</td>
<td>C$_{11}$N$_1$O$<em>3$H$</em>{19}$</td>
<td>1W0O</td>
<td>6.17$^{100}$</td>
</tr>
<tr>
<td><em>V. cholerae</em></td>
<td>Sialidase</td>
<td>Sialic acid</td>
<td>C$_{11}$N$_1$O$<em>3$H$</em>{19}$</td>
<td>1W0P</td>
<td>6.17$^{100}$</td>
</tr>
<tr>
<td><em>Zea mays</em></td>
<td>BGlc mut. E191D</td>
<td>Dhurrin</td>
<td>C$_{14}$N$_1$O$<em>3$H$</em>{17}$</td>
<td>1E55</td>
<td>5.62$^{101}$</td>
</tr>
</tbody>
</table>

\(^{(1)}\)\(n+1\) different chains/ligand pairs in the PDB are modeled.

Results

Once constructed, the training set contained 225 protein-ligand pairs. Of the binding data, 40% were \(K_d\)'s, 36% were \(K_i\)'s, 13% were \(K_a\)'s, 8% were \(IC_{50}\)'s, and the remaining 3% were generic \(K\) values, typically the Michaelis-Menten \(K_m\), but sometimes a solution partition constant (Figure 2). \(K_d\)'s are dissociation constants, typically measured by isothermal titration calorimetry (ITC). \(K_i\)'s are kinetic inhibition binding constants. \(K_a\)'s are association constants, typically also measured by ITC. \(IC_{50}\) is the concentration at which 50% of the enzyme
Figure 2. Histograms of: a) type of binding data, b) published $\Delta G_{\text{bind}}$ of training set, c) haRMSDs of docked structures, and d) total energy of the docked structures, using the AD1 force field.
activity is inhibited. The thermodynamic meaning of $IC_{50}$ depends on the inhibition model, but in many cases it has the same meaning as $K_i$. $\Delta G_{bind}$ values ranged from $-1.29$ kcal/mol to $-16.26$ kcal/mol, with the first and third quartiles being $-4.6$ and $-8.1$ kcal/mol, respectively. Finally, for most of the dockings $-90 < E_{bind} < -150$ kcal/mol and skewed from normal.

Of the 225,920 docked configurations, 50% docked to within 3 Å haRMSD of the crystallized ligand, with the highest haRMSD being 50 Å (Figure 2). Many of the final docked conformations could be clustered into groups that deviated from one another by <1 Å, resulting in 673 unique clusters that have very low energy. Of the lowest energy dockings, 37% are within 1.0 Å haRMSD of the crystallized ligand. A total of 249 of the dockings are both the lowest energy member of a cluster and within 1.0 Å haRMSD of the reference structure; it is composed of 115 unique protein–ligand pairs. This set was used as the data for computing the linear regression.

The best fitting model has a root mean squared error (RMSE) of 2.02 kcal/mol, while the worst has an RMSE of 2.40 kcal/mol (Table 2). The three L–R models listed in Table 3 of the 2003 paper were also applied to the data, with RMSE’s ranging from 4.84, 5.16, and 5.23 kcal/mol (Table 2). There is a bias in the models, as very strongly binding protein–carbohydrate pairs are predicted to not bind strongly enough, while weakly binding protein–carbohydrate pairs are predicted to bind too strongly (Figure 3).

**Table 2.** Parameter estimates for models that best estimated $\Delta G_{bind}$ values. All models shown use the AD1 force field. L–R models originally published by Laederach and Reilly.

<table>
<thead>
<tr>
<th>Model</th>
<th>$f_{vdW}$</th>
<th>$f_{hb}$</th>
<th>$f_{elec}$</th>
<th>$f_{entropy}$</th>
<th>$f_{sol}$</th>
<th>RMSE (kcal/mol)</th>
</tr>
</thead>
<tbody>
<tr>
<td>L–R A (1,2,4,7,14)</td>
<td>0.0737</td>
<td>0.0566</td>
<td>0.333</td>
<td>0.0497</td>
<td>0.119</td>
<td>4.84</td>
</tr>
<tr>
<td>L–R B (1,2,4,7,14)</td>
<td>0.0801</td>
<td>0.0526</td>
<td>0.345</td>
<td>0.114</td>
<td>0.140</td>
<td>5.23</td>
</tr>
<tr>
<td>L–R C (1,6,7,14)</td>
<td>0.0307</td>
<td>0.0332</td>
<td>0.277</td>
<td>0.0976</td>
<td>0.0274</td>
<td>5.16</td>
</tr>
<tr>
<td>Q (3,8,9)</td>
<td>0.0582</td>
<td>0.0483</td>
<td>0.0474</td>
<td>–</td>
<td>0.0779</td>
<td>2.05</td>
</tr>
<tr>
<td></td>
<td>R (3,8,10)</td>
<td>S (3,8,11)</td>
<td>T (3,8,12)</td>
<td>JA (2,3,8,13)</td>
<td>JB (2,4,8,13)</td>
<td>JC (2,5,8,13)</td>
</tr>
<tr>
<td>---</td>
<td>------------</td>
<td>------------</td>
<td>------------</td>
<td>--------------</td>
<td>--------------</td>
<td>--------------</td>
</tr>
<tr>
<td></td>
<td>0.0601</td>
<td>0.0603</td>
<td>0.0603</td>
<td>0.0740</td>
<td>0.0796</td>
<td>0.0861</td>
</tr>
<tr>
<td></td>
<td>0.0508</td>
<td>0.0510</td>
<td>0.0510</td>
<td>0.0526</td>
<td>0.0447</td>
<td>0.0227</td>
</tr>
<tr>
<td></td>
<td>0.0496</td>
<td>0.0498</td>
<td>0.0498</td>
<td>0.0511</td>
<td>0.0631</td>
<td>0.0783</td>
</tr>
<tr>
<td></td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>0.510</td>
<td>0.364</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.0777</td>
<td>0.0777</td>
<td>0.0777</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2.06</td>
<td>2.06</td>
<td>2.06</td>
</tr>
</tbody>
</table>

(1) Charges computed using MOPAC rather than GAMESS; (2) torsional model counting only heavy-atom dihedrals; (3) $\Delta G_{p,\text{water}} = 0.0$; (4) $\Delta G_{p,\text{water}} = 1.0$; (5) $\Delta G_{p,\text{water}} = 2.5$; (6) $\Delta G_{p,\text{water}} = 5.0$; (7) solvation model, carbon only, AD1; (8) solvation model, heavy atom, AD1; (9) mixed entropic model (MEM), $\xi = 0.1$; (10) MEM, $\xi = 0.33$; (11) MEM, $\xi = 0.67$; (12) MEM, $\xi = 1$; (13) unmixed entropic model; (14) Linear entropic model; —, parameter not statistically significant ($\alpha = 0.1$).

The first trend in the Table 2 data is that the parameters for vdW and hydrogen bonding (hb) terms affect the model error the most strongly. Models using the AD1 parameters have an average error of 2.06 kcal/mol, while AMBER99, CHARMM22, and MM3Pro parameter sets average 2.20, 2.38, and 2.36 kcal/mol RMSE, respectively. Furthermore, $f_{vdW}$ is statistically insignificant for every model using the CHARMM22 and MM3Pro parameters.

Second, the average RMSE is unaffected by the remaining parts of the model. It is between 2.22 and 2.25 kcal/mol when holding any of the non-force field model components constant and switching between the models.

The $f_{\text{entropy}}$ and $f_{\text{solv}}$ coefficients are the most problematic. Of the 288 models, only eleven produce $f_{\text{entropy}}$ values that are both positive and significantly different than zero. However,
compared to linear entropic models (data not shown), the logarithmic entropic model produces many more models with an entropic coefficient statistically different from zero. In fact, for only 13% of the models is the logarithmic entropy coefficient insignificant. Unfortunately, the mixed entropic models produce only negative coefficients. Such values are not physically meaningful, since the entropy change upon binding is always negative, leading to an increase in Gibbs free energy, and a negative entropic coefficient does not agree with this reality.

Similarly, only 54 models produce an $f_{solv}$ value both positive and significantly different than zero. Of the 54, 44 correspond to the heavy-atom AD1 solvation model. The remaining ten correspond to solvation model 1, an all-atom model where burying of hydrophilics is also punished. Of the 44 meaningful heavy-atom AD1 solvation models, 40 correspond to models

**Figure 3.** Residual error of the models plotted against published $\Delta G_{bind}$ values. Model JA (dark blue), L–R Model A (light blue), L–R Model B (magenta), and L–R Model C (green).
using the AD1 or AMBER99 force fields. Nine of the ten meaningful solvation model results use the CHARMM22 force field, and the remaining one uses the MM3Pro force field.

Thus, of the 288 models trained, ten result in coefficients that make sense physically (Table 2), with RMSE’s between 2.03 and 2.08 kcal/mol. Two of the ten have $\Delta G_{tor}$ values statistically significant at $\alpha = 0.1$. All ten of the best models use the AD1 force field parameters and the AD1 heavy-atom solvation model. They all contain $f_{vdW}$ and $f_{hb}$ coefficients similar in value to those found in the L–R models. Values of the $f_{elec}$ and $f_{solv}$ coefficients are greatly reduced in all ten models from those found by Laederach and Reilly. The $f_{entropy}$ value is much larger in models JA and JB than the corresponding coefficient in the L–R models. This $f_{entropy}$ value is also much larger than any of the other estimated coefficients.

**Discussion**

In general, models using the AD1 force field result in much smaller RMSE’s than any model previously published. In the range $-4$ kcal/mol $\leq \Delta G_{bind} \leq -8.5$ kcal/mol, Model JA produces an error of 1.47 kcal/mol, 34% the error of the best L–R model in that range. Even over the entire binding energy range, Model JA has an error that is less than half of the best L–R model. Furthermore, Model JA has a smaller range of residuals than any of the L–R models. As much as 20 kcal/mol is needed to capture the residual of the L–R models around the $-4.5$ kcal/mol $\Delta G_{bind}$ range. The free energies predicted by Model JA never deviate more than 10 kcal/mol from one another. Furthermore, Model JA also never predicts a positive $\Delta G_{bind}$ value, while each of the L–R models predicts positive $\Delta G_{bind}$ values for at least one protein–ligand pair, and L–R model C predicts positive $\Delta G_{bind}$ values for 24 protein–ligand pairs.

Model JA also has less bias than the L–R models (Figure 3). This is mainly due to the correct modeling of the entropic term, since models at the low-energy range of the scale also have many rotatable bonds, and logarithmic modeling of their contribution to the free
ligand’s entropy punishes the binding of that ligand less severely. This nonlinear entropic change lessens the positive trend of tightly binding conformations.

Proper modeling of the entropic term is very important to a robust model. Without the logarithmic dependence on the DoF, it was not possible to perform a linear regression in which all five coefficients were both physically meaningful and statistically significant.

The $\Delta G_{\text{bind}}$ values of the L–R models are largely inaccurate on this larger data set, with systemically incorrect predictions (Figure 3). Much of this systemic error seems to originate from a too small data set, since the errors arise upon its expansion. Although the models presented in this paper do not have the same systemically over- or under-predicted values, they do appear to bias the predictions so that the plot of the residuals has a negative slope when plotted against actual $\Delta G_{\text{bind}}$ values. Since Figure 3 is merely a two-dimensional projection of six-dimensional regression space, the bias seen is not an indication of a bad regression. When the residual was plotted against each of the five model terms, the data had no bias (not shown). The residual plot is flat over the entire range of the component’s values.

These systemic problems are not corrected with any linear regression performed on the data, either with or without an intercept, regardless of solvation parameters, entropic model, and force field. The problem partially stems from the lack of correlation between model components and the $\Delta G$ being modeled. $R^2$ correlation values vary between $1.3 \times 10^{-3}$ and 0.16 for all model components and the experimental value of $\Delta G_{\text{bind}}$. This problem was not present in the data used by Laederach and Reilly in 2003, where $R^2$ values reached 0.42 for vdW and solvation terms against $\Delta G_{\text{bind}}$. Despite the low individual correlation coefficients, the model is significant at an $\alpha$ threshold of $<<0.01$.

The problem of bias is also partially due to poor modeling of the solvation and entropic terms. In reality, the desolvation of a ligand and binding pocket not only changes the volume of the system, but it also frees water molecules of solvation that were in the binding contact area. Entropy increases as these water molecules join the bulk water. Taking this additional
entropy into account should affect models at the tighter-binding end of the scale more, since while internal energy changes linearly with the number of contributing atoms, entropy does not. Thus, the bias seen in Figure 3 should decrease with proper modeling of the entropy of solvation, which would use the full three-parameter Souten et al. solvation model.\textsuperscript{4} In addition it should be possible to fit the entropic terms to that data, since the volume change upon binding should be easily calculated from the Lennard-Jones correlation, and many binding studies include a measurement of $\Delta S_{\text{bind}}$.

The data set is largely composed of complexes with moderate binding constants. An expanded data set would contain more tightly bound carbohydrates, but these data are not yet available. Further crystallographic and thermodynamic studies of strongly bound systems would mainly be of inhibitors. Several crystal structures of acarbose-derived inhibitors complexed with enzymes exist, but no thermodynamic information accompanies them.

Model JA does not solve the problem of “holes” in the energetic objective function, where the docked complex has a lower energy than the reference, but a higher haRMSD from the crystallized ligand. In this case the reference is the lowest-energy cluster member whose haRMSD is $<1.0$ Å from the crystallized ligand. The problem is similar to that with unweighted binding energies (all $f$-coefficients set to 1.0 and $\Delta G_{\text{p,water}} = 0$). For the latter case, 50 of the 107 protein–ligand pairs yielding the lowest energies when docked also have the smallest haRMSD values, compared to 39 for Model JA. To further evaluate this result, we counted the number of conformations that fall into a “hole” for each protein–ligand pair in the training set. In 42% of the cases, the unweighted binding energy allowed fewer conformations in a “hole” than Model JA, compared to 20% giving the opposite, and 38% giving the same number.

To actually treat the problem of “holes”, the energetics of nonoptimal configurations must be addressed. Given the difficulty in obtaining both physically and statistically signif-
icant thermodynamic parameters, this problem was not treated here. Before fixing the “holes” in the energy landscape can considered, the bias in the model must be solved.

Finally, despite the difficulties in modeling $\Delta G_{\text{bind}}$, we have accomplished two important results. First, we have compiled and corrected a large binding dataset, composed primarily of diffusion and inhibition data, for carbohydrate crystal structures. Secondly, we have presented a model fitting reality with half the error of previously published models. Particularly in the range $-4 \text{ kcal/mol} \leq \Delta G_{\text{bind}} \leq -8.5 \text{ kcal/mol}$, the error rate is as low as 1.38 kcal/mol.

Finally, the models presented in this paper can be used to as an effective energy function for docking. Appropriate AD parameter files can be found in Appendix E.

Acknowledgments

The authors gratefully thank the National Science Foundation for funding this research and the Iowa State University High Performance Computing Facility for making its equipment available to conduct it.

References


Chapter 5: Computational Analysis of Glycoside Hydrolase Family 1 Selectivities

A manuscript to be submitted for publication to Biopolymers

ANTHONY D. HILL AND PETER J. REILLY

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

Abstract

Glycoside hydrolase family 1 consists of β-glucosidases, β-galactosidases, 6-phospho-β-galactosidases, myrosinases, and other enzymes having similar primary and tertiary structures yet diverse specificities. Among these enzymes, β-glucosidases hydrolyze cellobiose to glucose, and therefore they are key players in any cellulose to glucose process. All members of this family attack β-glycosidic bonds between a pyranosyl glycon and an aglycon, but they have little specificity for the aglycon or for the bond configuration. Furthermore, glycon specificity is not absolute. Seventeen family members, seven β-glucosidases, two cyanogenic β-glucosidases, one 6-phospho-β-galactosidase, two myrosinases, and five β-glycosidases, now have known tertiary structures. We have used automated docking to computationally bind disaccharides with mannosyl, allosyl, glucosyl, galactosyl, 6-phosphogalactosyl, and 6-phosphoglucosyl pyranosyl glycons, all linked by β-(1,2), β-(1,3), β-(1,4), and β-(1,6)-glycosidic bonds to β-glucopyranoside aglycons, along with β-(1,1-thio)-allopypyransyl, -galactopyranosyl, -glucopyranosyl, and -mannopyranosyl β-glucopyranosides, into all of these structures to investigate the structural determinants of these enzyme specificities. Five active-site residues, Thr194, Phe205, Asn285, Arg336, and Asn376 (Zea mays β-glucosidase numbering), control a significant amount of glycon specificity.
Introduction

D-Glucose has a large market in the food industry and is also a precursor to high-fructose syrup, ethanol, and other chemicals. It is mainly made by α-amylase-catalyzed hydrolysis of starch to maltooligosaccharides followed by glucoamylase-catalyzed hydrolysis of the latter to glucose. The use of starch to produce glucose in the United States is severely limited by its availability, as the maize from which nearly all starch is derived is also used for animal feed.

Another glucose source is cellulose, which is in much larger supply but whose hydrolysis is greatly slowed by its inaccessibility, as it is complexed with lignin and hemicelluloses, and by the low rate by which either acid or enzymes hydrolyze the β-(1,4) glycosidic bonds linking the glucosyl residues in cellulose. A very large effort is underway to remove barriers preventing the economical conversion of cellulose to glucose. Some of this involves increasing the rate at which cellulases hydrolyze this conversion.

Enzymes having glycoside hydrolase activity are divided into over 100 families based upon amino acid sequence similarities. Members of glycoside hydrolase family 1 (GH1) cleave β-glycosidic bonds in cellobioigosaccharides and other small substrates to produce monosaccharides. Nearly all of them catalyze hydrolysis by a retaining mechanism with both a catalytic proton donor/base and a catalytic nucleophile. Amino acid sequences around these residues for the seventeen GH1 members with crystal structures are shown in Figure 1. They
are the most highly conserved residues among GH1 enzymes, indicating their primary responsibility for catalytic activity.

GH1 enzymes have very diverse substrate specificities despite their high catalytic domain sequence homology (Figures 1 and 2, Table 1). Among them are β-glucosidases (EC 3.2.1.21), β-galactosidases (EC 3.2.1.23), β-mannosidases (EC 3.2.1.25), β-glucuronidases (EC 3.2.1.31), β-D-fucosidases (EC 3.2.1.38), phlorizin hydrolases (EC 3.2.1.62), 6-phospho-β-galactosidases (EC 3.2.1.85), 6-phospho-β-glucosidases (EC 3.2.1.86), strictosidine β-glu-
cosidases (EC 3.2.1.105), lactases (EC 3.2.1.108), prunasin β-glucosidases (EC 3.2.1.118), raucaffricine β-glucosidases (EC 3.2.1.125), thioglucosidases (EC 3.2.1.147), β-primeverosidases (EC 3.2.1.149), isoflavonoid 7-O-β-apiosyl β-glucosidases (EC 3.2.1.161); and hydroxyisourate hydrolases (EC 3.–.–.–).²

The kinetics of six of the seventeen enzymes have not been characterized with glycons other than glucose. Specificity on pyranosyl glycons and β-glycosidic bond configurations varies widely among the other eleven. *B. circulans* β-glucosidase prefers glucose as the glycon over galactose, although only when the bond is β-(1,4) rather than β-(1,6). *L. lactis*

---

**Figure 2.** Homology of GH1 structures is displayed with atom representation while looking down the active-site well of the *Zea mays* β-glucosidase main chain. Residues with <50% homology are shown as a ribbon, while homology of ≥50% is shown as spheres. Residues that are 100% conserved are shown in blue, while others are in green.
<table>
<thead>
<tr>
<th>Organism</th>
<th>Glucose</th>
<th>Galactose</th>
<th>Other</th>
<th>Other</th>
<th>% of Cel2 activity</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$k_{cat}$ ($s^{-1}$)</td>
<td>$K_M$ (mM)</td>
<td>$k_{cat}$</td>
<td>$K_M$</td>
<td>$k_{cat}$</td>
</tr>
<tr>
<td>Bacillus circulans</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Brevicoryne brassicae</td>
<td>2.076$^b$</td>
<td>0.359$^b$</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Lactococcus lactis</td>
<td>2$^g$</td>
<td>0.1$^g$</td>
<td>22$^g$</td>
<td>0.09$^g$</td>
<td>-</td>
</tr>
<tr>
<td>Paenibacillus polymyxa</td>
<td>13$^a$</td>
<td>0.9$^a$</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Pyrococcus horikoshii</td>
<td>194</td>
<td>1698</td>
<td>123$^a$</td>
<td>1.3$^a$</td>
<td>2$^h$</td>
</tr>
<tr>
<td>Sinapis alba</td>
<td>-</td>
<td>0.4$^b$</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Sorghum bicolor</td>
<td>13.8$^c$</td>
<td>3.83$^c$</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Streptomyces sp. QM-B81413</td>
<td>-</td>
<td>7.9</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Sulfolobus solfataricus</td>
<td>1333</td>
<td>30</td>
<td>1020$^a$</td>
<td>4.7$^a$</td>
<td>908$^l$</td>
</tr>
<tr>
<td>Thermosphaera aggregans</td>
<td>542$^a$</td>
<td>0.5$^a$</td>
<td>-</td>
<td>-</td>
<td>284$^a$</td>
</tr>
<tr>
<td>Thermotoga maritima</td>
<td>-</td>
<td>-</td>
<td>4MUFe$^c$</td>
<td>4MUGlc$^c$</td>
<td>4MUGalf$^c$</td>
</tr>
<tr>
<td>Thermus nonproteolyticus</td>
<td>91.6$^a$</td>
<td>0.9$^a$</td>
<td>83.0$^a$</td>
<td>10.0$^a$</td>
<td>100.0$^a$</td>
</tr>
<tr>
<td>Thermus thermophilus</td>
<td>23.4$^a$</td>
<td>0.10$^a$</td>
<td>82.9$^a$</td>
<td>5.6$^a$</td>
<td>29.7$^a$</td>
</tr>
<tr>
<td>Trifolium repens</td>
<td>3703$^a$</td>
<td>0.11$^a$</td>
<td>2043$^a$</td>
<td>0.22$^a$</td>
<td>121$^a$</td>
</tr>
<tr>
<td>Triticum aestivum</td>
<td>29.2$^a$</td>
<td>1.70$^a$</td>
<td>2.6$^a$</td>
<td>1.78$^a$</td>
<td>2.0$^a$</td>
</tr>
<tr>
<td>Zea mays</td>
<td>230$^f$</td>
<td>0.272$^f$</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Zea mays p60.1$^{22}$</td>
<td>29.5$^a$</td>
<td>0.64$^a$</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

**Aglycon:** $^a$p-nitrophenyl; $^b$-SC(C2H3)(=NOSO3$^-$); $^c$-nitrophenyl; $^d$p-hydroxy-(S)-mandelonitrile-; $^e$-4-methylumbelliferyl-; $^f$2,4-dihydroxy-7-methoxy-1,4-benzoxazin-3- (DIMBOA)

**Other glycon:** $^g$6-phospho-; $^h$mannose; $^i$xylose; $^j$fucose; $^k$arabinose

**Substrate:** $^l$laminaribiose; $^m$gentibiose; $^n$lactose

Table 1. Known activities on disaccharides and analogues of the GH1 enzymes studied in this chapter. Unless otherwise noted, the aglycon is glucose attached by a β(1,4) bond. A dash is shown for enzyme/substrate pairs for which activity has not been studied.
6-phospho-β-galactosidase is tenfold more active on 6-phospholactose than on 6-phosphocellobiose, but is not active at all on lactose. P. horikoshii alkyl β-glycosidase strongly attacks the β-(1,3) and β-(1,4) bonds in laminaribiose and cellobiose, respectively, as well as p-nitrophenyl β-galactopyranoside but not p-nitrophenyl β-mannopyranoside, although it binds tightly to the latter. Streptomyces sp. β-glucosidase cleaves cellobiose, laminaribiose, and lactose almost equally well. S. sulfataricus β-glycosidase strongly attacks laminaribiose, cellobiose, and gentiobiose, with β-(1,3), β-(1,4), and β-(1,6) bonds, respectively, as well as p-nitrophenyl β-glucopyranoside, β-galactopyranoside, β-fucopyranoside, and β-xylopyranoside. T. aggregans β-glycosidase is more active on a fucosyl glycon than on a glucosyl or a galactosyl one, while T. maritima β-glucosidase attacks substrates with glucose, galactose, and fucose glycons well, but not one with a xylose glycon. T. nonproteolyticus β-glycosidase lives up to its name in being roughly equally active on p-nitrophenyl β-glucopyranoside, β-galactopyranoside, and β-xylopyranoside, and only somewhat less on p-nitrophenyl β-mannopyranoside. T. thermophilis β-glycosidase cleaves substrates with glucosyl, galactosyl, and fucosyl glycons. T. repens cyanogenic β-glucosidase has very high $k_{cat}$ values on substrates with glucosyl and galactosyl glycons, along with lower $k_{cat}$ values but extremely low $K_M$ levels with arabinosyl and xylosyl glycons. T. aestivum β-glucosidase is over ten times as active on substrates whose glycon is glucose or fucose than it is when the glycon is galactose or xylose, and has a much higher $k_{cat}$ (and even lower $K_M$) when the aglycon is DIMBOA rather than p-nitrophenyl.

GH1 enzymes have been classified by multisequence alignment into fourteen subfamilies, and the classification has been updated here. Of the seventeen members with crystal structures, T. aestivum and Z. mays β-glucosidases, S. bicolor and T. repens cyanogenic β-glycosidases, and S. alba myrosinase are part of a plant subfamily (E1) of mainly β-glucosidases and myrosinases. Myrosinase from the insect species B. brassicae is also in the E1 subfamily, but is somewhat close to a bacterial subfamily (B6). B. circulans, P. polymyxa, Strept-
tomycées sp., and T. maritima β-glucosidases, along with T. nonproteolyticus and T. thermodenitrificans β-glucosidases, belong to a bacterial subfamily (B2) of nearly all β-glucosidases. L. lactis 6-P-β-galactosidase is part of a third subfamily (B4) composed almost completely of bacterial 6-P-β-galactosidases and 6-P-β-glucosidases. P. horikoshii, S. solfataricus, and T. aggregans β-glycosidases belong to a pair of closely related subfamilies (A1 and A2) with mainly archaeal but a few bacterial β-galactosidases, β-glucosidases, β-mannosidases, and general β-glycosidases.

Given this difficulty in determining specificities of GH1 members and the general lack of kinetic data associated with those having crystal structures, we decided to use computational methods, specifically automated docking, in an attempt to better understand the catalytic properties of these enzymes. Automated docking of carbohydrates to glycoside hydrolases with AutoDock24 nearly always yields complexes within 1-Å root mean square deviation (RMSD) of crystallized ligand–protein structures, as shown with glucoamylase,25 β-amylase,26 α-1,2-mannosidase,27 GH6 cellobiohydrolase and endoglucanase,28 and GH7 cellobiohydrolase29 and endoglucanase.30 In addition, the Gibbs free energy of binding (ΔGbind) can be obtained by automated docking to within 2 kcal/mol of empirically determined values of ΔGbind.31,32 The free energy function used by AutoDock is shown in eq. 1:

\[
\Delta G_{\text{bind}} = f_{\text{vdW}} \Delta G_{\text{vdW}} + f_{\text{hb}} \Delta G_{\text{hb}} + f_{\text{elec}} \Delta G_{\text{elec}} + f_{\text{solv}} \Delta G_{\text{solv}} + \Delta G_{\text{tors}} \sum_{\text{tors}}
\]

To better understand GH1 member specificities, 28 ligands, including β-(1,2)-, β-(1,3)-, β-(1,4)-, and β-(1,6)-alloyranosyl, -galactopyranosyl, -glucopyranosyl, -mannopyranosyl, -6-phosphogluconopyranosyl, and -6-phosphogalactopyranosyl β-glucopyranosides, as well as β-(1,1-thio)-alloyranosyl, -galactopyranosyl, -glucopyranosyl, and -mannopyranosyl β-glucopyranosides, were docked into the seventeen available GH1 crystal-structure active sites. The resulting ΔGbind values and conclusions as to which residues are responsible for the specificities of these enzymes are presented in this chapter.
**Materials and methods**

*Enzyme Construction*

The tertiary structure of each enzyme was downloaded from the Protein Data Bank (PDB). The PDB IDs are 1BGG, 1CBG, 1E1F, 1E6X, 1GNX, 1GOW, 1HXJ, 1NP2, 1OD0, 1PBG, 1QOX, 1QVB, 1UG6, 1V02, 1VFF, 1WCG, and 2DGA. All chains but the first that contained a glycosidic ligand were deleted from each structure file (if no glycosidic ligand were present, then the first chain was retained). These single-chain ligands were then superimposed using PyMSS, a program to minimize the sum of all pairwise root mean squared deviations (sRMSD). The basis for this minimization was the three most functionally important residues, the catalytic nucleophile, catalytic proton donor/base, and a totally conserved tryptophan in the active site. They were Glu191, Glu406, and Trp452, respectively, in *Zea mays* β-glucosidase, with the corresponding residues in other species being identified by multiple sequence alignment.

After superposition, the crystal-structure ligands were separated into new files. Hydrogen atoms were added to the enzymes using the web interface of What If. Water molecules were removed. Charges were assigned to the enzyme using the procedure of Laederach et al. The C-terminal oxygen charges were then adjusted so that the enzyme carried an integer charge. Solvation parameters were added to the file by the procedure of Morris et al. Docking grid maps were calculated using AutoGrid, part of the AutoDock 3.0.5 package. The grid parameter files (GPF) corresponded to those of Laederach and Reilly.

*Ligand Preparation*

Each ligand was constructed in PCModel 8.0 (Serena Software, Bloomington, IN). The atomic coordinates were optimized using the MMX force field in PCModel. Atomic partial charges were calculated by solving the restricted Hartree–Fock equation in GAMESS.
ligands were placed in enzyme active sites by minimizing the RMSD between the newly created ligands and \( p \)-nitrophenyl 1-thio-\( \beta \)-D-glucopyranoside crystallized in structure 1E1F, a \( Z. \) \textit{mays} \( \beta \)-glucosidase.

Torsions were defined using the procedure of Morris et al.,\textsuperscript{24} allowing those between the two rings as well as those between the rings and their hydroxyl groups to rotate. The glycon and aglycon were maintained in \( ^4C_1 \) conformations during docking.

\textit{Docking}

The Lamarckian genetic algorithm of the AutoDock package performed 1000 iterations with 500 generations per iteration, a population size of 50, and a maximum of 2,000,000 energy evaluations per generation. The docking parameter file (DPF) is in accordance with the procedure of Laederach and Reilly.\textsuperscript{31} The results of the 1000 iterations were then clustered so that no cluster member deviated \( >1.0 \) \( \AA \) from other cluster members. The lowest-energy member of each cluster was considered catalytically viable if the hydrogen atom attached to its glycon C1 was \( <3.0 \) \( \AA \) from the catalytic nucleophile and its glycosidic oxygen atom was \( <3.0 \) \( \AA \) from the catalytic proton donor/base. The six lowest-energy catalytically viable clusters, if available, were selected for further optimization, with the lowest-energy member of each cluster as a cluster representative. These clusters were minimized using 30 successive local minimizations, carried out using the Solis and Wets gradient search of the AutoDock package.\textsuperscript{9} Each local minimization was allowed to iterate 300 times. The ending conformation of the \( (n-1)^{th} \) iteration was used as the starting conformation of the \( n^{th} \) iteration. The iterated local search yields lower energy conformations.\textsuperscript{29}

\textit{ANOVA}

To screen for the residues that cause changes in specificity, three-way ANOVA was performed on the 20 residues within 5 \( \AA \) of the docked ligand. An unweighted free energy (\( E_{\text{bind}} \))
for the remainder of this chapter) calculated using eq. 1 ($f_{vdW}, f_{hb}, f_{solv} = 1.0; f_{elec} = 0.3113; 
\Delta G_{tors} = 0$) was normalized (eq. 2) and then used as the ANOVA response variable:

$$E_{normalized, enzyme, glycon, bond} = \frac{E_{binding, enzyme, glycon, bond}}{E_{binding, enzyme, glucose, 1-4}}$$

(2)

In this analysis, the three independent variables are the glycon type, the glycosidic bond configuration, and the amino acid residue. Since many of the positions within 5 Å of the docked ligand are substituted differently in enzymes from various species, a treatment pool exists with sufficient degrees of freedom to examine all first- and second-order interactions. When ANOVA shows a second-order interaction above the critical threshold between either the glycon or the glycosidic bond configuration and the amino acid residue, significant evidence exists for the amino acid substitution affecting substrate specificity. A least squares mean analysis using Tukey’s method in the JMP 5.1 statistical package was used to find distinct interactions.

To obtain the list of amino acid residues used in each species’ active site, the superposition of structures created earlier was used to hand-align the residues within 5 Å of the docked ligands.

Results and discussion

Values of $E_{bind}$ of the two lowest-energy glycons are summarized in Table 2. The glycon after which the enzyme is named does not always correspond to the glycon that docks with the lowest energy, although galactose and glucose are preferred in most cases and 6-P-galactose is the lowest-energy conformer in 6-P-β-galactosidase. It is also worth noting that most enzymes docked the (1,6)-linked ligand with the lowest energy. The complete tables of docking results are included in Appendix F.
Table 2. Glycons with lowest and second-lowest $-E_{\text{bind}}$ values, along with values of $\mu_E$ (the average docked energy), $\sigma_E$ (the standard deviation of docked ligands), and $-\Delta G_{\text{bind}}$ for the glycon with the lowest value. Energies are in kcal/mol. The ligand column lists the bond configuration and glycon. For all ligands, the aglycon is glucose. An $-S-$ is shown if the glycosidic bond is formed with sulfur rather than oxygen.

<table>
<thead>
<tr>
<th>Enzyme</th>
<th>Ligand</th>
<th>$-E_{\text{bind}}$</th>
<th>$-\mu_E/\sigma_E$</th>
<th>$-\Delta G_{\text{bind}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>B. circulans</em> β-glucosidase</td>
<td>(1,1)-S-Glc, (1,3)-Man</td>
<td>179.5, 168.12</td>
<td>147.0/17.6</td>
<td>8.31</td>
</tr>
<tr>
<td><em>B. brassicae</em> myrosinase</td>
<td>(1,1)-S-Man, (1,1)-S-Gal</td>
<td>244.9, 237.9</td>
<td>196.3/25.0</td>
<td>10.07</td>
</tr>
<tr>
<td><em>L. lactis</em> 6-P-β-galactosidase</td>
<td>(1,1)-S-Glc, 6-P-(1,6)-Gal</td>
<td>173.9, 173.1</td>
<td>147.5/14.2</td>
<td>8.52</td>
</tr>
<tr>
<td><em>P. horikoshii</em> β-glucosidase</td>
<td>(1,4)-Gal, (1,4)-Glc</td>
<td>240.5, 239.9</td>
<td>193.5/27.7</td>
<td>9.81</td>
</tr>
<tr>
<td><em>P. polymyxa</em> β-glucosidase</td>
<td>(1,3)-Gal, (1,1)-S-Glc</td>
<td>194.8, 190.4</td>
<td>158.5/31.8</td>
<td>8.08</td>
</tr>
<tr>
<td><em>S. alba</em> myrosinase</td>
<td>(1,6)-Gal, (1,6)-Glc</td>
<td>193.3, 189.0</td>
<td>158.9/19.4</td>
<td>8.64</td>
</tr>
<tr>
<td><em>S. bicolor</em> cyanogenic β-glucosidase</td>
<td>(1,1)-S-Glc, (1,6)-Gal</td>
<td>197.6, 196.9</td>
<td>160.9/28.2</td>
<td>10.20</td>
</tr>
<tr>
<td><em>Streptomyces sp.</em> β-glucosidase</td>
<td>(1,3)-Gal, (1,1)-S-Glc</td>
<td>191.8, 187.4</td>
<td>164.1/24.1</td>
<td>7.82</td>
</tr>
<tr>
<td><em>S. solfataricus</em> β-glycosidase</td>
<td>(1,3)-Gal, (1,3)-Man</td>
<td>194.8, 189.1</td>
<td>162.9/17.9</td>
<td>8.15</td>
</tr>
<tr>
<td><em>T. aggregans</em> β-glycosidase</td>
<td>(1,6)-Gal, (1,3)-Gal</td>
<td>209.8, 208.3</td>
<td>168.2/33.4</td>
<td>9.30</td>
</tr>
<tr>
<td><em>T. maritima</em> β-glucosidase</td>
<td>(1,6)-Gal, (1,2)-Glc</td>
<td>198.9, 194.3</td>
<td>166.4/32.5</td>
<td>8.84</td>
</tr>
<tr>
<td><em>T. nonproteolyticus</em> β-glyco-</td>
<td>(1,1)-S-Glc, (1,1)-S-Man</td>
<td>190.0, 187.1</td>
<td>160.4/22.6</td>
<td>9.39</td>
</tr>
<tr>
<td></td>
<td>sisidase</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>T. thermophilus</em> β-glycosidase</td>
<td>(1,3)-Glc, (1,3)-Gal</td>
<td>195.2, 194.6</td>
<td>169.4/26.0</td>
<td>8.14</td>
</tr>
<tr>
<td><em>T. repens</em> cyanogenic β-glucosidase</td>
<td>(1,6)-Gal, (1,6)-Glc</td>
<td>205.8, 184.9</td>
<td>158.4/30.9</td>
<td>9.37</td>
</tr>
<tr>
<td><em>T. aestivum</em> β-glucosidase</td>
<td>(1,4)-Gal, (1,4)-Glc</td>
<td>243.9, 241.3</td>
<td>204.2/23.9</td>
<td>9.76</td>
</tr>
<tr>
<td><em>Z. mays</em> β-glucosidase</td>
<td>(1,6)-Gal, (1,6)-Glc</td>
<td>223.0, 218.3</td>
<td>193.5/15.4</td>
<td>10.24</td>
</tr>
<tr>
<td>(ZMGlut1)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
The standard deviation of $E_{\text{bind}}$ for all enzymes, $\sigma_E$, is low compared to the difference between the lowest $E_{\text{bind}}$ and $\mu_E$ obtained with each particular enzyme, confirming that GH1 enzymes bind $\beta$-linked disaccharides in a fairly nonspecific manner. In fact, for each enzyme, nearly all of the ligands docked, and most of the ligands docked in a catalytically viable manner. The exception is hydrolysis of 6-P-$\beta$-glycosides. It was very rare for these compounds to dock in catalytically viable positions, indicating that the majority of GH1 enzymes do not hydrolyze them. The ANOVA results are summarized in Table 3, with full results appearing in Appendix G.

Unfortunately, there were not enough degrees of freedom to probe third-order interactions. Of the 50 amino acids within 5 Å of the docked ligands, 20 have significant interactions with bond configuration. For the most part, this interaction takes the form of one amino acid residue having a standout performance with (1,6)-linked disaccharides, while the remainder of the interactions cluster into mostly indistinguishable Tukey groups.

**Table 3.** Three-way ANOVA of the 20 amino acid residues within 5 Å of the docked ligand with significant second-level interactions. Residue numbering is based on *Z. mays* $\beta$-glucosidase unless the residue is not present in that enzyme, in which case the PDB ID of the residue source is shown. “gly.bond” is the probability of the observed interaction between the glycon and bond configuration happening randomly, and “aa.gly” is the same between the amino acid residue and the glycon. The probability of the interaction between the amino acid residue and the bond occurring randomly is always $<0.001$. Interactions significant at $\alpha = 0.10$ are in bold. (del) denotes an amino acid deletion.

<table>
<thead>
<tr>
<th>Residue</th>
<th>Residue substitutions</th>
<th>gly.bond</th>
<th>aa.gly</th>
</tr>
</thead>
<tbody>
<tr>
<td>Glu191</td>
<td>Gln, Glu</td>
<td>0.937</td>
<td><strong>0.074</strong></td>
</tr>
<tr>
<td>Residue</td>
<td>Residue Interaction</td>
<td>Proportion</td>
<td>T249_0.005</td>
</tr>
<tr>
<td>---------</td>
<td>---------------------</td>
<td>------------</td>
<td>------------</td>
</tr>
<tr>
<td>Gln193</td>
<td>Asn, Gln, Glu, Gly, Trp, Tyr</td>
<td>0.902</td>
<td>0.454</td>
</tr>
<tr>
<td>Thr194</td>
<td>Cys, Gly, Ser, Thr, Val</td>
<td>0.948</td>
<td><strong>0.099</strong></td>
</tr>
<tr>
<td>Phe198</td>
<td>Arg, Asn, Gln, Gly, Leu, Phe, Val</td>
<td>0.872</td>
<td>0.177</td>
</tr>
<tr>
<td>Val204</td>
<td>Glu, Gly, Lys, Ser, Thr, Val, (del)</td>
<td>0.909</td>
<td>0.276</td>
</tr>
<tr>
<td>Phe205</td>
<td>Asp, His, Leu, Phe</td>
<td>0.905</td>
<td><strong>0.091</strong></td>
</tr>
<tr>
<td>Phe260</td>
<td>His, Leu, Met, Phe, Pro, Tyr</td>
<td>0.870</td>
<td>0.118</td>
</tr>
<tr>
<td>Asp261</td>
<td>Ala, Asn, Asp, Ile, Val</td>
<td>0.883</td>
<td>0.105</td>
</tr>
<tr>
<td>Met263</td>
<td>Ala, Arg, His, Gln, Gly, Met, Phe, Ser, Pro</td>
<td>0.872</td>
<td>0.500</td>
</tr>
<tr>
<td>Asn285</td>
<td>Asn, Cys, His, Met, Phe, Ser, (del)</td>
<td>0.726</td>
<td><strong>0.020</strong></td>
</tr>
<tr>
<td>Thr334</td>
<td>Ala, Met, Phe, Ser, Thr</td>
<td>0.912</td>
<td>0.124</td>
</tr>
<tr>
<td>Ser335</td>
<td>Arg, Pro, Ser, Thr, (del)</td>
<td>0.937</td>
<td>0.248</td>
</tr>
<tr>
<td>Arg336</td>
<td>Arg, Asp, Gln, His, Leu, Thr, Tyr, Val, (del)</td>
<td>0.656</td>
<td><strong>0.093</strong></td>
</tr>
<tr>
<td>Gly341 (1GOW)</td>
<td>Gly, Phe, Tyr, (del)</td>
<td>0.943</td>
<td>0.418</td>
</tr>
<tr>
<td>Asn376</td>
<td>Asn, Arg, Glu, Ile, Met, Phe, Ser</td>
<td>0.864</td>
<td><strong>0.041</strong></td>
</tr>
<tr>
<td>Trp378</td>
<td>Ile, Trp</td>
<td>0.937</td>
<td><strong>0.074</strong></td>
</tr>
<tr>
<td>Ile379</td>
<td>Ile, Leu, Phe, (del)</td>
<td>0.953</td>
<td>0.141</td>
</tr>
<tr>
<td>Trp465</td>
<td>Phe, Trp</td>
<td>0.937</td>
<td><strong>0.074</strong></td>
</tr>
<tr>
<td>Phe466</td>
<td>Ala, Asn, Asp, Phe, Ser, (del)</td>
<td>0.863</td>
<td>0.223</td>
</tr>
<tr>
<td>Ala467</td>
<td>Ala, His, Glu, Gln, Lys, Ser, Phe</td>
<td>0.868</td>
<td>0.427</td>
</tr>
</tbody>
</table>

The first listed significant glycon–residue interaction is the catalytic proton donor/base, which should not be a mutation target. *S. alba* myrosinase is the only member of GH1 to have a Gln residue at this position. The remaining seven significant glycon–residue interactions show that some residues produce indistinguishable Tukey groups between glycons, while other substitutions show a glycon preference (Table 4). In particular, mutating Thr194 to Gly194, Phe205 to His205, Asn285 to Met285, and Asn376 to Ser376 in *Z. mays* β-gluco-
sidase could produce a less glycon-specific enzyme. In *S. solfataricus* β-glycosidase, mutating Val209 to Gly209, Phe222 to His222, Thr325 to Tyr325, and Phe359 to Ser359 could do the same.

The residues that significantly interact with glycon specificity are proximal to conformational changes in the bound enantiomers (Figure 3). Thus, it makes sense that focusing site-directed mutagenesis on these residues could produce a less specific enzyme. These residues are different than those mutated earlier by Corbett et al.,\(^4^9\) who found that mutating Met439 to Cys439 increases D-xylosidase specificity while decreasing D-fucosidase activity, and that mutating Glu432 to Cys432 and Trp433 to Cys433 significantly decreased β-galacosidase and β-glucosidase activity.

To create a more specific enzyme, appropriate residues should be mutated to ones shown in the first column (Table 4). For a nonspecific enzyme like *S. solfataricus*, mutating Thr325 to Gln and Phe359 to Glu should produce a more specific enzyme.

**Table 4.** Summary of Tukey’s LSQ analysis on the amino acids that significantly interact with the glycon. Substitutions are categorized by the degree to which the second-order interactions among that substitution are indistinguishable.

<table>
<thead>
<tr>
<th>Amino acid</th>
<th>Most specific</th>
<th>Moderately specific</th>
<th>Least specific</th>
</tr>
</thead>
<tbody>
<tr>
<td>Thr194</td>
<td>Val</td>
<td>Ser, Thr</td>
<td>Cys, Gly</td>
</tr>
<tr>
<td>Phe205</td>
<td>Asp, Phe</td>
<td>Leu</td>
<td>His</td>
</tr>
<tr>
<td>Asn285</td>
<td>Asn, Phe, (del)</td>
<td>Cys</td>
<td>His, Met, Ser</td>
</tr>
<tr>
<td>Arg336</td>
<td>Gln</td>
<td>Leu, Thr</td>
<td>Arg, His, Tyr, Val, (del)</td>
</tr>
<tr>
<td>Asn376</td>
<td>Glu</td>
<td>Asn, Met, Phe</td>
<td>Ile, Ser</td>
</tr>
<tr>
<td>Trp378</td>
<td>Trp</td>
<td>Ile</td>
<td></td>
</tr>
<tr>
<td>Trp465</td>
<td>Trp</td>
<td>Phe</td>
<td></td>
</tr>
</tbody>
</table>
Conclusion

The binding of 28 moieties of cellobiose to GH1 β-glycosidases has been computationally explored. As a result, we now know how specific the constituent enzymes are to those compounds. Using this information, better candidates for protein engineering can be selected. We also have shown that β-(1,6)-linked as well as galactosidic disaccharides are often tighter binders to GH1 family members. We can now state with some statistical certainty which residue substitutions specifically interact with either bond configuration or glycon. Using this information, it is now possible to suggest which residues should be subjected to site-directed mutagenesis to engineer the specificity of highly active and thermostable GH1 members. Focusing future mutagenesis on these residues is likely to produce a highly nonspecific enzyme. Coupling engineering of specificity to advances in GH1 activity and thermostability can create an economically viable route for conversion of cellulose to glucose.

Figure 3. Stereogram of cellobiose (green) docked in the Z. mays ZMGlu1 active site (cyan). Residues within 5 Å of the docked ligands that interact significantly with either the glycon or bond specificity are shown.
References


29. Mulakala, C.; Reilly, P. J. *Proteins* 2005, 60, 598.


47. Vriend, G. J Mol Graph 1990, 8, 52.

Chapter 6: Puckering Coordinates of Monocyclic Rings by Triangular Decomposition

A manuscript submitted to *The Journal of Chemical Information and Modeling*

ANTHONY D. HILL AND PETER J. REILLY*

Department of Chemical and Biological Engineering, 2114 Sweeney Hall, Iowa State University, Ames, Iowa 50011

We describe a new method of describing the pucker of an *N*-member monocyclic ring using *N* − 3 parameters. To accomplish this, three ring atoms define a reference plane, and the remainder of the ring is decomposed into triangular flaps. The angle of incidence for each flap upon the reference plane is then measured. The combination of these angles is characteristic of the ring’s pucker. This puckering coordinate system is compared to existing reduced parameter systems to describe rings using a cyclohexane molecule. We show that this method has the same descriptive power of previous systems while offering advantages in molecular simulations.

___________________________________________________________________________

*Author to whom correspondence should be sent. Phone: +1-515-294-5968; fax: +1-515-294-2689; e-mail: reilly@iastate.edu.*
Introduction

It has been previously shown that $N - 3$ parameters can meaningfully describe $N$-member non-aromatic monocyclic rings.$^{1-4}$ These reduced parameter representations are attractive because their mathematics are not difficult, while their parameters retain much of the information about ring conformation. However, as non-aromatic rings are typically puckered, it is necessary to describe not only the bonding pattern of these rings, but also their pucker.

Beside cyclic alkanes, several biologically active aliphatic monocycles exist with one or more ring positions substituted by either oxygen, nitrogen, or sulfur. Describing the pucker of cyclic aldoses is useful in studies of their reaction mechanisms and molecular interactions.$^{5,6}$ Nearly always carbohydrate–protein associations, as well as the reactions involving the carbohydrate, require a change of ring conformation. Six-member carbohydrate rings may be described as moving through 38 canonical states (Figure 1), previously mapped in spherical coordinates. This spherical projection connects these states as continuous movements in each of the $N - 3$ dimensions. The intuitiveness of this concept has nearly cemented the Cremer–Pople (CP) formalism$^1$ as the de facto descriptor of puckered rings.

Recently, an alternative reduced parameter set proposed by Bercés et al. uses endocyclic dihedral angles to describe ring pucker.$^3$ The parameters allow the direct use of NMR coupling constants to resolve the pucker. Additionally, the endocyclic torsions allow one to describe a particular pucker as the linear combination of three ideal puckers, $^{1}C_4$, $^{1,4}B$, and $^{o}S_2$, using IUPAC nomenclature.$^7,8$

For all the advantages of the CP and Bercés et al. systems, each has drawbacks. With the former, it is relatively difficult to calculate the Cartesian coordinates of all the ring atoms based solely on the reduced parameters. Three coordinates, $\theta$, $\phi$, and $Q$, giving six relative elevations above and below an average plane, exist for a six-member ring. It is possible to use these six elevations and twelve additional system constraints to simultaneously solve
eighteen equations to yield the full eighteen Cartesian coordinates that describe the positions of the six ring atoms. Instead, in general the elevation constraints are used in an \textit{ab initio} simulation to find the Cartesian coordinates of each atom.\textsuperscript{5,6}

With the Bercés et al. system, a straightforward algorithm can calculate all the Cartesian coordinates by using sequential dihedral angles to sequentially place each of the atoms around the ring. However, this system lacks some intuitiveness. Although any three dihedrals can be translated into a linear combination of three ideal puckered states, this is not easily done without a computer. In addition, any three of the six endocyclic dihedrals may be specified. Although useful for calculating a pucker from NMR coupling constants, it is difficult to connect three arbitrary dihedrals to one of the 38 canonical puckering states without resorting to a table or computer.

\textbf{Figure 1.} The five canonical shapes of a puckered pyranosyl ring: chairs (\textit{C}, two states), envelopes (\textit{E}, twelve states), boats (\textit{B}, six states), skew–boats (\textit{S}, six states), and half-chairs (\textit{H}, twelve states). Four atoms are coplanar in four of the conformations, and a fifth is coplanar in envelopes. Superscripts and subscripts denote the atoms not in the plane.
Finally, neither the CP formalism nor the Bercés et al. system helps one to deduce the change induced in a ring’s pucker by external forces. For this reason, we propose a system based upon the decomposition of a monocylic ring into $N - 2$ triangles, with one triangle forming a reference plane against which the angle of elevation is measured for the remaining triangles (Figure 2). Thus, a ring can be completely described as three atoms in a plane, with $N - 3$ flaps that have some angle of orientation to that plane. Formulating the puckered state in this way is not only intuitive, since it is easy to picture the positions of flaps above or below the plane, but it allows both easy computation of the Cartesian coordinates of the atoms as well as a quantitative way to measure the effect of external forces on the ring conformation.

Figure 2. A $^4C_1$ aldopyranosyl ring. The yellow lines are vectors used to compute the angles $\theta_i$. To simplify the illustration, the opposite of $\vec{p}_1$ and $\vec{n}$ are shown. Vectors $\vec{p}_1$, $\vec{q}_0$, and $\vec{n}$ are not to scale so that intersections can be demonstrated, although their directions are correct. The three vectors shown must be computed for each angle $\theta_i$. 


**Calculation of the angles**

The calculation of each of the \(N-3\) puckering angles requires that we first define the axes of puckering, \(\tilde{a}_i\). These are based on the Cartesian coordinates of each ring atom, \(\tilde{x}_j\) (numbering starts at zero to simplify the subscripts):

\[
\tilde{a}_i = \tilde{x}_{2(i+1)} - \tilde{x}_{2i}
\]  

(1)

Using two of the axes, we calculate the vector normal to the reference plane, \(\tilde{n}\), using the cross product:

\[
\tilde{n} = \tilde{a}_1 \otimes \tilde{a}_0
\]  

(2)

We also calculate the vectors representing each of the bonds between atoms, denoting a bond vector as \(\tilde{r}_i\):

\[
\tilde{r}_i = \tilde{x}_{i+1} - \tilde{x}_i
\]  

(3)

Using the bond vectors on either side of an atom, we compute an atom’s orientation vector relative to the plane, \(\tilde{p}_i\). This vector will be orthogonal to both bond vectors, as well as to the axis about which this atom puckers:

\[
\tilde{p}_i = \tilde{r}_{i-1} \otimes \tilde{r}_i
\]  

(4)

To calculate the angle of puckering with the appropriate sign, we create a vector \(\tilde{q}_i\) orthogonal to \(\tilde{p}_{2i+1}\) and \(\tilde{a}_i\):

\[
\tilde{q}_i = \tilde{a}_i \otimes \tilde{p}_{2i+1}
\]  

(5)

The angles of intersection between this vector, \(\tilde{q}_i\), and \(\tilde{n}\) are \(90^\circ - \theta_i\), yielding a positive \(\theta_i\) when the flap is above the plane, and \(90^\circ + \theta_i\), yielding a negative \(\theta_i\) when the flap is below the plane (eq. 6):
\[ \theta_i = 90 - \cos^{-1}\left( \frac{\vec{q}_i \cdot \vec{n}}{||\vec{q}_i|| \cdot ||\vec{n}||} \right) \]  

(6)

These calculations are valid for \( N \leq 6 \) and produce \( N - 3 \) angles of puckering with respect to the reference plane. For \( N > 6 \), the first three axes do not form a triangle (Figure 3). For every each odd/even pair of numbers when \( N > 6 \), i.e. \( \{7, 8\} \), there will be an additional axis cutting the rectangle formed by the first three axes to form two triangles. In these cases, eqs. (1) and (5) must be modified. The remaining equations still produce valid puckering angles about the nearest hinge. This parameterization will be used primarily on monocyclic rings of five and six atoms, so the modified equations are not shown.

**Conversion to Cartesian coordinates**

Any restoration to a full coordinate system from a reduced system requires significant additional information. Eighteen Cartesian coordinates must be specified for six-member rings.

**Figure 3.** A seven–member ring with the four axes shown by dotted lines. The first three axes (computed using the unmodified eq. 5) are shown in heavy dotted lines, and the fourth axis is shown with a light dotted line. As can be seen, the first three axes do not form a triangle, and the fourth axis is needed to bisect the surface created by the first three axes.
From an $N-3$ parameter system, we must specify fifteen additional degrees of freedom to have a fully defined system. One method is to use MM3 or a similar program to solve for the remaining coordinates. However, a reconstruction algorithm can be proposed that should be reasonably accurate and sufficiently fast that it could be included as part of a molecular simulation. Including the puckering coordinates, three positional coordinates, and three orientation coordinates leaves nine more coordinates to fully define the system. We could then use bond lengths and bond angles ($\phi$), although neither can be known with perfect accuracy, since both change from ideal values when a ring is puckered. Thus, we must either use all twelve values in an over-specified system or choose nine for a system that is neither over- nor under-defined. For expediency, we choose six bond lengths and three bond angles, since a low-energy change in bond angle would yield a larger change in the atomic positions than a change in bond length of similar energy. This yields the following algorithm:

1. Place one atom ($\tilde{x}_1$) using three Cartesian coordinates.
2. Use $\phi_0$, $\|\tilde{r}_0\|$, $\|\tilde{r}_1\|$, and an orientation vector to place $\tilde{x}_0$ and $\tilde{x}_2$.
3. Use $\tilde{x}_0$, $\tilde{x}_2$, and $\theta_0$ to calculate the normal vector, $\tilde{n}$.
4. Use $\phi_1$, $\phi_2$, and $\|\tilde{r}_{2-3}\|$ to calculate $\|\tilde{a}_1\|$ and $\|\tilde{a}_2\|$.
5. Use $\tilde{n}$, $\tilde{x}_0$, $\tilde{x}_2$, $\|\tilde{a}_1\|$, and $\|\tilde{a}_2\|$ to place $\tilde{x}_4$.
6. Use $\theta_1$, $\tilde{n}$, $\tilde{x}_2$, $\tilde{x}_4$, $\phi_1$, $\|\tilde{r}_2\|$, and $\|\tilde{r}_3\|$ to place $\tilde{x}_3$.
7. If $N = 6$, repeat step 6 to place $\tilde{x}_5$, using the appropriate variables.
8. Check the nonspecified $\phi$'s for consistency with those specified and adjust the specified $\phi$'s as necessary; then iterate steps 2–7 until the $\phi$'s converge.

**Implications of puckering coordinates**

Expressing the pucker of an aldopyranose or aldofuranose ring as a plane with puckered flaps has a large advantage in molecular simulations. Since one of the planar axes can be seen as a
puckering axis, one can compute the moment of puckering from the atoms pendant to that axis. For molecular dynamics, rather than moving each of the atoms in the ring at each time step, the moment of puckering can be calculated and the puckering flap rotated accordingly.

Forces can be computed for a docked conformation of atoms.\textsuperscript{10,11} This has been used to elucidate the transition-state pathway of a pyranose conformation.\textsuperscript{12} However, this method is very subjective. To understand which direction the ring moves in puckering coordinates requires comparing the force components on the ring and pendant atoms, inspecting a chart of puckering pathways, comparing the binding energy of the nearby docked conformations, and then subjectively deciding which puckered state is next in the transition-state pathway.

Representing a ring with flexible flaps and puckering moments allows one to reduce the subjectivity by calculating the latter about each axis and determining how the conformation will pucker.

A small Python script has been included in Appendix H to demonstrate the straightforward nature of converting between Cartesian coordinates and puckering angles. A further step would be to incorporate this code into a docking suite to allow non-aromatic rings to pucker realistically.

**Compatibility with other reduced coordinate sets**

Table 1 demonstrates equivalent reduced coordinate sets for the 38 canonical puckers in CP parameters, Béces et al. dihedrals, and puckering angles. All three systems completely and uniquely describe all 38 canonical puckers, and each system is compatible with the other two.

This system demonstrates a pleasing symmetry among similar conformations: the $^4C_1$ angles are exactly opposite the $^1C_4$ angles, the $^{1,4}B$ angles are opposite the $B_{1,4}$ angles, the $^6E$ angles are opposite the $E_6$ angles, etc. Also pleasing is the conservation of patterns among the same types of pucker, i.e. the chairs all have angles in the same direction and of the same magnitude, the boats all have two moderately-sized angles of the same sign and one large
Table 1. Puckering coordinates of each reduced coordinate system for the 38 canonical ring puckers. Structures were generated using the coordinates of Bercès et al.\(^3\) (using the 6ring.py program included in Appendix H), and then coordinates were measured using both the CP system and the puckering angles described in this paper, assuming a cyclohexane ring. Position 6 is signified as O in aldopyranosyl rings.

<table>
<thead>
<tr>
<th>Conformer</th>
<th>Bercès et al. dihedrals</th>
<th>Cremer–Pople coordinates</th>
<th>This work</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(\tau_1) ((^\circ))</td>
<td>(\tau_2) ((^\circ))</td>
<td>(\tau_3) ((^\circ))</td>
</tr>
<tr>
<td>1\text{C}_4</td>
<td>60</td>
<td>–60</td>
<td>60</td>
</tr>
<tr>
<td>4\text{C}_1</td>
<td>–60</td>
<td>60</td>
<td>–60</td>
</tr>
<tr>
<td>1,4\text{B}</td>
<td>0</td>
<td>60</td>
<td>–60</td>
</tr>
<tr>
<td>B_{1,4}</td>
<td>0</td>
<td>–60</td>
<td>60</td>
</tr>
<tr>
<td>2,5\text{B}</td>
<td>–60</td>
<td>0</td>
<td>60</td>
</tr>
<tr>
<td>B_{2,5}</td>
<td>60</td>
<td>0</td>
<td>–60</td>
</tr>
<tr>
<td>3,6\text{B}</td>
<td>60</td>
<td>–60</td>
<td>0</td>
</tr>
<tr>
<td>B_{3,6}</td>
<td>–60</td>
<td>60</td>
<td>0</td>
</tr>
<tr>
<td>1\text{H}_2</td>
<td>45</td>
<td>–15</td>
<td>0</td>
</tr>
<tr>
<td>2\text{H}_1</td>
<td>–45</td>
<td>15</td>
<td>0</td>
</tr>
<tr>
<td>2\text{H}_3</td>
<td>–60</td>
<td>45</td>
<td>–15</td>
</tr>
<tr>
<td>3\text{H}_2</td>
<td>60</td>
<td>–45</td>
<td>15</td>
</tr>
<tr>
<td>3\text{H}_4</td>
<td>45</td>
<td>–60</td>
<td>45</td>
</tr>
<tr>
<td>4\text{H}_3</td>
<td>–45</td>
<td>60</td>
<td>–45</td>
</tr>
<tr>
<td>4\text{H}_5</td>
<td>–15</td>
<td>45</td>
<td>–60</td>
</tr>
<tr>
<td>5\text{H}_4</td>
<td>15</td>
<td>–45</td>
<td>60</td>
</tr>
<tr>
<td>5\text{H}_6</td>
<td>0</td>
<td>–15</td>
<td>45</td>
</tr>
<tr>
<td>6\text{H}_5</td>
<td>0</td>
<td>15</td>
<td>–45</td>
</tr>
<tr>
<td>Conformer</td>
<td>Bercés et al. dihedrals</td>
<td>Cremer–Pople coordinates</td>
<td>This work</td>
</tr>
<tr>
<td>-----------</td>
<td>--------------------------</td>
<td>---------------------------</td>
<td>-----------</td>
</tr>
<tr>
<td></td>
<td>$\tau_1$ (°) $\tau_2$ (°) $\tau_3$ (°) $\phi$ (°) $\theta$ (°) $Q$ (Å) $\theta_0$ (°) $\theta_1$ (°) $\theta_2$ (°)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$^6H_1$</td>
<td>-15 0 -15 30 51 0.42 17.83 -9.07 42.16</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$^1H_6$</td>
<td>15 0 15 210 129 0.42 -17.83 9.07 -42.16</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$^1S_3$</td>
<td>-30 60 -30 210 88 0.62 0 50.84 -50.84</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$^3S_1$</td>
<td>30 -60 30 30 92 0.62 0 -50.84 50.84</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$^5S_1$</td>
<td>-30 -30 60 90 92 0.62 50.84 -50.84 0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$^1S_5$</td>
<td>30 30 -60 270 88 0.62 -50.84 50.84 0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$^6S_2$</td>
<td>60 -30 -30 330 88 0.62 -50.84 0 50.84</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$^2S_6$</td>
<td>-60 30 30 150 92 0.62 50.84 0 -50.84</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$^1E$</td>
<td>30 0 0 240 125 0.45 -35.26 17.37 -35.26</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$E_1$</td>
<td>-30 0 0 60 55 0.45 35.26 -17.37 35.26</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$^2E$</td>
<td>-60 30 0 120 55 0.45 46.86 0 0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$E_2$</td>
<td>60 -30 0 300 125 0.45 -46.86 0 0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$^3E$</td>
<td>60 -60 30 360 125 0.45 -35.26 -35.26 17.37</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$E_3$</td>
<td>-60 60 -30 180 55 0.45 35.26 35.26 -17.37</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$^4E$</td>
<td>-30 60 -60 240 55 0.45 0 46.86 0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$E_4$</td>
<td>30 -60 60 60 125 0.45 0 -46.86 0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$^5E$</td>
<td>0 -30 60 120 125 0.45 17.37 -35.26 -35.26</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$E_5$</td>
<td>0 30 -60 300 55 0.45 -17.37 35.26 35.26</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$^6E$</td>
<td>0 0 -30 360 55 0.45 0 0 46.86</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$E_6$</td>
<td>0 0 30 180 125 0.45 0 0 -46.86</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

angle of the opposite sign, and the envelopes all have either two zero angles and one large angle, or two moderately-sized angles and a smaller angle about half their magnitude and
opposite in sign. In addition, one can easily visualize the flaps of the puckered ring elevated above and below the planar atoms.

The Bérces et al. dihedral angles associated with the 38 canonical puckers are all multiples of 15°, as are all the $\phi$ values and those $\theta$ values associated with chair and boat rings in the CP system. This is not so for the angles occurring in the proposed method. However, several values do appear frequently; for instance, $35.26^\circ$ appears in all chair and boat and some envelope conformations. This angle is significant in tetrahedral geometry (Figure 4), so its occurrence is not surprising. Only chair and boat conformations have perfect tetrahedral bond geometries, so this angle does not appear elsewhere except in some envelopes, due to the geometry of their pitch with regard to the reference plane caused by their out-of-plane atoms. Thus the angles observed in the proposed system are reasonable and are merely a reflection of the geometry of the carbon atom’s tetrahedral bonding.

The principle of least motion can be applied to ring structures by stating that two ring conformations can be adjacent in a transition-state pathway if the movement required to transform one to the other is minimal. This principle is observed in the proposed coordinate system, as puckered rings related by least motion are also related by minimal change in puckering coordinates, measured by summing the changes in each of the coordinates. In some transitions, such as $B_{3,6} \rightarrow ^1S_3 \rightarrow ^{1,4}B$, all three coordinates change simultaneously, while in others, like $^1C_4 \rightarrow ^3E \rightarrow ^{3,6}B$, only one coordinate changes. Regardless, the total change in puckering coordinates is smallest between structures participating in feasible transition pathways.
Conclusions

We have shown that an alternate set of reduced coordinates exists for monocyclic non-aromatic rings. We have also highlighted the application of these coordinates to describing aldopyranosyl rings. We have further demonstrated that these coordinates are especially useful in molecular simulations. These coordinates can be used in either descriptive or prescriptive modes.

It would be very useful to integrate these parameters into molecular mechanics or molecular dynamics simulation packages. To do this, it would not only be necessary to modify the source code of the relevant software, but also the force fields of these programs to accommo-

Figure 4. A diagram of tetrahedral geometry. In considering a flap’s pucker, $\phi$ is the bond angle between the exocyclic groups, the three atoms at the bottom are in the reference plane, and $\theta$ is the angle of puckering. Using triangle ABC, $\omega$ is easily computed as $0.5(180^\circ - \phi)$. Using triangle ADE, $\theta + \omega + 90^\circ - 2\omega = 90^\circ$. Consequently, we observe that $\theta = \omega$, and is directly related to the bond angle between exocyclic groups. For a tetrahedron, $\phi = 109.47^\circ$, giving $\theta = 35.26^\circ$. 
date the puckering terms. The energy changes caused by variations in puckering have been studied,\textsuperscript{5,6} but no effort to fit this data to these puckering coordinates has been made. Such work is beyond the scope of this article, but not out of the realm of future research.

Meanwhile, the proposed puckering angle system provides a new and useful method to describe monocyclic ring conformation. In particular, breaking a ring down into groups of flaps that move as the ring puckers allows a quantitative assessment of the effect that intermolecular forces on pendant atoms\textsuperscript{10–12} will have upon the conformation of the ring to which they are attached.

**Acknowledgment**

The authors gratefully thank the National Science Foundation for funding this research, and Dr. Shinya Fushinobu (University of Tokyo), whose research inspired it.

**References**

Chapter 7: Future research directions

The work published in this thesis has enabled several new areas of research. Some of them were mentioned in the chapters. In this chapter, I briefly describe areas of research that I would like to pursue given funding, students, and collaborators.

Mutation prediction

Chapter 5 lists several potential sites for mutagenesis of GH1 enzymes. However, none of these mutation sites have been confirmed by site directed mutagenesis. With appropriate collaborators, it would be possible to test the proposed mutations in a few of the GH1 enzymes.

Confirmation of this prediction technique would allow one to speed enzyme engineering. Using the Gibbs free energy function detailed in Chapter 4, it would be possible to then predict sites for mutagenesis of other glycoside hydrolases. This could then enable a public effort to engineer enzymes for cellulosic ethanol production.

Carbohydrate puckering energy

To allow a carbohydrate ring to pucker in automated docking, it would be best to incorporate one of the reduced coordinate methods mentioned in Chapter 6 into an automated docking suite, like AutoDock.\(^1\) Since the emergy of ring puckering is affected by more than steric and hydrogen bonding interactions, it would be necessary to include a puckering term into the free energy function of Chapter 4. Using the work of Dowd et al.,\(^2,3\) or new work on docking energy using MM4, it would be possible to create a function for the free energy based on the puckering coordinates.
**Glycoside hydrolase transition state prediction**

Since it is much faster to perform automated docking than molecular dynamics, it is very desirable to use automated docking to do simulations normally done with molecular dynamics. One of the big problems that can be solved with molecular dynamics is the prediction of the pre-transition state complex of a carbohydrate and carbohydrate-active enzymes. This has been done with automated docking by Mulakala et al. by docking the 38 different puckers of a carbohydrate into the active site of an enzyme. This complex has also been studied by Fushinobu et al. using molecular dynamics.

Upon incorporation of puckering code into an automated docking suite, like AutoDock, it would be possible to perform some comparisons between molecular dynamics and automated docking. Additionally, it would be possible to compare the binding of natural carbohydrates and inhibitors to confirm that the natural carbohydrates assume a similar pucker. If successful, automated docking could allow researchers to study bound carbohydrate conformations with much less computation time than molecular dynamics.

**RNA folding and ring puckering**

Recently, there has been a desire to study RNA folding and structure to gain insight into the mechanics of catalytic RNA. Since RNA is formed on a phospho-ribose backbone, it is very likely that carbohydrate ring puckering is important to this folding. Using an appropriate energy function, and modeling of ring puckering, it would be interesting to study how ring puckering affects RNA folding and the ability to predict folded RNA structure.

**References**


5. Fushinobu, S; Mertz, J. B.; Petersen, L; Hill, A. D.; Reilly, P. J. *unpublished*
Appendix A: Hybrid Organic-Inorganic Catalyst for Cellobiose Hydrolysis

A grant proposal submitted to the Center for Catalysis, Iowa State University.

Written by Anthony D. Hill, Peter J. Reilly, Brent H. Shanks, and Mark S. Hargrove

Peter J. Reilly, a Brent H. Shanks, a and Mark S. Hargrove, b Co-Principal Investigators

aDepartment of Chemical and Biological Engineering

bDepartment of Biochemistry, Biophysics, and Molecular Biology

Total funding request: $90,000

Abstract

We propose to produce an enzyme mimetic to hydrolyze cellobiose, first to probe hydrolytic enzyme mechanisms but ultimately for commercial purposes. We will make an artificial β-glycosidase by synthesizing two different oligopeptides, each containing a catalytic nucleophile, a cysteine residue, a zipper monomer, and a proton donor, the second oligopeptide in reverse order of its parts from the first; attaching the two oligopeptides to each other through hydrophobic interaction by forming a zipper dimer; and attaching them with either disulfide or peptide bonds to mesoporous silica beads with pendent amino groups under oxidizing conditions.
**Objectives**

We wish to make a functioning artificial enzyme, until now an almost unattainable goal, by:

1) synthesizing the oligopeptide containing the highly homologous region around the catalytic nucleophile of a Family 1 $\beta$-glycosidase, an optional cysteine residue, a zipper monomer, and the highly homologous region around the proton donor of the same enzyme;

2) synthesizing the oligopeptide of the same items, but with the different parts in reverse order;

3) attaching the oligopeptide chains to each other by hydrophobic interaction between their identical zipper monomers, forming a nucleotide-proton donor pair at each end of the chain;

4) crystallizing the paired oligopeptides and determining their structures by X-ray diffraction;

5) synthesizing mesoporous silica beads with pendent amino groups;

6) attaching the dimer to the pore walls of the silica beads by either disulfide or peptide bonds between its cysteine residues and the amino groups of the pores;

7) assaying $\beta$-glycosidase activity with $p$-nitrophenyl $\beta$-D-glucoside to make $\beta$-glucose and $p$-nitrophenol and with cellobiose to make two $\beta$-glucose molecules.

**Procedures**

Enzymes having glycosidic or glycosyltransferase activity have been divided into >90 families based upon catalytic domain amino acid sequence similarity. Family 1 enzymes cleave $\beta$-glycosidic bonds between certain disaccharides, including cellobiose. This family has fifteen members with known crystal structures. All but one member contain two residues for glycosidic bond hydrolysis. One glutamic acid residue acts as a nucleophile, donating its
electrons to the 1-carbon of the nonreducing glucosyl residue. A second glutamic acid residue donates a proton to the glycosidic oxygen atom. The separated residues are then hydrated by a water molecule. These residues are among the most highly conserved ones in any enzyme, as would be expected if their primary responsibility is for catalytic activity and catalytic residue stabilization.

No other residues in the proton chain adjacent to the catalytic nucleophile and proton donor are on the active-site surface, indicating that they do not interact significantly with the disaccharide molecule but that residues in other parts of the enzyme do. Since glutamic acid is a relatively weak acid, it appears that its neighboring residues change the chemical environment so that one residue is basic and the other is acidic. We therefore propose to make a structure containing two nucleophile-proton donor pairs with its adjacent residues on two separate strands at catalytically active distances from each other. Ligating each conserved sequence to a linker to maintain proximity should thus not strongly affect the chemical activity of the glutamic acid residues, making the selected sequences good candidates for enzymatic activity.

Synthesis

Synthesis requires four steps: 1) generating the catalytic oligopeptides, 2) bringing these oligopeptides into proximity, 3) synthesizing mesoporous silica and 4) attaching the oligopeptide dimer to mesoporous silica. Each of the four steps has multiple options.

Catalytic oligopeptides

Various species produce slightly different Family 1 amino acid sequences, all of which hydrolyze β-glycosidic bonds. Listed below are samples of the expressed catalytic oligopeptides, A and C being sequences around the catalytic nucleophile and B and D being sequences around the catalytic proton donor. Each pairing of the members of the A and B
lists and of the members of the C and D lists should be explored for activity after dimer-
ization (see next section). High-throughput protein synthesis based on FMOC chemistry is
available at the ISU Protein Facility.

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>ITLNEPWG</td>
<td>IYITENGR</td>
<td>TLNEPW</td>
<td>YITENG</td>
</tr>
<tr>
<td>LTFNEPQT</td>
<td>IYITENGI</td>
<td>TFNEPQ</td>
<td>YVTENG</td>
</tr>
<tr>
<td>LTFNEPET</td>
<td>MYITENGM</td>
<td>TFNEPE</td>
<td>VITENG</td>
</tr>
<tr>
<td>ITFNEPWC</td>
<td>LYITENGA</td>
<td>TFNEPW</td>
<td>IVTENG</td>
</tr>
<tr>
<td>LTFNEPWC</td>
<td>IYITENGA</td>
<td>TFNEIG</td>
<td></td>
</tr>
<tr>
<td>ATLNEPWC</td>
<td>LYVTENGA</td>
<td>TMNEPN</td>
<td></td>
</tr>
<tr>
<td>TTLNEPWC</td>
<td>LVITENGA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ITLNEPWV</td>
<td>VYITENGA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TTFNEIGP</td>
<td>IYITENGL</td>
<td></td>
<td></td>
</tr>
<tr>
<td>STMNEPNV</td>
<td>MYVTENGI</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>LIVTENGV</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Dimerization linker**

The catalytic oligopeptides can be dimerized by various techniques. The most facile is to use
the heterodimerizing Jun-Fos leucine zipper motif. Additionally, it is possible to use a
homodimerizing leucine zipper system at lower efficiency to bring the two catalytic
oligopeptides together. A third option is to use a commonly available dicarboxylic linker
with FMOC/tert-butyl chemistry.

The Jun-Fos protein system is a heterodimerizing coiled-coil motif that binds to a specific
DNA sequence. The system catalyzes heterodimerization preferentially over
homodimerization. The amino terminus and basic region can be swapped in a modular
fashion, with no effect on the dimerization of Jun and Fos. Fixing of one catalytic
Oligopeptide to the leucine zipper of Fos and the other to the zipper motif of Jun could provide an easy heterodimerization system.

One can also use any other of the proteins of the B-ZIP family, including the CNC/small-MAF, large-MAF/Jun, and large-MAF/Fos heterodimers. In addition, PAR, CREB, and GCN4 would be acceptable homodimer systems that are less efficient than the heterodimer systems.

Fos, Jun, CREB, and PAR proteins in particular bind with greater stability to a basic/basic N-terminal region if an acidic/basic N-terminal region is present. Further, this A-ZIP (acidic N-terminally extended leucine zipper)/B-ZIP (native leucine zipper family member) differentiation provides a heterodimerization system. Therefore, in addition to the naked α-helix motif, a leucine zipper plus an acidic/basic N-terminal region should be studied as a linker.

Therefore we propose to synthesize the following oligopeptides:

<table>
<thead>
<tr>
<th>Part A</th>
<th>Protein</th>
<th>Sequence</th>
<th>Part B</th>
<th>Protein</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Part A</td>
<td>Protein</td>
<td>Sequence</td>
<td>Part B</td>
<td>Protein</td>
<td>Sequence</td>
</tr>
<tr>
<td>Fos-ZIP</td>
<td>ETDQLEE EKAELES</td>
<td>EIAELQK EKERLEF</td>
<td>VLVAHKP</td>
<td>Jun-ZIP</td>
<td>KVKTLKA ENAGLSS</td>
</tr>
<tr>
<td>CNC-ZIP</td>
<td>DVCNLQA KKTTLKR</td>
<td>EQAQCNK AINI MKQ</td>
<td>KLHDLYH</td>
<td>S-MAF-ZIP</td>
<td>QKSELER EVDKLAR</td>
</tr>
<tr>
<td>L-MAF-ZIP</td>
<td>EKTLIQ QVEQLKQ</td>
<td>EVSRLAR ERDAYKV</td>
<td>KCEKL AN</td>
<td>Fos-ZIP</td>
<td>ETDQLEE EKAELES</td>
</tr>
<tr>
<td>L-MAF-ZIP</td>
<td>EKTLIQ QVEQLKQ</td>
<td>EVSRLAR ERDAYKV</td>
<td>Jun-ZIP</td>
<td>KVKTLKA ENAGLSS</td>
<td>TAGLLRE QVAQLKQ</td>
</tr>
</tbody>
</table>
Additionally, a chemical linker may be used in place of a leucine zipper. Two commercially available candidates are 2,7-di-\textit{tert}-butyl-9,9-dimethylxanthene-4,5-dicarboxylic acid (\textit{A}) and mono(\(\alpha\)-ethyl-\(\alpha\)-methylbenzyl)phthalate (\textit{B}):
As can be seen, both molecules offer carboxylic acids for peptide bond formation. \(B\) has the advantage of controlling which carboxylic acid is bound to which oligopeptide. It would be possible to first attach one oligopeptide to the free carboxyl group and then to deprotect the second carboxyl group with \(\text{H}_2/\text{Pd-C}\), as in \(Z\) peptide synthesis chemistry.

**Fixation to silica**

Mesoporous silica will be synthesized with amino groups on the pore walls by co-condensation of (3-aminopropyl)trimethoxysilane (APTMS) with tetraethoxysilane (TEOS).\(^5\) High-surface-area structures with tunable narrow pore size distributions from 2 to 30 nm diameter, the most advantageous size to support the oligopeptide dimers, can be achieved by proper choice of surfactants and synthesis conditions. The amino group concentration within the pores will be controlled by the amount of APTMS introduced during the synthesis. The oligopeptide complex may be fixed to silica either with disulfide linkages through cysteine or with a peptide linkage. In the former case, a cysteine would be added to the oligopeptides to be attached. The disulfide linkage could then be created by subjecting the system to oxidizing conditions using DMSO.\(^6\) In the latter case, solid-phase peptide synthesis techniques could be employed to directly attach an oligopeptide by a peptide bond.

**Characterization**

The synthesized oligopeptide dimers will be crystallized and their structures will be determined by X-ray diffraction. The atomic coordinates of the catalytic regions will show the distances between nucleophiles and proton donors at each chain end. Several leucine
zipper crystal structures have been solved, which should facilitate a molecular replacement solution to the phase problem. However, if molecular replacement fails, we will exploit the Cys residues to facilitate heavy atom isomorphous replacement phasing using mercury. The free Cys side chain is very reactive toward mercury, providing a high probability for isomorphous replacement. By taking advantage of multi-wavelength X-rays at a synchrotron facility, we could solve structures using a single mercury derivative.

Oligopeptide catalytic activity will be measured by \( p \)-nitrophenyl \( \beta \)-D-glucoside hydrolysis to \( \beta \)-glucose and \( p \)-nitrophenol and by cellobiose hydrolysis to two glucose molecules. \( p \)-Nitrophenol production will be followed at 405 nm and glucose production will be measured with a glucose oxidase-peroxidase-\( o \)-dianisidine kit.

Mesoporous silica will be characterized by X-ray diffraction, with its surface area and pore size distribution determined by the BET and BJH methods, respectively. Scanning electron micrography will be used to examine silica particle morphology. The presence of amino groups will be verified by solid-state NMR.

**Caveats**

Beside potential problems with homodimer formation by hydrophobic interaction between oligopeptide chains and by disulfide bond formation between cysteine residues, two very significant issues exist in directly mimicking \( \beta \)-glycosidases. Family 1 \( \beta \)-glycosidases strongly bind only one of the two sugar residues in cellobiose or related disaccharides, less interaction than with most hydrolases, but even this is absent in the enzyme mimetic proposed here. This lack of a binding domain makes the construct more like cation exchange resins that less efficiently hydrolyze disaccharides, although two carboxyl groups in close proximity rather than one will be present. Furthermore, it is uncertain that the two partially unsecured nucleophile-proton donor pairs at each end of the chain will be the correct distances from each other to be efficient catalysts. Two hydrolytic mechanisms occur: In one,
found in Family 1 β-glycosidases, a covalent bond is formed between the nucleophile and the nonreducing C-1, a double displacement reaction ensues, and anomeric configuration is retained. The nucleophile and proton donor are 5.5–6.5 Å apart. In the second, no covalent bond is formed, a single displacement reaction occurs, configuration is inverted, and the residues are 9–10 Å apart. An interesting observation if activity is found will be the configuration of the former nonreducing glucosyl residue.

**Justification**

D-Glucose is a precursor to high-fructose syrup, ethanol, and other chemicals and is itself a food. Currently almost all glucose production is achieved by enzymatic starch hydrolysis.

The use of starch to produce glucose is severely limited by the amount of starch available for use, as the corn from which nearly all starch is derived in this country is also used for animal feed. Another glucose source is cellulose, which is in much larger supply but whose hydrolysis is limited much more than that of starch by its accessibility to enzymes and by the slow rate by which either acid or enzymes hydrolyze its β-(1,4) glycosidic bond.

A very large effort, much of it funded by the U.S. Department of Energy, is underway to remove the barriers preventing the economical conversion of cellulose to glucose. Much of this has dealt with increasing the rate at which cellulases hydrolyze cellulose to glucose.

The discovery of functioning enzyme mimetics is one of the great goals of enzymology. Because of the problems mentioned above, it is unlikely that this first-generation construct will be as active as native Family 1 β-glycosidases, which themselves are not very active. However, beyond the intellectual challenge of making mimetics are other rewards: much higher catalytic concentrations because of the low molecular weight of the linked oligopeptides and much higher stability because of the lower probability of unfolding at higher temperatures.
References


Appendix B: Progress Report on CCAT Project

August 18, 2006

Synthesis of peptides

Two pairs of catalytic amino acids, referred to as A1, B1, C1, and D1 (Table 1), along with Jun and Fos dimerization domains were synthesized by solid-phase organic synthesis (SPOS). These domains were structured so that four peptides were formed: A1-Fos-B1 (Pep1), B1-Jun-A1 (Pep2), C1-Fos-D1 (Pep3), and D1-Jun-C1 (Pep4). The four peptides formed two potential catalytic pairs Pep1/Pep2 (Cat1) and Pep3/Pep4 (Cat2).

The peptides could be formed in either one large SPOS procedure, or in several smaller SPOS procedures. When the several small procedures are used, oligo-peptides are formed that must then be stitched together. The advantage of the large SPOS synthesis is that all the burden of the synthesis rests on the ISU Protein Synthesis facility. The disadvantage is that as the size of the peptide increases, the purity of the final peptide decreases. The Pep1 peptide had a final purity of only 40%. Additionally, for each combination of catalytic domain and dimerization domain, a entirely new peptide would need to be synthesized. This last problem became large, as wait times at the ISU Protein Synthesis Facility became nearly one month long. The disadvantage of the oligo-peptide block synthesis is that aseveral more peptide synthesis steps are required of the researcher.

Pep1, Pep2, Pep3, and Pep4 were are synthesized using the large SPOS procedure. Additionally, the blocks to mix and match A1, B1, C1, D1, Jun, and Fos were all synthesized using the small block SPOS procedure. Purity ranged from 40% to 70% desired peptide among undesired peptides.

At the same time that we wanted to assay for catalytic activity, we also wanted to perform solution NMR to observe the quaternary structure of the peptidic catalysts. In order
to have pure blocks for the secondary steps of the block synthesis procedure and we needed to develop a protein purification method.

**Protein purification**

The proteins were purified using reverse-phase HPLC. A method was developed using a C-12 semi-preparatory column to separate milligrams of the synthesized peptides. The collected fractions were then lyophilized at -80 °C. It was found that a more than half of the original peptide was lost when purified by this procedure.

**Catalytic activity**

Pep₃ and Pep₄ were each dissolved in a dimerization buffer until saturation. The solutions were then diluted by half and solutions containing each monomer, the dimer, and a control were created. Cellobiose was added to each of the solutions in concentrations ranging from 0.1 mg/mL to 5 mg/mL. The catalysis was done in duplicate for 72 hours at room temperature, with periodic sampling. Glucose levels were measured using a kit from Sigma-Aldrich. No catalytic activity of Cat₂ was detected.

**Table 1.** Amino acid sequences of constituent oligopeptides.

<table>
<thead>
<tr>
<th>Peptide</th>
<th>Amino acid sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>A₁</td>
<td>ITLNEPWG</td>
</tr>
<tr>
<td>B₁</td>
<td>IYITENGR</td>
</tr>
<tr>
<td>C₁</td>
<td>TLNEPW</td>
</tr>
<tr>
<td>D₁</td>
<td>YITENG</td>
</tr>
<tr>
<td>Jun</td>
<td>ETDQLEE EKAELES EIAELQK EKERLEF VLVAHKP</td>
</tr>
<tr>
<td>Fos</td>
<td>KVKTTLKA ENAGLSS TAGLLRE QVAQLK Q VMTHVS</td>
</tr>
</tbody>
</table>
Appendix C: User Manual for PyMSS v0.8

Anthony D. Hill 1/18/2006

Quick info

PyMSS is a program to be run on the command line of any POSIX compatible operating system. This includes Microsoft Windows, Macintosh OS X, Linux, and many UNIX variants. This means that the first step in running PyMSS is getting a command line open; in Mac OS X, this means opening the Terminal, and in Windows you need to open a DOS box.

PyMSS reads a parameter file specifying the filenames as well as the list of atoms to be used for the superpositioning of proteins. The format of this file is described in the section entitled, “File Formats”.

PyMSS accepts options on the command line in the manner shown in the section entitled, “Usage”. The first statement is a description of what should be typed on the command line to run PyMSS. Directives in brackets, [ ], indicate optional directives. file.param directs PyMSS to the parameter file described in the section entitled “File Formats”. More detailed descriptions of the directives are contained in the sections entitled “Initial Superposition” and “Genetic Algorithm”. Note: On Microsoft Windows, the command line statement must be preceded by the Python Interpreter’s path. Commonly, this path is C:\Python24\Python.exe, which yields “C:\Python24\Python.exe pymss.py [optional directives] file.param”.

Finally, all of PyMSS’s output will be directed into the working directory. An .out.pdb file will be output for every file listed in the parameters file.

Usage

where

- $i =$ The number of genetic iterations (integer)
- $g =$ The number of generations per iteration (integer)
- $c =$ The number of children per generation (integer)
- $o =$ The number of organisms kept between generations (integer)
- $fm =$ The initial amplitude of genetic mutation (float)
- $fi =$ The magnitude of the initial random deviation from the pseudo-star superposition (float)
- $fd =$ The multiplier applied to the genetic mutation between iterations (float)
- `-debug` = Toggles whether or not the initial superposition is output
- `-noinit` = Toggles whether or not an initial superposition is performed before the genetic algorithm is run.
- `-timing` = Toggles whether or not timing of the program is performed and output.
- `-psinit` = Use the pseudo-star algorithm for the initial superposition instead of neighbor-joining.
- `-njinit (default)` = Use the neighbor-joining algorithm for the initial superposition instead of pseudo-star.
- `file.param` = The file containing information about protein file names and key atoms.

**Introduction**

PyMSS is a program for the multiple superposition of protein structures. To accomplish this task, three algorithms have been built in. Two of the algorithms, called pseudo-star (PSA) and neighbor-joining (NJA), expand on optimal pairwise superpositions. Their algorithms are explained in more detail in the section “Initial Superposition”. Neither algorithm is guaranteed to give the best possible superposition. However, both algorithms are fast, and in use
they produce results that are difficult to beat, even when using the genetic algorithm component of PyMSS. This genetic algorithm (GA) is the third algorithm that PyMSS can use to solve a superposition; its methodology and options are explained in the section entitled “Genetic Algorithm”.

Whichever algorithm is used, PyMSS requires the user to preselect some key atoms that correspond to the same part of each protein to be superimposed. For instance, good candidates would be atoms that bind to a heme group, or atoms involved in binding to a critical metabolite or cofactor. These atoms are listed in a parameter file that PyMSS must read every time it is run. The format of this file is specified in the section entitled, “File Formats”.

As mentioned previously, PyMSS can apply three different algorithms in the superpositioning of proteins. Of the three, the genetic algorithm offers the most theoretical promise, since it can simultaneously optimize the all transformation coordinates of all the proteins. In contrast, the PSA and NJA use heuristics and pairwise transformations to solve the problem three transformation coordinates at a time. In practice, this simultaneous optimization results in a highly dimensional problem space that is not necessarily easily minimized. The genetic algorithm can provide an efficient means of searching that space, yet certain requirements should be met if that efficient search is to be done in a timely fashion. The first requirement is that the problem space not be too large. Since the problem space grows in dimensionality in proportion to the number of structures, the GA should not be used for large numbers of structures (in experience this means >50). The second requirement is that the GA start with a solution somewhat close to the optimal solution. Practically, this is achieved by starting with either PSA or NJA (at the user’s discretion) and then piggy-backing a GA solution on top of this. While it is possible to not use an initial superposition (with the –noinit option), this is not recommended unless some other program has been used to superimpose the structures.
File formats

A parameter file sequentially lists the each protein to be superimposed followed by its key atoms, grouped by residue number. The format is as follows:

protein 1 filename
residue 1 number
space separated list of atom names in residue 1>
residue 2 number
space separated list of atom names in residue 2 >
residue 3 number
space separated list of atom names in residue 3>
...
residue n number
space separated list of atom names in residue n>
protein 2 filename
residue 1 number
space separated list of atom names in residue 1>
...
residue n number
space separated list of atom names in residue n>
...
protein n filename
...space separated list of atom names in residue n>

For example:

1cbg.pdb
33
CA CB OE1 NE2 >
183
CA CB OE1 OE2 >
397
CA CB OE1 OE2 >
1e6x.pdb
39
CA CB OE1 NE2 >
187
CA CB NE2 OE1 >
409
CA CB OE1 OE2 >
1e1f.pdb
38
CA CB OE1 NE2 >
191
CA CB OE1 OE2 >
406
would superimpose proteins found in 1cbg.pdb, 1e6x.pdb, and 1e1f.pdb using 12 atoms, named: [CA, CB, OE1, and NE2 in residue number 33 of 1cbg], [CA, CB, OE1, and OE2 in residue number 183 of 1cbg], and [CA, CB, OE1, and OE2 in residue number 397 of 1cbg]. The residue number is found in columns 23–26, while the atom name is found in columns 13–16 in a Protein Data Bank (PDB) file according to the PDB file format description (http://www.rcsb.org/pdb/file_formats/pdb/pdbguide2.2/ guide2.2_frame.html).

Initial superpositions

Genetic algorithms can search a very large space for an optimal solution. However, in order to do this expediently, the initial solution should be somewhat close to optimal. PyMSS has two methods built in to calculate an initial superposition in a quick manner.

Both methods use a quick method for computing the optimal pair-wise rotation. This method is reported by Simon K. Kearsley in *Acta Crystallographa*, Volume A45, page 208.

Pseudo-Star

In the pseudo-star method, one structure of the group is selected as a reference structure. All the remaining structures are optimally superimposed upon that structure. Since it cannot be guaranteed that any chosen structure will produce the best result, this process is repeated, iteratively selecting each structure to be the reference structure. The superposition with the best sum of all pair-wise Root Mean Squared Deviations (sRMSD) is used.

Neighbor-Joining

The neighbor-joining process is an iteratively solved one, where in each iteration, the most similar proteins (defined as having the lowest residual Root Mean Squared Deviation (RMSD) are paired and superimposed. From each pair-wise superposition, a pair-wise aver-
age structure is computed and replaces the pair in the set of possible structures. After all the possible pairs have been replaced by an average structure, the algorithm moves onto the next iteration with approximately half the original structures to be superimposed. This process is iterated until there is only one remaining structure. The iterations left a history of transformations that can then be applied in series to each original structure to produce a superposition.

*Genetic Algorithm*

The GA optimizes the sRMSD by simultaneously operating on the rigid-body transformation parameters for all structures. These parameters are a pair of angles defining a unit vector in spherical coordinates and a third angle that specifies the rotation about this vector. Each structure has an independent set of these parameters. This simplified transformation is feasible, as the spatial center of each molecule is translated to the origin. Unlike the other algorithms described here, the GA requires computation of the sRMSD with each update of the superposition. The runtime then increases linearly with $M^2$, where $M$ is the total number of key atoms selected [the number of key atoms per structure ($M/N$) times the number of structures]. This is a higher order than with the other algorithms. Although non-deterministic and having no guarantee of optimality, GAs are efficient at optimizing highly dimensional problem spaces.

As a first step, PyMSS reads the parameter file and translates all structures so that the center of the basis atoms lies at the origin. The PSA or NJA is employed to create an approximate superposition. The superposition is then refined if necessary with the GA. A population consisting of ten organisms whose genetic codes are randomly selected is created. Each organism has two chromosomes consisting of $3N$ angles, organized so that the angles $3i$ through $(3i + 2)$ correspond to the $\theta$, $\phi$, and $\psi$ angles of the $i^{th}$ structure in a zero-based array. The proteins are transformed according to Eq. (1):

$$x_{R,i} = R_i \times x_i$$
where \( x_{R,i} \) and \( x_i \) are the rotated and unrotated coordinates of protein \( i \), respectively, and \( R_i \) is the rotation matrix for protein \( i \), defined as

\[
R_i = \begin{bmatrix}
1 - 2(q_{i,0}^2 + q_{i,2}^2) & 2(q_{i,0}q_{i,2} - q_{i,1}q_{i,3}) & 2(q_{i,0}q_{i,2} + q_{i,1}q_{i,3}) \\
2(q_{i,0}q_{i,3} + q_{i,1}q_{i,2}) & 1 - 2(q_{i,0}^2 + q_{i,2}^2) & 2(q_{i,0}q_{i,3} - q_{i,1}q_{i,2}) \\
2(q_{i,0}q_{i,1} - q_{i,3}q_{i,2}) & 2(q_{i,0}q_{i,1} + q_{i,2}q_{i,3}) & 1 - 2(q_{i,0}^2 + q_{i,3}^2)
\end{bmatrix}
\]  

(2)

where \( q_i \) is a rotation quaternion, defined as

\[
q_i = \begin{bmatrix}
 v_{i,0} \sin \left( \frac{v_{i,3}}{2} \right) & v_{i,1} \sin \left( \frac{v_{i,3}}{2} \right) & v_{i,2} \sin \left( \frac{v_{i,3}}{2} \right) & \cos \left( \frac{v_{i,3}}{2} \right)
\end{bmatrix}
\]  

(3)

with \([v_{i,0} \ v_{i,1} \ v_{i,2} \ v_{i,3}]\) \( v_{i,3} \) being a unit vector:

\[
v_i = \begin{bmatrix}
 \cos(\theta_i)\cos(\phi_i) & \sin(\theta_i)\cos(\phi_i) & \sin(\phi_i) & \psi_i
\end{bmatrix}
\]  

(4)

and \( \psi_i \) being the rotation about that unit vector.

The quaternion method of developing a rotation matrix is used because the alternative, progressive multiplication of rotation matrices about the three axes

\[
x_{R,i} = R_{i,\theta} \times R_{i,\phi} \times R_{i,\psi} \times x_i
\]  

(5)

is subject to gimbal lock.

The initial amplitude of the angles is configurable by the user, though best results are found when the amplitude is kept small, with \(-0.1\) and \(0.1\) radians being the default.

The GA runs for some number of user-configurable generations (with a default of 500), with some number of user-configurable organisms in each generation (with a default of 25). The best organisms (defined as those containing angles that produce the lowest sRMSD) are allowed to survive to the next generation (again defined by the user, with a default of 10), and they breed to populate the remainder of the generation. Every few generations the best phenotypes are selected to populate a homozygous population of the same size as that start-
ing a generation. As with all GA parameters, the number of generations between creation of a homozygous population can be set by the user. The mutation rate and amplitude are initially set to values chosen by the user (defaults of 0.5 and $\pi/6$, respectively), but they are allowed to mutate with each generation. When the homozygous population is created, the mutation rate and amplitude are reset, the latter to its initial value multiplied by a geometric scalar, so that the mutation amplitude decreases with each homozygous population. Elitism and crossover are allowed, but chromosome deletion and duplication are not. After the total number of generations, the best phenotype is selected, and the appropriate transforms are made to all the input structures, with the transformed structures outputted into separate files.
Appendix D: Files Needed to Reproduce the Results in Chapter 3

---- Begin mass.config --------

# Configuration file for MASS (Multiple Structural Alignment by Secondary Structures) program
#

SSE (Secondary Structure Elements) Parameters
#
# The format type of the file that assign secondary structure elements (SSEs). Currently, three formats are supported: DSSP, DSSPC, PDB
sse-file-type = DSSP
#
# The endpoints of a secondary structure element are fuzzy. Therefore, the margins of the SSEs can be ignored by setting the sse-margin parameter to a value greater than 0.
sse-margin = 0
#
# Only helices whose size is bigger than this threshold will be considered. Note that the margins are also included in the overall size of a helix
helix-min-size = 5
#
# Only strands whose size is bigger than this threshold will be considered. Note that the margins are also included in the overall size of a strand
strand-min-size = 2
#

Parameters for Defining Similar Bases
#
# Maximal allowed difference in line distance between two bases to be considered as similar. The line distance of a base is computed between his two SSEs.
base-line-distance-tolerance = 1.5
#
# Maximal allowed difference in mid-point distance between two bases to be considered as similar. The mid-point distance of a base is computed between his two SSEs.
base-midPoint-distance-tolerance = 1.5
#
# Maximal allowed angle difference distance between two bases to be considered as similar. The angle of a base is computed between the vectors of his two SSEs. The value is in radians (0.3 ~ 20 degrees)
base-angle-tolerance = 0.3
#
# The maximal RMSD for two bases to be considered as matched
max-rmsd = 6.0
#

------------------------------------------------------------------

#
Multiple Alignment Parameters

Alignments that align less than the value of this parameter will be ignored.

minimum-number-of-aligned-molecules = 2

When set to false, the program runs much faster.

pivot-iteration = true

The program can be run in two alternative modes:
1. sse - using SSE information only. This mode can be used for cases where only SSE information exists.
2. atom - using both SSE and atomic information, where the algorithm fine-tunes the transformations at the atomic level.

alignment-type = atom

max-num-of-best-pairwise-matches = 1

A match is not considered as a poor if it's size is at least this percentage of the smallest molecule in the match. Setting the value of this parameter to 0.0 disables this filter.

min-percentage-of-matched-atoms = 0.02

Clustering Parameters

cluster-bin-size = 3.5
cluster-max-dist = 3.5

Extension Parameters

extension-max-atom-distance = 3.0

The type of match list generated for an alignment.
The available values:
1. hausdorff -
2. one-to-one -
3. sequential - The match list will be computed according to the sequence order of the polypeptide chain.

match-list-type = one-to-one

extension-num-of-refinements = 2

Output Parameters

The number of multiple structural alignments that will be reported for each number of molecules

max-num-of-solutions = 10

Ranking of solutions. The available values are:
1. absolute
2. relative

ranking = absolute
# Log Parameters

# The available log levels are:
# 0 - only error messages
# 1 - error and warning messages
# 2 - error, warning and general information messages
# 3 - error, warning, general information and complexity messages
# 4 - error, warning, general information, complexity and debug messages
debug-level = 2

---- End mass.config ----------

---- Begin superpose.mod ----------
log.verbose()
env = environ()

aln=alignment(env)

# Edit below. Copy the following three lines for
# each structure file, replacing zyme with the
# actual pdb filename.

zzyme=model(env)
zzyme.read(file='zyme.pdb')
aln.append_model(mdl=zzyme, align_codes='zyme', atom_files='zyme.pdb')

aln.malign(local_alignment=False)
aln.malign3d(local_alignment=False, write_fit=True)

---- End superpose.mod ----------

---- Begin zyme.com ----------
output_file=zyme_pairs.out
mode=pairwise
matrix_file=/lockers/bioinformatics/stampdir/blosum62.mat
pairwise_random=100,100,1
gap_penalty=8.0
seq_file=zyme.seqs

---- Emd zyme.com ----------

---- Begin zyme_order.com ----------
zyme_pairs.out
n
n
3
n
---- Emd zyme_order.com ---------

---- Begin zyme_tree.com ---------
output_file=zyme_mult_tree.out
mode=multiple
matrix_file=/lockers/bioinformatics/stampdir/blosum62.mat
gap_penalty=8.0
consplot=mz
print_vertical=
seq_file=zyme.seqs
order_file=zyme.tord
tree_file=zyme.tree

---- Emd zyme_tree.com ---------
--- Begin proteinhxn.gpf ------
receptor proteinhxn.pdbqs #macromolecule
gridfld proteinhxn.AD1.fld #grid_data_file
npts 70 70 70 #num.grid points in xyz
spacing .375 #spacing (Angstroms)
gridcenter 5.607 62.956 63.147 #xyz-coordinates or "auto"
types CNOHX #atom type names
smooth 0.500 #store minimum energy within radius (Angstroms)

map proteinhxn.AD1.C.map #filename of grid map
nbp_coeffs 1272653.000 1127.684 12 6 #C-C_non_bond_parameters/Cn,Cm,n,m
nbp_coeffs 610155.100 783.345 12 6 #C-N_non_bond_parameters/Cn,Cm,n,m
nbp_coeffs 588883.800 633.754 12 6 #C-O_non_bond_parameters/Cn,Cm,n,m
nbp_coeffs 1569268.000 1476.364 12 6 #C-S_non_bond_parameters/Cn,Cm,n,m
nbp_coeffs 88604.240 226.910 12 6 #C-H_non_bond_parameters/Cn,Cm,n,m
nbp_coeffs 88604.240 226.910 12 6 #C-X_non_bond_parameters/Cn,Cm,n,m *
nbp_coeffs 297936.289 631.990 12 6 #C-M_non_bond_parameters/Cn,Cm,n,m *
sol_par 0.00 0.00 #C atomic fragmental volume, solvation param.
constant 0.000 #C grid map constant energy

map proteinhxn.AD1.N.map #filename of grid map
nbp_coeffs 610155.100 783.345 12 6 #N-C_non_bond_parameters/Cn,Cm,n,m
nbp_coeffs 266862.2 546.7653 12 6 #N-N_non_bond_parameters/Cn,Cm,n,m
nbp_coeffs 249961.0 445.9175 12 6 #N-O_non_bond_parameters/Cn,Cm,n,m
nbp_coeffs 721128.6 1036.932 12 6 #N-S_non_bond_parameters/Cn,Cm,n,m
nbp_coeffs 39093.66 155.9833 12 6 #N-H_non_bond_parameters/Cn,Cm,n,m *
nbp_coeffs 39093.66 155.9833 12 6 #N-X_non_bond_parameters/Cn,Cm,n,m *
nbp_coeffs 126818.72 361.362 12 6 #N-M_non_bond_parameters/Cn,Cm,n,m *
sol_par 0.00 0.00 #N atomic fragmental volume, solvation param.
constant 0.000 #N grid map constant energy

map proteinhxn.AD1.O.map #filename of grid map
nbp_coeffs 588883.800 633.754 12 6 #O-C_non_bond_parameters/Cn,Cm,n,m
nbp_coeffs 249961.400 445.918 12 6 #O-N_non_bond_parameters/Cn,Cm,n,m
nbp_coeffs 230584.400 368.677 12 6 #O-O_non_bond_parameters/Cn,Cm,n,m
nbp_coeffs 675844.100 854.687 12 6 #O-S_non_bond_parameters/Cn,Cm,n,m
nbp_coeffs 75570.000 23850.000 12 10 #O-H_non_bond_parameters/Cn,Cm,n,m *
nbp_coeffs 38919.640 124.049 12 6 #O-X_non_bond_parameters/Cn,Cm,n,m *
nbp_coeffs 126818.72 361.362 12 6 #O-M_non_bond_parameters/Cn,Cm,n,m *
sol_par 0.00 0.00 #O atomic fragmental volume, solvation param.
constant 0.000 #O grid map constant energy

map proteinhxn.AD1.H.map #filename of grid map
nbp_coeffs 88604.240 226.910 12 6 #H-C_non_bond_parameters/Cn,Cm,n,m
nbp_coeffs 39093.660 155.983 12 6 #H-N_non_bond_parameters/Cn,Cm,n,m *
nbp_coeffs 75570.000 23850.000 12 10 #H-O_non_bond_parameters/Cn,Cm,n,m *
nbp_coeffs 2657200.000 354290.000 12 10 #H-S_non_bond_parameters/Cn,Cm,n,m

nbp_coeffs 1908.578 46.738 12 6 #H-H_non_bond_parameters/Cn,Cm,n,m
nbp_coeffs 1908.578 46.738 12 6 #H-X_non_bond_parameters/Cn,Cm,n,m *
nbp_coeffs 11537.8093 128.6638 12 6 #H-M_non_bond_parameters/Cn,Cm,n,m *
nbp_coeffs 39093.660 155.983 12 6 #H-n_non_bond_parameters/Cn,Cm,n,m *
sol_par 0.00 0.0000 #H atomic fragmental volume, solvation param.
constant 0.000 #H grid map constant energy

map proteinhx.AD1.X.map
nbp_coeffs 88604.240 226.910 12 6 #X-C_non_bond_parameters/Cn,Cm,n,m *
nbp_coeffs 39093.660 155.983 12 6 #X-N_non_bond_parameters/Cn,Cm,n,m *
nbp_coeffs 38919.640 124.049 12 6 #X-O_non_bond_parameters/Cn,Cm,n,m *
nbp_coeffs 126821.300 290.076 12 6 #X-S_non_bond_parameters/Cn,Cm,n,m *
nbp_coeffs 1908.578 46.738 12 6 #X-H_non_bond_parameters/Cn,Cm,n,m *
nbp_coeffs 1908.578 46.738 12 6 #X-X_non_bond_parameters/Cn,Cm,n,m *
nbp_coeffs 11537.8093 128.6638 12 6 #X-M_non_bond_parameters/Cn,Cm,n,m *
nbp_coeffs 39093.660 155.983 12 6 #X-n_non_bond_parameters/Cn,Cm,n,m *
sol_par 0.00 0.0000 #D atomic fragmental volume, solvation param.
constant 0.000 #D grid map constant energy
elecmap proteinhx.AD1.e.map #electrostatic potential map
dielectric -1.0 #<0,distance-dep.diel; >0,constant
#fmap proteinhx.AD1.f.map #floating grid
--- End proteinhxn.gpf -------

--- Begin ligandmin.dpf -------
seed time pid # for random number generator
types CNOHX # atom type names
fld proteinhx.AD1.fld # grid data file
map proteinhx.AD1.C.map # C-atomic affinity map file
map proteinhx.AD1.N.map # N-atomic affinity map file
map proteinhx.AD1.O.map # O-atomic affinity map file
map proteinhx.AD1.H.map # H-atomic affinity map file
map proteinhx.AD1.X.map # X-atomic affinity map file
map proteinhx.AD1.e.map # electrostatics map file
move ligandhx.t.pdbqs # small molecule file
about 3.544 65.710 64.546 # small molecule center

# Initial Translation, Quaternion and Torsions
tran0 3.544 65.710 64.546 # initial coordinates/A or "random"
quat0 1.0 0.0 0.0 # initial quaternion or "random"
ndihe 4 # number of initial torsions
dihe0 0.0 0.0 0.0 # initial torsions
torsdof 1 0.3113 # num. non-Hydrogen torsional DOF & coeff.

# Initial Translation, Quaternion and Torsion Step Sizes and Reduction Factors
tstep 0.1 # translation step/A
qstep 1.0 # quaternion step/deg
dstep 1.0 # torsion step/deg
tnrf 0.95 # trans reduction factor/per cycle
quarf 0.95 # quat reduction factor/per cycle
dihrf 0.95 # tors reduction factor/per cycle

# Internal Non-Bonded Parameters
intnbp_coeffs 1272653.000 1127.684 12 6 # C-C internal energy non-bond parameters/Cn
intnbp_coeffs 610155.100 783.345 12 6 # C-N non bond_parameters/Cn,Cm,n,m
intnbp_coeffs 588883.800 633.754 12 6 # C-O internal energy non-bond parameters/Cn
intnbp_coeffs 88604.240 226.910 12 6 # C-H internal energy non-bond parameters/Cn
intnbp_coeffs 88604.240 226.910 12 6 # C-X internal energy non-bond parameters/Cn
intnbp_coeffs 266862.2 546.7653 12 6 # N-N non_bond_parameters/Cn,Cm,n,m
intnbp_coeffs 249961.0 445.9175 12 6 # N-O non_bond_parameters/Cn,Cm,n,m
intnbp_coeffs 75570.000 23850.000 12 10 # N-H non_bond_parameters/Cn,Cm,n,m
intnbp_coeffs 39093.66 155.9833 12 6 # N-X non_bond_parameters/Cn,Cm,n,m
intnbp_coeffs 230584.400 368.677 12 6 # O-O internal energy non-bond parameters/Cn
intnbp_coeffs 75570.000 23850.000 12 10 # O-H internal energy non-bond parameters/Cn
intnbp_coeffs 38919.640 124.049 12 6 # O-X internal energy non-bond parameters/Cn
intnbp_coeffs 1908.578 46.738 12 6 # H-H internal energy non-bond parameters/Cn
intnbp_coeffs 1908.578 46.738 12 6 # H-X internal energy non-bond parameters/Cn
intnbp_coeffs 1908.578 46.738 12 6 # X-X internal energy non-bond parameters

#intelec # calculate internal electrostatic energy
#watch sman1.qamb.tors.watch.pdb # real-time monitoring file
outlev 1 # diagnostic output level

# Docked Conformation Clustering Parameters for "analysis" command
rmstol 1.0 # cluster tolerance (Angstroms)
rmsref qligandhx.t.pdbqs # reference structure file for RMS calc.
#rmsnosym # do no symmetry checking in RMS calc.
write_all # write all conformations in a cluster
extrng 1000. # external grid energy
e0max 0.4 # max. allowable initial energy, max. num. retries
sw_max_its 15000 # number of iterations of Solis & Wets local search
sw_max_succ 4 # number of consecutive successes before changing rho
sw_max_fail 4 # number of consecutive failures before changing rho
sw_rho 1.0 # size of local search space to sample
sw_lb_rho 0.01 # lower bound on rho
ls_search_freg 1.0 # probability of performing local search on an indiv.
set_sw1 # set the above pseudo-Solis & Wets parameters
do_Local 100
analysis
--- End ligandmin.dpf -----
--- Begin proteinhxn.allsolv.gpf ------
receptor proteinhxn.pdbqs #macromolecule
gridfld proteinhxn.AD1.allsolv.fld #grid_data_file
npts 70 70 70 #num.grid points in xyz
spacing .375 #spacing (Angstroms)
gridcenter 0.0 0.0 0.0 #xyz-coordinates or "auto"
types CNOHXSP #atom type names
smooth 0.500 #store minimum energy within radius (Angstroms)

map proteinhxn.AD1.allsolv.C.map #filename of grid map
nbp_coeffs 0.00 0.00 12 6 #C-C_non_bond_parameters/Cn,Cm,n,m
nbp_coeffs 0.00 0.00 12 6 #C-N_non_bond_parameters/Cn,Cm,n,m
nbp_coeffs 0.00 0.00 12 6 #C-O_non_bond_parameters/Cn,Cm,n,m
nbp_coeffs 0.00 0.00 12 6 #C-S_non_bond_parameters/Cn,Cm,n,m
nbp_coeffs 0.00 0.00 12 6 #C-H_non_bond_parameters/Cn,Cm,n,m
nbp_coeffs 0.00 0.00 12 6 #C-X_non_bond_parameters/Cn,Cm,n,m *
nbp_coeffs 0.00 0.00 12 6 #C-n_non_bond_parameters/Cn,Cm,n,m *
sol_par 12.77 4.00 #C atomic fragmental volume, solvation param.
constant 0.000 #C grid map constant energy

map proteinhxn.AD1.allsolv.N.map #filename of grid map
nbp_coeffs 0.00 0.00 12 6 #N-C_non_bond_parameters/Cn,Cm,n,m
nbp_coeffs 0.00 0.00 12 6 #N-N_non_bond_parameters/Cn,Cm,n,m
nbp_coeffs 0.00 0.00 12 6 #N-O_non_bond_parameters/Cn,Cm,n,m
nbp_coeffs 0.00 0.00 12 6 #N-S_non_bond_parameters/Cn,Cm,n,m
nbp_coeffs 0.00 0.00 12 6 #N-H_non_bond_parameters/Cn,Cm,n,m *
nbp_coeffs 0.00 0.00 12 6 #N-X_non_bond_parameters/Cn,Cm,n,m *
nbp_coeffs 0.00 0.00 12 6 #N-n_non_bond_parameters/Cn,Cm,n,m *
sol_par 9.00 -17.40 #N atomic fragmental volume, solvation param.
constant 0.000 #N grid map constant energy

map proteinhxn.AD1.allsolv.O.map #filename of grid map
nbp_coeffs 0.00 0.00 12 6 #O-C_non_bond_parameters/Cn,Cm,n,m
nbp_coeffs 0.00 0.00 12 6 #O-N_non_bond_parameters/Cn,Cm,n,m
nbp_coeffs 0.00 0.00 12 6 #O-O_non_bond_parameters/Cn,Cm,n,m
nbp_coeffs 0.00 0.00 12 6 #O-S_non_bond_parameters/Cn,Cm,n,m
nbp_coeffs 0.00 0.00 12 6 #O-H_non_bond_parameters/Cn,Cm,n,m *
nbp_coeffs 0.00 0.00 12 6 #O-X_non_bond_parameters/Cn,Cm,n,m *
nbp_coeffs 0.00 0.00 12 6 #O-n_non_bond_parameters/Cn,Cm,n,m *
sol_par 11.04 -17.40 #O atomic fragmental volume, solvation param.
constant 0.000 #O grid map constant energy

map proteinhxn.AD1.allsolv.H.map #filename of grid map
nbp_coeffs 0.00 0.00 12 6 #H-C_non_bond_parameters/Cn,Cm,n,m
nbp_coeffs 0.00 0.00 12 6 #H-N_non_bond_parameters/Cn,Cm,n,m *
nbp_coeffs 0.00 0.00 12 6 #H-O_non_bond_parameters/Cn,Cm,n,m
nbp_coeffs 0.00 0.00 12 6 #H-S_non_bond_parameters/Cn,Cm,n,m *
nbp_coeffs 0.00 0.00 12 6 #H-H_non_bond_parameters/Cn,Cm,n,m *
nbp_coeffs 0.00 0.00 12 6 #H-X_non_bond_parameters/Cn,Cm,n,m *
sol_par 11.04 -17.40 #H atomic fragmental volume, solvation param.
constant 0.000 #H grid map constant energy
nbp_coeffs 0.00 0.00 12 6 #H-M_non_bond_parameters/Cn,Cm,n,m *
nbp_coeffs 0.00 0.00 12 6 #H-n_non_bond_parameters/Cn,Cm,n,m *
sol_par 0.00 0.0000 #H atomic fragmental volume, solvation param.
constant 0.000 #H grid map constant energy

map proteinhx.AD1.allsolv.X.map
nbp_coeffs 0.00 0.00 12 6 #X-C_non_bond_parameters/Cn,Cm,n,m *
nbp_coeffs 0.00 0.00 12 6 #X-N_non_bond_parameters/Cn,Cm,n,m *
nbp_coeffs 0.00 0.00 12 6 #X-O_non_bond_parameters/Cn,Cm,n,m *
nbp_coeffs 0.00 0.00 12 6 #X-S_non_bond_parameters/Cn,Cm,n,m *
nbp_coeffs 0.00 0.00 12 6 #X-H_non_bond_parameters/Cn,Cm,n,m *
nbp_coeffs 0.00 0.00 12 6 #X-X_non_bond_parameters/Cn,Cm,n,m *
nbp_coeffs 0.00 0.00 12 6 #X-M_non_bond_parameters/Cn,Cm,n,m *
sol_par 0.00 0.0000 #X atomic fragmental volume, solvation param.
constant 0.000 #X grid map constant energy

map proteinhx.AD1.allsolv.S.map
nbp_coeffs 0.00 0.00 12 6 #S-C_non_bond_parameters/Cn,Cm,n,m *
nbp_coeffs 0.00 0.00 12 6 #S-N_non_bond_parameters/Cn,Cm,n,m *
nbp_coeffs 0.00 0.00 12 6 #S-O_non_bond_parameters/Cn,Cm,n,m *
nbp_coeffs 0.00 0.00 12 6 #S-S_non_bond_parameters/Cn,Cm,n,m *
nbp_coeffs 0.00 0.00 12 6 #S-H_non_bond_parameters/Cn,Cm,n,m *
nbp_coeffs 0.00 0.00 12 6 #S-X_non_bond_parameters/Cn,Cm,n,m *
nbp_coeffs 0.00 0.00 12 6 #S-M_non_bond_parameters/Cn,Cm,n,m *
sol_par 16.39 -6.40 #S atomic fragmental volume, solvation param.
constant 0.000 #S grid map constant energy

map proteinhx.AD1.allsolv.P.map
nbp_coeffs 0.00 0.00 12 6 #P-C_non_bond_parameters/Cn,Cm,n,m *
nbp_coeffs 0.00 0.00 12 6 #P-N_non_bond_parameters/Cn,Cm,n,m *
nbp_coeffs 0.00 0.00 12 6 #P-O_non_bond_parameters/Cn,Cm,n,m *
nbp_coeffs 0.00 0.00 12 6 #P-S_non_bond_parameters/Cn,Cm,n,m *
nbp_coeffs 0.00 0.00 12 6 #P-H_non_bond_parameters/Cn,Cm,n,m *
nbp_coeffs 0.00 0.00 12 6 #P-X_non_bond_parameters/Cn,Cm,n,m *
nbp_coeffs 0.00 0.00 12 6 #P-M_non_bond_parameters/Cn,Cm,n,m *
sol_par 6.00 -4.00 #P atomic fragmental volume, solvation param.
constant 0.000 #P grid map constant energy

elecmapt proteinhx.AD1.allsolv.e.map #electrostatic potential map
dielectric -1.0 #<0,distance-dep.diel; >0,constant
#fmap proteinhx.AD1.solv.f.map #floating grid
--- End proteinhxn.allsolv.gpf -------

--- Begin proteinhxn.allsolv.dpf -------

seed time pid # for random number generator
types CNOPSHX # atom type names

fld proteinhx.AD1.allsolv.fld # grid data file
map proteinhx.AD1.allsolv.C.map # C-atomic affinity map file
map proteinhx.AD1.allsolv.N.map # N-atomic affinity map file
map proteinhx.AD1.allsolv.O.map # O-atomic affinity map file
map proteinhx.AD1.allsolv.P.map # P-atomic affinity map file
map proteinhx.AD1.allsolv.S.map # S-atomic affinity map file
map proteinhx.AD1.allsolv.H.map # H-atomic affinity map file
map proteinhx.AD1.allsolv.X.map # X-atomic affinity map file
map proteinhx.AD1.allsolv.e.map # electrostatics map file

--- End proteinhxn.allsolv.dpf ------

--- Begin proteinhxn.amber99.126.gpf ------
receptor proteinhx.pdbqs #macromolecule
gridfld proteinhx.amber99.126.fld #grid_data_file
npts 70 70 70 #num.grid points in xyz
spacing .375 #spacing (Angstroms)
gridcenter 0.0 0.0 0.0 #xyz-coordinates or "auto"
types CNOHXSP #atom type names
smooth 0.500 #store minimum energy within radius (Angstroms)

map proteinhx.amber99.126.C.map
nbp_coeffs 819971.662 531.103 12 6 # C-C
nbp_coeffs 882619.071 653.361 12 6 # C-N
nbp_coeffs 701803.794 614.503 12 6 # C-O
nbp_coeffs 1860689.430 1044.664 12 6 # C-S
nbp_coeffs 2275.776 18.289 12 6 # C-H
nbp_coeffs 86154.188 112.530 12 6 # C-X
nbp_coeffs 436131.076 96.257 12 6 # C-K
sol_par 0.00 0.00
constant 0.000

map proteinhx.amber99.126.N.map
nbp_coeffs 882619.071 653.361 12 6 # N-C
nbp_coeffs 944293.233 801.324 12 6 # N-N
nbp_coeffs 744975.864 750.714 12 6 # N-O
nbp_coeffs 2015621.900 1289.234 12 6 # N-S
nbp_coeffs 436131.076 96.257 12 6 # N-H
nbp_coeffs 89677.699 136.132 12 6 # N-X
nbp_coeffs 490708.532 121.066 12 6 # N-K
sol_par 0.00 0.00
constant 0.000

map proteinhx.amber99.126.O.map
nbp_coeffs 701803.794 614.503 12 6 # O-C
nbp_coeffs 744975.864 750.714 12 6 # O-N
nbp_coeffs 581803.229 699.747 12 6 # O-O
nbp_coeffs 1615879.281 1217.533 12 6 # O-S
nbp_coeffs 0.000 0.000 12 6 # O-H
nbp_coeffs 68278.663 125.288 12 6 # O-X
nbp_coeffs 413008.635 117.149 12 6 # O-K
sol_par 0.00 0.00
constant 0.000

map proteinhx.amber99.126.H.map
nbp_coeffs 2275.776 18.289 12 6 # H-C
nbp_coeffs 2126.012 20.960 12 6 # H-N
nbp_coeffs 0.000 0.000 12 6 # H-O
nbp_coeffs 0.000 0.000 12 6 # H-S
nbp_coeffs 0.140 0.094 12 6 # H-H
nbp_coeffs 107.194 2.595 12 6 # H-X
nbp_coeffs 3245.635 5.428 12 6 # H-K
sol_par 0.00 0.00
constant 0.000

map_proteinhx.amber99.126.X.map
nbp_coeffs 86154.188 112.530 12 6 # X-C
nbp_coeffs 89677.699 136.132 12 6 # X-N
nbp_coeffs 68278.663 125.288 12 6 # X-O
nbp_coeffs 202461.849 225.248 12 6 # X-S
nbp_coeffs 107.194 2.595 12 6 # X-H
nbp_coeffs 7516.077 21.726 12 6 # X-X
nbp_coeffs 58368.854 23.018 12 6 # X-K
sol_par 0.00 0.00
constant 0.000

map_proteinhx.amber99.126.S.map
nbp_coeffs 1860689.430 1044.664 12 6 # S-C
nbp_coeffs 2015621.900 1289.234 12 6 # S-N
nbp_coeffs 1615879.281 1217.533 12 6 # S-O
nbp_coeffs 4194304.000 2048.000 12 6 # S-S
nbp_coeffs 0.000 0.000 12 6 # S-H
nbp_coeffs 202461.849 225.248 12 6 # S-X
nbp_coeffs 944714.838 184.984 12 6 # S-K
sol_par 0.00 0.00
constant 0.000

map_proteinhx.amber99.126.P.map
nbp_coeffs 2253703.493 1087.328 12 6 # P-C
nbp_coeffs 2457465.576 1346.305 12 6 # P-N
nbp_coeffs 1986837.360 1276.821 12 6 # P-O
nbp_coeffs 5045349.812 2124.311 12 6 # P-S
nbp_coeffs 8410.658 43.419 12 6 # P-H
nbp_coeffs 254238.516 238.717 12 6 # P-X
nbp_coeffs 1090297.891 187.944 12 6 # P-K
sol_par 0.00 0.00
constant 0.000
elecmap_proteinhx.amber99.126.e.map #electrostatic potential map
dielectric -1.0 #<0,distance-dep.diel; >0,constant
#fmap_proteinhx.AD1.solv.f.map #floating grid
--- End proteinhx.amber99.126.gpf -------

--- Begin proteinhx.amber99.hb.gpf -------
receptor_proteinhx.pdbqs #macromolecule
gridfld_proteinhx.amber99.hb.fld #grid_data_file
npts 70 70 70 #num.grid points in xyz
spacing .375 #spacing (Angstroms)
gridcenter 0.0 0.0 0.0 #xyz-coordinates or "auto"
types CNOHXSP #atom type names

smooth 0.500 #store minimum energy within radius (Angstroms)

map proteinhx.amber99.hb.C.map
nbp_coeffs 0.000 0.000 12 6 # C-C
nbp_coeffs 0.000 0.000 12 6 # C-N
nbp_coeffs 0.000 0.000 12 6 # C-O
nbp_coeffs 0.000 0.000 12 6 # C-S
nbp_coeffs 0.000 0.000 12 6 # C-H
nbp_coeffs 0.000 0.000 12 6 # C-X
nbp_coeffs 0.000 0.000 12 6 # C-K
sol_par 0.00 0.00
constant 0.000

map proteinhx.amber99.hb.N.map
nbp_coeffs 0.000 0.000 12 6 # N-C
nbp_coeffs 0.000 0.000 12 6 # N-N
nbp_coeffs 0.000 0.000 12 6 # N-O
nbp_coeffs 0.000 0.000 12 6 # N-S
nbp_coeffs 0.000 0.000 12 6 # N-H
nbp_coeffs 0.000 0.000 12 6 # N-X
nbp_coeffs 0.000 0.000 12 6 # N-K
sol_par 0.00 0.00
constant 0.000

map proteinhx.amber99.hb.O.map
nbp_coeffs 0.000 0.000 12 6 # O-C
nbp_coeffs 0.000 0.000 12 6 # O-N
nbp_coeffs 0.000 0.000 12 6 # O-O
nbp_coeffs 0.000 0.000 12 6 # O-S
nbp_coeffs 1404.670 328.524 12 10 # O-H
nbp_coeffs 0.000 0.000 12 6 # O-X
nbp_coeffs 0.000 0.000 12 6 # O-K
sol_par 0.00 0.00
constant 0.000

map proteinhx.amber99.hb.H.map
nbp_coeffs 0.000 0.000 12 6 # H-C
nbp_coeffs 0.000 0.000 12 6 # H-N
nbp_coeffs 1404.670 328.524 12 10 # H-O
nbp_coeffs 5978.607 1114.286 12 10 # H-S
nbp_coeffs 0.000 0.000 12 6 # H-H
nbp_coeffs 0.000 0.000 12 6 # H-X
nbp_coeffs 0.000 0.000 12 6 # H-K
sol_par 0.00 0.00
constant 0.000

map proteinhx.amber99.hb.X.map
nbp_coeffs 0.000 0.000 12 6 # X-C
nbp_coeffs 0.000 0.000 12 6 # X-N
nbp_coeffs 0.000 0.000 12 6 # X-O
nbp_coeffs 0.000 0.000 12 6 # X-S
nbp_coeffs 0.000 0.000 12 6 # X-H
nbp_coeffs 0.000 0.000 12 6 # X-X
nbp_coeffs 0.000 0.000 12 6 # X-K
sol_par 0.00 0.00
constant 0.000

map proteinhx.amber99.hb.S.map
nbp_coeffs 0.000 0.000 12 6 # S-C
nbp_coeffs 0.000 0.000 12 6 # S-N
nbp_coeffs 0.000 0.000 12 6 # S-O
nbp_coeffs 0.000 0.000 12 6 # S-S
nbp_coeffs 5978.607 1114.286 12 10 # S-H
nbp_coeffs 0.000 0.000 12 6 # S-X
nbp_coeffs 0.000 0.000 12 6 # S-K
sol_par 0.00 0.00
constant 0.000

elecmap proteinhx.amber99.hb.e.map #electrostatic potential map
dielectric -1.0 #<0,distance-dep.diel; >0,constant
#fmap proteinhx.AD1.solv.f.map #floating grid

--- End proteinhxn.amber99.hb.gpf -------

--- Begin proteinhxn.amber99.126.dpf -------
seed time pid # for random number generator
types CNOHXSP # atom type names
fld proteinhx.amber99.126.fld #grid_data_file
map proteinhx.amber99.126.C.map
map proteinhx.amber99.126.N.map
map proteinhx.amber99.126.O.map
map proteinhx.amber99.126.H.map
map proteinhx.amber99.126.X.map
map proteinhx.amber99.126.S.map
map proteinhx.amber99.126.P.map
map proteinhx.amber99.126.e.map #electrostatic potential map
#fmap proteinhx.amber99.126.f.map #floating grid

--- End proteinhxn.amber99.126.dpf -------

--- Begin proteinhxn.amber99.hb.dpf -------
seed time pid # for random number generator
types CNOHXSP  # atom type names

fld proteinhx.amber99 hb.fld  #grid_data_file
map proteinhx.amber99 hb C.map
map proteinhx.amber99 hb N.map
map proteinhx.amber99 hb O.map
map proteinhx.amber99 hb H.map
map proteinhx.amber99 hb X.map
map proteinhx.amber99 hb S.map
map proteinhx.amber99 hb P.map
map proteinhx.amber99 hb e map  #electrostatic potential map
#fmap proteinhx AD1 solv f.map  #floating grid

--- End proteinhxn.amber99 hb dpf -------

--- Begin proteinhxn.charmm22.126 gpf -------
receptor proteinhx pdbqs  #macromolecule
gridfld proteinhx.charmm22.126.fld  #grid_data_file
npts 70 70 70  #num.grid points in xyz
spacing .375  #spacing (Angstroms)
gridcenter 0.0 0.0 0.0  #xyz-coordinates or "auto"

types CNOHXSP  # atom type names

smooth 0.500  #store minimum energy within radius (Angstroms)

map proteinhx.charmm22.126 C.map
nbp_coeffs 6298349.364 709.837 12 6  # C-C
nbp_coeffs 6140127.956 1246.332 12 6  # C-N
nbp_coeffs 4233132.940 966.386 12 6  # C-O
nbp_coeffs 14138869.136 2316.317 12 6  # C-S
nbp_coeffs 7135.410 29.262 12 6  # C-H
nbp_coeffs 390996.677 181.126 12 6  # C-X
nbp_coeffs 203954.741 118.871 12 6  # C-Mg
sol_par 0.00 0.00
constant 0.000

map proteinhx.charmm22.126 N.map
nbp_coeffs 6140127.956 1246.332 12 6  # N-C
nbp_coeffs 5266361.605 2052.581 12 6  # N-N
nbp_coeffs 3532990.280 1569.969 12 6  # N-O
nbp_coeffs 12726529.840 3907.920 12 6  # N-S
nbp_coeffs 2410.696 30.246 12 6  # N-H
nbp_coeffs 273210.557 269.241 12 6  # N-X
nbp_coeffs 133821.213 171.227 12 6  # N-Mg
sol_par 0.00 0.00
constant 0.000

map proteinhx.charmm22.126 O.map
nbp_coeffs 4233132.940 966.386 12 6  # O-C
nbp_coeffs 3532990.280 1569.969 12 6  # O-N
nbp_coeffs 2356287.882 1197.316 12 6  # O-O
nbp_coeffs 8626349.658 3004.547 12 6  # O-S
nbp_coeffs 0.000 0.000 12 6 # O-H
nbp_coeffs 175322.515 201.413 12 6 # O-X
nbp_coeffs 84695.444 127.208 12 6 # O-Mg
sol_par 0.00 0.00
constant 0.000

map proteinhx.charmm22.126.H.map
nbp_coeffs 7135.410 29.262 12 6 # H-C
nbp_coeffs 2410.696 30.246 12 6 # H-N
nbp_coeffs 0.000 0.000 12 6 # H-O
nbp_coeffs 0.000 0.000 12 6 # H-S
nbp_coeffs 0.000 0.001 12 6 # H-H
nbp_coeffs 23.192 1.708 12 6 # H-X
nbp_coeffs 6.390 0.815 12 6 # H-Mg
sol_par 0.00 0.00
constant 0.000

map proteinhx.charmm22.126.X.map
nbp_coeffs 390996.677 181.126 12 6 # X-C
nbp_coeffs 273210.557 269.241 12 6 # X-N
nbp_coeffs 175322.515 201.413 12 6 # X-O
nbp_coeffs 713731.054 532.974 12 6 # X-S
nbp_coeffs 23.192 1.708 12 6 # X-H
nbp_coeffs 10086.252 29.792 12 6 # X-X
nbp_coeffs 4436.186 17.954 12 6 # X-Mg
sol_par 0.00 0.00
constant 0.000

map proteinhx.charmm22.126.S.map
nbp_coeffs 14138869.136 2316.317 12 6 # S-C
nbp_coeffs 12726529.840 3907.920 12 6 # S-N
nbp_coeffs 8626349.658 3004.547 12 6 # S-O
nbp_coeffs 3019888.800 7372.800 12 6 # S-S
nbp_coeffs 0.000 0.000 12 6 # S-H
nbp_coeffs 713731.054 532.974 12 6 # S-X
nbp_coeffs 358117.149 343.059 12 6 # S-Mg
sol_par 0.00 0.00
constant 0.000

map proteinhx.charmm22.126.P.map
nbp_coeffs 24384408.352 3248.124 12 6 # P-C
nbp_coeffs 22954763.262 5604.190 12 6 # P-N
nbp_coeffs 15708522.399 4329.312 12 6 # P-O
nbp_coeffs 53557628.912 10484.139 12 6 # P-S
nbp_coeffs 20850.202 116.326 12 6 # P-H
nbp_coeffs 1382938.803 792.184 12 6 # P-X
nbp_coeffs 709305.966 515.536 12 6 # P-Mg
sol_par 0.00 0.00
constant 0.000

elecmap proteinhx.charmm22.126.e.map #electrostatic potential map
dielectric -1.0 #<0,distance-dep.diel; >0,constant
#fmap proteinhx.AD1.solv.f.map #floating grid
--- End proteinhxn.charmm22.126.gpf -------
--- Begin proteinhxn.charmm22.hb.gpf ------
receptor proteinhxn.pdbqs #macromolecule
gridfld proteinhxn.charmm22.hb.fld #grid_data_file
npts 70 70 70 #num.grid points in xyz
spacing .375 #spacing (Angstroms)
gridcenter 0.0 0.0 0.0 #xyz-coordinates or "auto"
types CNOHXSP #atom type names
smooth 0.500 #store minimum energy within radius (Angstroms)

map proteinhxn.charmm22.hb.C.map
nbp_coeffs 0.000 0.000 12 6 # C-C
nbp_coeffs 0.000 0.000 12 6 # C-N
nbp_coeffs 0.000 0.000 12 6 # C-O
nbp_coeffs 0.000 0.000 12 6 # C-S
nbp_coeffs 0.000 0.000 12 6 # C-H
nbp_coeffs 0.000 0.000 12 6 # C-X
nbp_coeffs 0.000 0.000 12 6 # C-Mg
sol_par 0.00 0.00
constant 0.000

map proteinhxn.charmm22.hb.N.map
nbp_coeffs 0.000 0.000 12 6 # N-C
nbp_coeffs 0.000 0.000 12 6 # N-N
nbp_coeffs 0.000 0.000 12 6 # N-O
nbp_coeffs 0.000 0.000 12 6 # N-S
nbp_coeffs 0.000 0.000 12 6 # N-H
nbp_coeffs 0.000 0.000 12 6 # N-X
nbp_coeffs 0.000 0.000 12 6 # N-Mg
sol_par 0.00 0.00
constant 0.000

map proteinhxn.charmm22.hb.O.map
nbp_coeffs 0.000 0.000 12 6 # O-C
nbp_coeffs 0.000 0.000 12 6 # O-N
nbp_coeffs 0.000 0.000 12 6 # O-O
nbp_coeffs 0.000 0.000 12 6 # O-S
nbp_coeffs 1311.413 329.664 12 10 # O-H
nbp_coeffs 0.000 0.000 12 6 # O-X
nbp_coeffs 0.000 0.000 12 6 # O-Mg
sol_par 0.00 0.00
constant 0.000

map proteinhxn.charmm22.hb.H.map
nbp_coeffs 0.000 0.000 12 6 # H-C
nbp_coeffs 0.000 0.000 12 6 # H-N
nbp_coeffs 1311.413 329.664 12 10 # H-O
nbp_coeffs 8357.212 1688.871 12 10 # H-S
nbp_coeffs 0.000 0.000 12 6 # H-H
nbp_coeffs 0.000 0.000 12 6 # H-X
nbp_coeffs 0.000 0.000 12 6 # H-Mg
sol_par 0.00 0.00
constant 0.000

map proteinhx.charmm22.hb.X.map
nbp_coeffs 0.000 0.000 12 6 # X-C
nbp_coeffs 0.000 0.000 12 6 # X-N
nbp_coeffs 0.000 0.000 12 6 # X-O
nbp_coeffs 0.000 0.000 12 6 # X-S
nbp_coeffs 0.000 0.000 12 6 # X-H
nbp_coeffs 0.000 0.000 12 6 # X-X
nbp_coeffs 0.000 0.000 12 6 # X-Mg
sol_par 0.00 0.00
constant 0.000

map proteinhx.charmm22.hb.S.map
nbp_coeffs 0.000 0.000 12 6 # S-C
nbp_coeffs 0.000 0.000 12 6 # S-N
nbp_coeffs 0.000 0.000 12 6 # S-O
nbp_coeffs 0.000 0.000 12 6 # S-S
nbp_coeffs 8357.212 1688.871 12 10 # S-H
nbp_coeffs 0.000 0.000 12 6 # S-X
nbp_coeffs 0.000 0.000 12 6 # S-Mg
sol_par 0.00 0.00
constant 0.000

map proteinhx.charmm22.hb.P.map
nbp_coeffs 0.000 0.000 12 6 # P-C
nbp_coeffs 0.000 0.000 12 6 # P-N
nbp_coeffs 0.000 0.000 12 6 # P-O
nbp_coeffs 0.000 0.000 12 6 # P-S
nbp_coeffs 0.000 0.000 12 6 # P-H
nbp_coeffs 0.000 0.000 12 6 # P-X
nbp_coeffs 0.000 0.000 12 6 # P-Mg
sol_par 0.00 0.00
constant 0.000

elecmap proteinhx.charmm22.hb.e.map #electrostatic potential map
dielectric -1.0 #<0,distance-dep.diel; >0,constant
#fmap proteinhx.AD1.solv.f.map #floating grid

--- End proteinhxn.charmm22.hb.gpf ------

--- Begin proteinhxn.charmm22.126.dpf ------
seed time pid # for random number generator
types CNOHXSP # atom type names
fld proteinhx.charmm22.126 fld #grid_data_file
map proteinhx.charmm22.126.C.map
map proteinhx.charmm22.126.N.map
map proteinhx.charmm22.126.O.map
map proteinhx.charmm22.126.H.map
map proteinhx.charmm22.126.X.map
map proteinhx.charmm22.126.S.map
map proteinhx.charmm22.126.P.map
map proteinhx.charmm22.126.e.map #electrostatic potential map
#fmap proteinhx.AD1.solv.f.map #floating grid

--- End proteinhx.charmm22.dpf ----- 

--- Begin proteinhx.charmm22.hb.dpf ----- 
seed time pid # for random number generator


types CNOHXSP # atom type names

fld proteinhx.charmm22.hb.fld #grid_data_file
map proteinhx.charmm22.hb.C.map
map proteinhx.charmm22.hb.N.map
map proteinhx.charmm22.hb.O.map
map proteinhx.charmm22.hb.H.map
map proteinhx.charmm22.hb.X.map
map proteinhx.charmm22.hb.S.map
map proteinhx.charmm22.hb.P.map
map proteinhx.charmm22.hb.e.map #electrostatic potential map

#fmap proteinhx.AD1.solv.f.map #floating grid

--- End proteinhx.charmm22.hb.dpf ----- 

--- Begin proteinhx.hsolv.gpf ----- 

receptor proteinhx.pdbqs #macromolecule
gridfld proteinhx.AD1.hsolv.fld #grid_data_file
npts 70 70 70 #num.grid points in xyz
spacing .375 #spacing (Angstroms)
gridcenter 0.0 0.0 0.0 #xyz-coordinates or "auto"

types CNOHXSP #atom type names

smooth 0.500 #store minimum energy within radius (Angstroms)

map proteinhx.AD1.hsolv.C.map #filename of grid map
nbp_coeffs 0.00 0.00 12 6 #C-C_non_bond_parameters/Cn,Cm,n,m
nbp_coeffs 0.00 0.00 12 6 #C-N_non_bond_parameters/Cn,Cm,n,m
nbp_coeffs 0.00 0.00 12 6 #C-O_non_bond_parameters/Cn,Cm,n,m
nbp_coeffs 0.00 0.00 12 6 #C-S_non_bond_parameters/Cn,Cm,n,m
nbp_coeffs 0.00 0.00 12 6 #C-H_non_bond_parameters/Cn,Cm,n,m
nbp_coeffs 0.00 0.00 12 6 #C-X_non_bond_parameters/Cn,Cm,n,m *
nbp_coeffs 0.00 0.00 12 6 #C-M_non_bond_parameters/Cn,Cm,n,m *
sol_par 12.77 -4.00 #C atomic fragmental volume, solvation param.
constant 0.000 #C grid map constant energy

map proteinhx.AD1.hsolv.N.map #filename of grid map
nbp_coeffs 0.00 0.00 12 6 #N-C_non_bond_parameters/Cn,Cm,n,m
nbp_coeffs 0.00 0.00 12 6 #N-N_non_bond_parameters/Cn,Cm,n,m
nbp_coeffs 0.00 0.00 12 6 #N-O_non_bond_parameters/Cn,Cm,n,m
nbp_coeffs 0.00 0.00 12 6 #N-S_non_bond_parameters/Cn,Cm,n,m
nbp_coeffs 0.00 0.00 12 6 #N-H_non_bond_parameters/Cn,Cm,n,m *
nbp_coeffs 0.00 0.00 12 6 #N-X_non_bond_parameters/Cn,Cm,n,m *
nbp_coeffs 0.00 0.00 12 6 #N-M_non_bond_parameters/Cn,Cm,n,m *
nbp_coeffs 0.00 0.00 12 6 #N-n_non_bond_parameters/Cn,Cm,n,m *
sol_par 9.00 17.40 #N atomic fragmental volume, solvation param.
constant 0.000 #N grid map constant energy

map proteinhx.AD1.hsolv.O.map #filename of grid map
nbp_coeffs 0.00 0.00 12 6 #O-C_non_bond_parameters/Cn,Cm,n,m
nbp_coeffs 0.00 0.00 12 6 #O-N_non_bond_parameters/Cn,Cm,n,m
nbp_coeffs 0.00 0.00 12 6 #O-O_non_bond_parameters/Cn,Cm,n,m
nbp_coeffs 0.00 0.00 12 6 #O-S_non_bond_parameters/Cn,Cm,n,m
nbp_coeffs 0.00 0.00 12 6 #O-H_non_bond_parameters/Cn,Cm,n,m *
nbp_coeffs 0.00 0.00 12 6 #O-X_non_bond_parameters/Cn,Cm,n,m *
nbp_coeffs 0.00 0.00 12 6 #O-M_non_bond_parameters/Cn,Cm,n,m *
nbp_coeffs 0.00 0.00 12 6 #O-n_non_bond_parameters/Cn,Cm,n,m *
sol_par -9.00 17.40 #O atomic fragmental volume, solvation param.
constant 0.000 #O grid map constant energy

map proteinhx.AD1.hsolv.H.map #filename of grid map
nbp_coeffs 0.00 0.00 12 6 #H-C_non_bond_parameters/Cn,Cm,n,m
nbp_coeffs 0.00 0.00 12 6 #H-N_non_bond_parameters/Cn,Cm,n,m *
nbp_coeffs 0.00 0.00 12 6 #H-O_non_bond_parameters/Cn,Cm,n,m
nbp_coeffs 0.00 0.00 12 6 #H-S_non_bond_parameters/Cn,Cm,n,m *
nbp_coeffs 0.00 0.00 12 6 #H-X_non_bond_parameters/Cn,Cm,n,m *
nbp_coeffs 0.00 0.00 12 6 #H-M_non_bond_parameters/Cn,Cm,n,m *
nbp_coeffs 0.00 0.00 12 6 #H-n_non_bond_parameters/Cn,Cm,n,m *
sol_par -4.00 3.0000 #H atomic fragmental volume, solvation param.
constant 0.000 #H grid map constant energy

map proteinhx.AD1.hsolv.X.map
nbp_coeffs 0.00 0.00 12 6 #X-C_non_bond_parameters/Cn,Cm,n,m *
nbp_coeffs 0.00 0.00 12 6 #X-N_non_bond_parameters/Cn,Cm,n,m *
nbp_coeffs 0.00 0.00 12 6 #X-O_non_bond_parameters/Cn,Cm,n,m *
nbp_coeffs 0.00 0.00 12 6 #X-S_non_bond_parameters/Cn,Cm,n,m *
nbp_coeffs 0.00 0.00 12 6 #X-X_non_bond_parameters/Cn,Cm,n,m *
nbp_coeffs 0.00 0.00 12 6 #X-M_non_bond_parameters/Cn,Cm,n,m *
nbp_coeffs 0.00 0.00 12 6 #X-n_non_bond_parameters/Cn,Cm,n,m *
sol_par 4.00 -3.0000 #X atomic fragmental volume, solvation param.
constant 0.000 #X grid map constant energy

map proteinhx.AD1.hsolv.S.map
nbp_coeffs 0.00 0.00 12 6 #S-C_non_bond_parameters/Cn,Cm,n,m *
nbp_coeffs 0.00 0.00 12 6 #S-N_non_bond_parameters/Cn,Cm,n,m *
nbp_coeffs 0.00 0.00 12 6 #S-O_non_bond_parameters/Cn,Cm,n,m *
nbp_coeffs 0.00 0.00 12 6 #S-S_non_bond_parameters/Cn,Cm,n,m *
nbp_coeffs 0.00 0.00 12 6 #S-X_non_bond_parameters/Cn,Cm,n,m *
nbp_coeffs 0.00 0.00 12 6 #S-M_non_bond_parameters/Cn,Cm,n,m *
nbp_coeffs 0.00 0.00 12 6 #S-n_non_bond_parameters/Cn,Cm,n,m *
sol_par -16.39 6.40 #S atomic fragmental volume, solvation param.
constant 0.000 #S grid map constant energy

map proteinhx.AD1.hsolv.P.map
nbp_coeffs 0.00 0.00 12 6 #P-C_non_bond_parameters/Cn,Cm,n,m *
nbp_coeffs 0.00 0.00 12 6  #P-N_non_bond_parameters/Cn,Cm,n,m *
nbp_coeffs 0.00 0.00 12 6  #P-O_non_bond_parameters/Cn,Cm,n,m *
nbp_coeffs 0.00 0.00 12 6  #P-S_non_bond_parameters/Cn,Cm,n,m *
nbp_coeffs 0.00 0.00 12 6  #P-H_non_bond_parameters/Cn,Cm,n,m *
nbp_coeffs 0.00 0.00 12 6  #P-X_non_bond_parameters/Cn,Cm,n,m *
nbp_coeffs 0.00 0.00 12 6  #P-M_non_bond_parameters/Cn,Cm,n,m *
sol_par -19.93 20.00  #D atomic fragmental volume, solvation param.
constant 0.000  #D grid map constant energy

elecmap proteinhx.AD1.hsolv.e.map  #electrostatic potential map
dielectric -1.0  #<0,distance-dep.diel; >0,constant
#fmap proteinhx.AD1.solv.f.map  #floating grid

--- End proteinhxn.hsolv.gpf -----

--- Begin proteinhxn.hsolv.dpf ------
seed time pid  # for random number generator
types CNOPSHX  # atom type names
fld proteinhx.AD1.hsolv.fld  # grid data file
map proteinhx.AD1.hsolv.C.map  #C-atomic affinity map file
map proteinhx.AD1.hsolv.N.map  #N-atomic affinity map file
map proteinhx.AD1.hsolv.O.map  #O-atomic affinity map file
map proteinhx.AD1.hsolv.P.map  #P-atomic affinity map file
map proteinhx.AD1.hsolv.S.map  #S-atomic affinity map file
map proteinhx.AD1.hsolv.H.map  #H-atomic affinity map file
map proteinhx.AD1.hsolv.X.map  #X-atomic affinity map file
map proteinhx.AD1.hsolv.e.map  #electrostatics map file

--- End proteinhxn.hsolv.dpf ------

--- Begin proteinhxn.AD1.126.gpf ------
receptor proteinhx.pdbqs  #macromolecule
gridfld proteinhx.AD1.126.fld  #grid_data_file
npts 70 70 70  #num.grid points in xyz
spacing .375  #spacing (Angstroms)
gridcenter 0.0 0.0 0.0  #xyz-coordinates or "auto"
types CNOHXSP  #atom type names
smooth 0.500  #store minimum energy within radius (Angstroms)
map proteinhx.AD1.126.C.map  #filename of grid map
nbp_coeffs 1272653.000 1127.684 12 6  #C-C_non_bond_parameters/Cn,Cm,n,m
nbp_coeffs 610155.100 783.345 12 6  #C-N_non_bond_parameters/Cn,Cm,n,m
nbp_coeffs 588883.800 633.754 12 6  #C-O_non_bond_parameters/Cn,Cm,n,m
nbp_coeffs 1569268.000 1476.364 12 6  #C-S_non_bond_parameters/Cn,Cm,n,m
nbp_coeffs 88604.240 226.910 12 6  #C-H_non_bond_parameters/Cn,Cm,n,m
nbp_coeffs 88604.240 226.910 12 6  #C-X_non_bond_parameters/Cn,Cm,n,m *
nbp_coeffs 297936.289 631.99 12 6  #C-M_non_bond_parameters/Cn,Cm,n,m *
sol_par 0.00 0.00  #C atomic fragmental volume, solvation param.
constant 0.000  #C grid map constant energy
map proteinhx.AD1.126.N.map #filename of grid map
nbp_coeffs 610155.100 783.345 12 6 #N-C_non_bond_parameters/Cn,Cm,n,m
nbp_coeffs 266862.2 546.7653 12 6 #N-N_non_bond_parameters/Cn,Cm,n,m
nbp_coeffs 249961.0 445.9175 12 6 #N-O_non_bond_parameters/Cn,Cm,n,m
nbp_coeffs 721128.6 1036.932 12 6 #N-S_non_bond_parameters/Cn,Cm,n,m
nbp_coeffs 39093.66 155.9833 12 6 #N-H_non_bond_parameters/Cn,Cm,n,m
nbp_coeffs 39093.66 155.9833 12 6 #N-X_non_bond_parameters/Cn,Cm,n,m
nbp_coeffs 126818.72 361.362 12 6 #N-M_non_bond_parameters/Cn,Cm,n,m
nbp_coeffs 266862.2 546.7653 12 6 #N-n_non_bond_parameters/Cn,Cm,n,m
sol_par 0.00 0.00 #O atomic fragmental volume, solvation param.
constant 0.000 #O grid map constant energy

map proteinhx.AD1.126.O.map #filename of grid map
nbp_coeffs 588883.80 633.754 12 6 #O-C_non_bond_parameters/Cn,Cm,n,m
nbp_coeffs 249961.400 445.918 12 6 #O-N_non_bond_parameters/Cn,Cm,n,m
nbp_coeffs 230584.400 368.677 12 6 #O-O_non_bond_parameters/Cn,Cm,n,m
nbp_coeffs 675844.100 854.687 12 6 #O-S_non_bond_parameters/Cn,Cm,n,m
nbp_coeffs 0.000 0.000 12 10 #O-H_non_bond_parameters/Cn,Cm,n,m
nbp_coeffs 38919.640 124.049 12 6 #O-X_non_bond_parameters/Cn,Cm,n,m
nbp_coeffs 126818.72 361.362 12 6 #O-M_non_bond_parameters/Cn,Cm,n,m
nbp_coeffs 249961.400 445.918 12 6 #O-n_non_bond_parameters/Cn,Cm,n,m
sol_par 0.00 0.00 #O atomic fragmental volume, solvation param.
constant 0.000 #O grid map constant energy

map proteinhx.AD1.126.H.map #filename of grid map
nbp_coeffs 88604.240 226.910 12 6 #H-C_non_bond_parameters/Cn,Cm,n,m
nbp_coeffs 39093.660 155.983 12 6 #H-N_non_bond_parameters/Cn,Cm,n,m
nbp_coeffs 0.000 0.000 12 10 #H-O_non_bond_parameters/Cn,Cm,n,m
nbp_coeffs 1908.578 46.738 12 6 #H-H_non_bond_parameters/Cn,Cm,n,m
nbp_coeffs 1908.578 46.738 12 6 #H-X_non_bond_parameters/Cn,Cm,n,m
nbp_coeffs 11537.8093 128.6638 12 6 #H-M_non_bond_parameters/Cn,Cm,n,m
nbp_coeffs 39093.660 155.983 12 6 #H-n_non_bond_parameters/Cn,Cm,n,m
sol_par 0.00 0.0000 #H atomic fragmental volume, solvation param.
constant 0.000 #H grid map constant energy

map proteinhx.AD1.126.X.map
nbp_coeffs 88604.240 226.910 12 6 #X-C_non_bond_parameters/Cn,Cm,n,m
nbp_coeffs 39093.660 155.983 12 6 #X-N_non_bond_parameters/Cn,Cm,n,m
nbp_coeffs 38919.640 124.049 12 6 #X-O_non_bond_parameters/Cn,Cm,n,m
nbp_coeffs 126821.300 290.076 12 6 #X-S_non_bond_parameters/Cn,Cm,n,m
nbp_coeffs 1908.578 46.738 12 6 #X-H_non_bond_parameters/Cn,Cm,n,m
nbp_coeffs 1908.578 46.738 12 6 #X-X_non_bond_parameters/Cn,Cm,n,m
nbp_coeffs 11537.8093 128.6638 12 6 #X-M_non_bond_parameters/Cn,Cm,n,m
nbp_coeffs 39093.660 155.983 12 6 #X-n_non_bond_parameters/Cn,Cm,n,m
sol_par 0.00 0.0000 #D atomic fragmental volume, solvation param.
constant 0.000 #D grid map constant energy

map proteinhx.AD1.126.S.map
nbp_coeffs 1569268.000 1476.364 12 6 #S-C_non_bond_parameters/Cn,Cm,n,m
nbp_coeffs 721128.6 1036.932 12 6 #S-N_non_bond_parameters/Cn,Cm,n,m
nbp_coeffs 675844.100 854.687 12 6 #S-O_non_bond_parameters/Cn,Cm,n,m
nbp_coeffs 3355443.200 1638.400000 12 6 #S-S_non-bond-parameters/Cn,Cm,n,m
nbp_coeffs 0.000 0.000 12 10 #S-H non_bond_parameters/Cn,Cm,n,m *
nbp_coeffs 126821.300 290.076 12 6 #S-X non_bond_parameters/Cn,Cm,n,m *
nbp_coeffs 297936.289 631.998 12 6  # S-
M_non_bond_parameters/Cn,Cm,n,m *
nbp_coeffs 610155.100 783.345 12 6  # S-
n_non_bond_parameters/Cn,Cm,n,m *
sol_par 0.00 0.0000 #D atomic fragmental volume, solvation param.
constant 0.000 #D grid map constant energy

map proteinhx.AD1.126.P.map
nbp_coeffs 3903483.822 1643.536 12 6  #P-
C_non_bond_parameters/Cn,Cm,n,m
nbp_coeffs 1898374.035 1165.863 12 6  #P-
N_non_bond_parameters/Cn,Cm,n,m
nbp_coeffs 1316590.401 1026.291 12 6  #P-
O_non_bond_parameters/Cn,Cm,n,m
nbp_coeffs 4512698.060 1900.042 12 6  #P-
S_non_bond_parameters/Cn,Cm,n,m
nbp_coeffs 49622.755 111.825 12 6  #P-
H_non_bond_parameters/Cn,Cm,n,m
nbp_coeffs 49622.755 111.825 12 6  #P-
X_non_bond_parameters/Cn,Cm,n,m *
nbp_coeffs 297936.289 631.998 12 6  #P-
M_non_bond_parameters/Cn,Cm,n,m *
nbp_coeffs 610155.100 783.345 12 6  #P-
n_non_bond_parameters/Cn,Cm,n,m *
sol_par 0.00 0.0000 #D atomic fragmental volume, solvation param.
constant 0.000 #D grid map constant energy
elecmap proteinhx.AD1.126.e.map #electrostatic potential map
dielectric -1.0 #<0,distance-dep.diel; >0,constant
#fmap proteinhx.AD1.126.f.map #floating grid
--- End proteinhxn.AD1.126.gpf ------

--- Begin proteinhxn.AD1.hb.gpf ------
receptor proteinhx.pdbqs #macromolecule
gridfld proteinhx.AD1.hb.fld #grid_data_file
npts 70 70 70 #num.grid points in xyz
spacing .375 #spacing (Angstroms)
gridcenter 0.0 0.0 0.0 #xyz-coordinates or "auto"
types CNOHXSP #atom type names
smooth 0.500 #store minimum energy within radius (Angstroms)

map proteinhxn.AD1.hb.C.map #filename of grid map
nbp_coeffs 0.00 0.00 12 6 #C-C_non_bond_parameters/Cn,Cm,n,m
nbp_coeffs 0.00 0.00 12 6 #C-N_non_bond_parameters/Cn,Cm,n,m
nbp_coeffs 0.00 0.00 12 6 #C-O_non_bond_parameters/Cn,Cm,n,m
nbp_coeffs 0.00 0.00 12 6 #C-S_non_bond_parameters/Cn,Cm,n,m
nbp_coeffs 0.00 0.00 12 6 #C-H_non_bond_parameters/Cn,Cm,n,m
nbp_coeffs 0.00 0.00 12 6 #C-X_non_bond_parameters/Cn,Cm,n,m *
nbp_coeffs 0.00 0.00 12 6 #C-M_non_bond_parameters/Cn,Cm,n,m *
nbp_coeffs 0.00 0.00 12 6 #C-n_non_bond_parameters/Cn,Cm,n,m *
sol_par 0.00 0.00 #C atomic fragmental volume, solvation param.
constant 0.000 #C grid map constant energy

map proteinhx.AD1 hb.N.map #filename of grid map
nbp_coeffs 0.00 0.00 12 6 #N-C_non_bond_parameters/Cn,Cm,n,m
nbp_coeffs 0.00 0.00 12 6 #N-N_non_bond_parameters/Cn,Cm,n,m
nbp_coeffs 0.00 0.00 12 6 #N-O_non_bond_parameters/Cn,Cm,n,m
nbp_coeffs 0.00 0.00 12 6 #N-S_non_bond_parameters/Cn,Cm,n,m
nbp_coeffs 0.00 0.00 12 6 #N-H_non_bond_parameters/Cn,Cm,n,m *
nbp_coeffs 0.00 0.00 12 6 #N-X_non_bond_parameters/Cn,Cm,n,m *
nbp_coeffs 0.00 0.00 12 6 #N-M_non_bond_parameters/Cn,Cm,n,m *
nbp_coeffs 0.00 0.00 12 6 #N-n_non_bond_parameters/Cn,Cm,n,m *
sol_par 0.00 0.00 #O atomic fragmental volume, solvation param.
constant 0.000 #O grid map constant energy

map proteinhx.AD1 hb.O.map #filename of grid map
nbp_coeffs 0.00 0.00 12 6 #O-C_non_bond_parameters/Cn,Cm,n,m
nbp_coeffs 0.00 0.00 12 6 #O-N_non_bond_parameters/Cn,Cm,n,m
nbp_coeffs 0.00 0.00 12 6 #O-O_non_bond_parameters/Cn,Cm,n,m
nbp_coeffs 0.00 0.00 12 6 #O-S_non_bond_parameters/Cn,Cm,n,m
nbp_coeffs 75570.000 23850.000 12 10 #O-H_non_bond_parameters/Cn,Cm,n,m *
nbp_coeffs 0.00 0.00 12 6 #O-X_non_bond_parameters/Cn,Cm,n,m *
nbp_coeffs 0.00 0.00 12 6 #O-M_non_bond_parameters/Cn,Cm,n,m *
nbp_coeffs 0.00 0.00 12 6 #O-n_non_bond_parameters/Cn,Cm,n,m *
sol_par 0.00 0.00 #O atomic fragmental volume, solvation param.
constant 0.000 #O grid map constant energy

map proteinhx.AD1 hb.H.map #filename of grid map
nbp_coeffs 0.00 0.00 12 6 #H-C_non_bond_parameters/Cn,Cm,n,m
nbp_coeffs 0.00 0.00 12 6 #H-N_non_bond_parameters/Cn,Cm,n,m *
nbp_coeffs 75570.000 23850.000 12 10 #H-O_non_bond_parameters/Cn,Cm,n,m
nbp_coeffs 2657200.000 354290.000 12 10 #H-S_non_bond_parameters/Cn,Cm,n,m *
nbp_coeffs 0.00 0.00 12 6 #H-H_non_bond_parameters/Cn,Cm,n,m *
nbp_coeffs 0.00 0.00 12 6 #H-X_non_bond_parameters/Cn,Cm,n,m *
nbp_coeffs 0.00 0.00 12 6 #H-M_non_bond_parameters/Cn,Cm,n,m *
nbp_coeffs 0.00 0.00 12 6 #H-n_non_bond_parameters/Cn,Cm,n,m *
sol_par 0.00 0.0000 #H atomic fragmental volume, solvation param.
constant 0.000 #H grid map constant energy

map proteinhx.AD1 hb.X.map
nbp_coeffs 0.00 0.00 12 6 #X-C_non_bond_parameters/Cn,Cm,n,m *
nbp_coeffs 0.00 0.00 12 6 #X-N_non_bond_parameters/Cn,Cm,n,m *
nbp_coeffs 0.00 0.00 12 6 #X-O_non_bond_parameters/Cn,Cm,n,m *
nbp_coeffs 0.00 0.00 12 6 #X-S_non_bond_parameters/Cn,Cm,n,m *
nbp_coeffs 0.00 0.00 12 6 #X-H_non_bond_parameters/Cn,Cm,n,m *
nbp_coeffs 0.00 0.00 12 6 #X-X_non_bond_parameters/Cn,Cm,n,m *
nbp_coeffs 0.00 0.00 12 6 #X-M_non_bond_parameters/Cn,Cm,n,m *
nbp_coeffs 0.00 0.00 12 6 #X-n_non_bond_parameters/Cn,Cm,n,m *
sol_par 0.00 0.0000 #D atomic fragmental volume, solvation param.
constant 0.000 #D grid map constant energy

map proteinhx.AD1 hb.S.map
nbp_coeffs 0.00 0.00 12 6 #S-C_non_bond_parameters/Cn,Cm,n,m *
nbp_coeffs 0.00 0.00 12 6 #S-N_non_bond_parameters/Cn,Cm,n,m *

nbp_coeffs 0.00 0.00 12 6 #S-O_non_bond_parameters/Cn,Cm,n,m *

nbp_coeffs 0.00 0.00 12 6 #S-S_non_bond_parameters/Cn,Cm,n,m *

nbp_coeffs 2657200.000 354290.000 12 10 #S-H_non_bond_parameters/Cn,Cm,n,m *

nbp_coeffs 0.00 0.00 12 6 #S-X_non_bond_parameters/Cn,Cm,n,m *

nbp_coeffs 0.00 0.00 12 6 #S-M_non_bond_parameters/Cn,Cm,n,m *

nbp_coeffs 0.00 0.00 12 6 #S-n_non_bond_parameters/Cn,Cm,n,m *

sol_par 0.00 0.0000 #D atomic fragmental volume, solvation param.
constant 0.000 #D grid map constant energy

map proteinhx.AD1.hb.P.map

nbp_coeffs 0.00 0.00 12 6 #P-C_non_bond_parameters/Cn,Cm,n,m *

nbp_coeffs 0.00 0.00 12 6 #P-N_non_bond_parameters/Cn,Cm,n,m *

nbp_coeffs 0.00 0.00 12 6 #P-O_non_bond_parameters/Cn,Cm,n,m *

nbp_coeffs 0.00 0.00 12 6 #P-S_non_bond_parameters/Cn,Cm,n,m *

nbp_coeffs 0.00 0.00 12 6 #P-H_non_bond_parameters/Cn,Cm,n,m *

nbp_coeffs 0.00 0.00 12 6 #P-X_non_bond_parameters/Cn,Cm,n,m *

nbp_coeffs 0.00 0.00 12 6 #P-M_non_bond_parameters/Cn,Cm,n,m *

nbp_coeffs 0.00 0.00 12 6 #P-n_non_bond_parameters/Cn,Cm,n,m *

sol_par 0.00 0.0000 #D atomic fragmental volume, solvation param.
constant 0.000 #D grid map constant energy

elecmap proteinhx.AD1.hb.e.map #electrostatic potential map
dielectric -1.0 #<0,distance-dep.diel; >0,constant
#fmap proteinhx.AD1.hb.f.map #floating grid
--- End proteinhxn.AD1.hb.gpf ------

--- Begin proteinhxn.AD1.126.dpf -------

seed time pid # for random number generator
types CNOPSHX # atom type names

fld proteinhx.AD1.126.fld # grid data file
map proteinhx.AD1.126.C.map # C-atomic affinity map file
map proteinhx.AD1.126.N.map # N-atomic affinity map file
map proteinhx.AD1.126.O.map # O-atomic affinity map file
map proteinhx.AD1.126.P.map # P-atomic affinity map file
map proteinhx.AD1.126.S.map # S-atomic affinity map file
map proteinhx.AD1.126.H.map # H-atomic affinity map file
map proteinhx.AD1.126.X.map # X-atomic affinity map file
map proteinhx.AD1.126.e.map # electrostatics maps file
--- End proteinhxn.AD1.126.dpf -------

--- Begin proteinhxn.AD1.hb.dpf -------

seed time pid # for random number generator
types CNOPSHX # atom type names

fld proteinhx.AD1.hb.fld # grid data file
map proteinhx.AD1.hb.C.map # C-atomic affinity map file
map proteinhx.AD1.hb.N.map # N-atomic affinity map file
map proteinhx.AD1.hb.O.map # O-atomic affinity map file
map proteinhx.AD1 hb P.map # P-atomic affinity map file
map proteinhx.AD1 hb S.map # S-atomic affinity map file
map proteinhx.AD1 hb H.map # H-atomic affinity map file
map proteinhx.AD1 hb X.map # X-atomic affinity map file
map proteinhx.AD1 hb e.map # electrostatics map file
--- End proteinhxn.AD1 hb.dpf -------

--- Begin proteinhxn.MM3PRO.126.gpf -------
receptor proteinhx.pdbqs #macromolecule
gridfld proteinhx.mm3pro.126.fld #grid_data_file
npts 70 70 70 #num.grid points in xyz
spacing .375 #spacing (Angstroms)
gridcenter 0.0 0.0 0.0 #xyz-coordinates or "auto"
types CNOHXSP #atom type names
smooth 0.500 #store minimum energy within radius (Angstroms)

map proteinhx.mm3pro.126 C.map
nbp_coeffs 574494.296 249.089 12 6 # C-C
nbp_coeffs 522278.552 266.802 12 6 # C-N
nbp_coeffs 436672.200 264.035 12 6 # C-O
nbp_coeffs 2162349.262 799.230 12 6 # C-S
nbp_coeffs 112449.746 96.690 12 6 # C-H
nbp_coeffs 134267.191 111.715 12 6 # C-X
nbp_coeffs 679387.873 233.858 12 6 # C-Mg
sol_par 0.00 0.00
constant 0.000

map proteinhx.mm3pro.126 N.map
nbp_coeffs 522278.552 266.802 12 6 # N-C
nbp_coeffs 470452.858 284.461 12 6 # N-N
nbp_coeffs 389525.099 280.142 12 6 # N-O
nbp_coeffs 1983041.804 859.807 12 6 # N-S
nbp_coeffs 98195.296 101.501 12 6 # N-H
nbp_coeffs 117486.905 117.395 12 6 # N-X
nbp_coeffs 625432.546 252.064 12 6 # N-Mg
sol_par 0.00 0.00
constant 0.000

map proteinhx.mm3pro.126 O.map
nbp_coeffs 436672.200 264.035 12 6 # O-C
nbp_coeffs 389525.099 280.142 12 6 # O-N
nbp_coeffs 319204.214 274.467 12 6 # O-O
nbp_coeffs 1673353.790 854.820 12 6 # O-S
nbp_coeffs 0.000 0.000 12 6 # O-H
nbp_coeffs 94328.367 113.847 12 6 # O-X
nbp_coeffs 529888.229 251.106 12 6 # O-Mg
sol_par 0.00 0.00
constant 0.000

map proteinhx.mm3pro.126 H.map
nbp_coeffs 112449.746 96.690 12 6 # H-C
nbp_coeffs 98195.296 101.501 12 6 # H-N
nbp_coeffs 0.000 0.000 12 6 # H-O
nbp_coeffs 0.000 0.000 12 6 # H-S
nbp_coeffs 18446.744 34.360 12 6 # H-H
nbp_coeffs 22225.188 39.879 12 6 # H-X
nbp_coeffs 140445.808 93.290 12 6 # H-Mg
sol_par 0.00 0.00
constant 0.000

map proteinhx.mm3pro.126.X.map
nbp_coeffs 134267.191 111.715 12 6 # X-C
nbp_coeffs 117486.905 117.395 12 6 # X-N
nbp_coeffs 94328.367 113.847 12 6 # X-O
nbp_coeffs 523946.429 364.980 12 6 # X-S
nbp_coeffs 22225.188 39.879 12 6 # X-H
nbp_coeffs 26765.177 46.273 12 6 # X-X
nbp_coeffs 167232.616 107.639 12 6 # X-Mg
sol_par 0.00 0.00
constant 0.000

map proteinhx.mm3pro.126.S.map
nbp_coeffs 2162349.262 799.230 12 6 # S-C
nbp_coeffs 1983041.804 859.807 12 6 # S-N
nbp_coeffs 1673353.790 854.820 12 6 # S-O
nbp_coeffs 8071845.421 2553.831 12 6 # S-S
nbp_coeffs 0.000 0.000 12 6 # S-H
nbp_coeffs 523946.429 364.980 12 6 # S-X
nbp_coeffs 2526925.265 745.911 12 6 # S-Mg
sol_par 0.00 0.00
constant 0.000

map proteinhx.mm3pro.126.P.map
nbp_coeffs 2405756.745 805.052 12 6 # P-C
nbp_coeffs 2218022.918 868.375 12 6 # P-N
nbp_coeffs 1882162.117 865.763 12 6 # P-O
nbp_coeffs 8935267.716 2565.954 12 6 # P-S
nbp_coeffs 500581.721 322.199 12 6 # P-H
nbp_coeffs 595860.136 371.695 12 6 # P-X
nbp_coeffs 2791049.267 748.624 12 6 # P-Mg
sol_par 0.00 0.00
constant 0.000

elecmap proteinhx.mm3pro.126.e.map # electrostatic potential map
dielectric -1.0 # <0,distance-dep.diel; >0,constant
# fmap proteinhx.AD1.solv.f.map # floating grid
--- End proteinhxn.MM3PRO.126.gpf ------

--- Begin proteinhxn.MM3PRO.hb.gpf ------
receptor proteinhx.pdbqs # macromolecule
gridfld proteinhx.mm3pro.hb.fld # grid data file
npts 70 70 70 # num.grid points in xyz
spacing .375 # spacing (Angstroms)
gridcenter 0.0 0.0 0.0 # xyz-coordinates or "auto"
types CNOHXSP #atom type names

smooth 0.500 #store minimum energy within radius (Angstroms)

map proteinhx.mm3pro.hb.C.map
nbp_coeffs 0.000 0.000 12 6 # C-C
nbp_coeffs 0.000 0.000 12 6 # C-N
nbp_coeffs 0.000 0.000 12 6 # C-O
nbp_coeffs 0.000 0.000 12 6 # C-S
nbp_coeffs 0.000 0.000 12 6 # C-H
nbp_coeffs 0.000 0.000 12 6 # C-X
nbp_coeffs 0.000 0.000 12 6 # C-Mg
sol_par 0.00 0.00
constant 0.000

map proteinhx.mm3pro.hb.N.map
nbp_coeffs 0.000 0.000 12 6 # N-C
nbp_coeffs 0.000 0.000 12 6 # N-N
nbp_coeffs 0.000 0.000 12 6 # N-O
nbp_coeffs 0.000 0.000 12 6 # N-S
nbp_coeffs 0.000 0.000 12 6 # N-H
nbp_coeffs 0.000 0.000 12 6 # N-X
nbp_coeffs 0.000 0.000 12 6 # N-Mg
sol_par 0.00 0.00
constant 0.000

map proteinhx.mm3pro.hb.O.map
nbp_coeffs 0.000 0.000 12 6 # O-C
nbp_coeffs 0.000 0.000 12 6 # O-N
nbp_coeffs 0.000 0.000 12 6 # O-O
nbp_coeffs 0.000 0.000 12 6 # O-S
nbp_coeffs 4358.837 1657.949 12 10 # O-H
nbp_coeffs 0.000 0.000 12 6 # O-X
nbp_coeffs 0.000 0.000 12 6 # O-Mg
sol_par 0.00 0.00
constant 0.000

map proteinhx.mm3pro.hb.H.map
nbp_coeffs 0.000 0.000 12 6 # H-C
nbp_coeffs 0.000 0.000 12 6 # H-N
nbp_coeffs 4358.837 1657.949 12 10 # H-O
nbp_coeffs 185332.377 32510.604 12 10 # H-S
nbp_coeffs 0.000 0.000 12 6 # H-H
nbp_coeffs 0.000 0.000 12 6 # H-X
nbp_coeffs 0.000 0.000 12 6 # H-Mg
sol_par 0.00 0.00
constant 0.000

map proteinhx.mm3pro.hb.X.map
nbp_coeffs 0.000 0.000 12 6 # X-C
nbp_coeffs 0.000 0.000 12 6 # X-N
nbp_coeffs 0.000 0.000 12 6 # X-O
nbp_coeffs 0.000 0.000 12 6 # X-S
nbp_coeffs 0.000 0.000 12 6 # X-H
nbp_coeffs 0.000 0.000 12 6 # X-X
nbp_coeffs 0.000 0.000 12 6 # X-Mg
sol_par 0.00 0.00
constant 0.000

map proteinhx.mm3pro hb S.map
nbp_coeffs 0.000 0.000 12 6 # S-C
nbp_coeffs 0.000 0.000 12 6 # S-N
nbp_coeffs 0.000 0.000 12 6 # S-O
nbp_coeffs 185332.377 32510.604 12 10 # S-H
nbp_coeffs 0.000 0.000 12 6 # S-X
nbp_coeffs 0.000 0.000 12 6 # S-Mg
sol_par 0.00 0.00
constant 0.000

elecmap proteinhx.mm3pro hb e.map #electrostatic potential map
dielectric -1.0 #<0,distance-dep.diel; >0,constant
#fmap proteinhx.AD1.solv.f.map #floating grid

--- End proteinhxn.MM3PRO hb.gpf ---

--- Begin proteinhxn.MM3PRO.126.dpf ---
seed time pid # for random number generator
types CNOHXSP # atom type names
fld proteinhx.mm3pro.126 fld #grid_data_file
map proteinhx.mm3pro.126 C.map
map proteinhx.mm3pro.126 N.map
map proteinhx.mm3pro.126 O.map
map proteinhx.mm3pro.126 H.map
map proteinhx.mm3pro.126 X.map
map proteinhx.mm3pro.126 S.map
map proteinhx.mm3pro.126 P.map
map proteinhx.mm3pro.126 e.map #electrostatic potential map
#fmap proteinhx.AD1.solv.f.map #floating grid

--- End proteinhxn.MM3PRO.126.dpf ---

--- Begin proteinhxn.MM3PRO hb.dpf ---
seed time pid # for random number generator
types CNOHXSP  # atom type names
fld proteinhx.mm3pro.hb.fld  #grid_data_file
map proteinhx.mm3pro.hb.C.map
map proteinhx.mm3pro.hb.N.map
map proteinhx.mm3pro.hb.O.map
map proteinhx.mm3pro.hb.H.map
map proteinhx.mm3pro.hb.X.map
map proteinhx.mm3pro.hb.S.map
map proteinhx.mm3pro.hb.P.map
map proteinhx.mm3pro.hb.e.map  #electrostatic potential map
#fmap proteinhx.AD1.solv.f.map  #floating grid

--- End proteinhxn.MM3PRO.hb.dpf -------

--- Begin proteinhxn.tsolv.gpf -------
receptor proteinhx.pdbqs  #macromolecule
gridfld proteinhx.AD1.tsolv.fld  #grid_data_file
npts 70 70 70  #num.grid points in xyz
spacing .375  #spacing (Angstroms)
gridcenter 0.0 0.0 0.0  #xyz-coordinates or "auto"
types CNOHXSP  #atom type names
smooth 0.500  #store minimum energy within radius (Angstroms)
map proteinhx.AD1.tsolv.C.map  #filename of grid map
nbp_coeffs 0.00 0.00 12 6  #C-C_non_bond_parameters/Cn,Cm,n,m
nbp_coeffs 0.00 0.00 12 6  #C-N_non_bond_parameters/Cn,Cm,n,m
nbp_coeffs 0.00 0.00 12 6  #C-O_non_bond_parameters/Cn,Cm,n,m
nbp_coeffs 0.00 0.00 12 6  #C-S_non_bond_parameters/Cn,Cm,n,m
nbp_coeffs 0.00 0.00 12 6  #C-H_non_bond_parameters/Cn,Cm,n,m
nbp_coeffs 0.00 0.00 12 6  #C-X_non_bond_parameters/Cn,Cm,n,m *
nbp_coeffs 0.00 0.00 12 6  #C-M_non_bond_parameters/Cn,Cm,n,m *
sol_par 12.77 -4.00  #C atomic fragmental volume, solvation param.
constant 0.000  #C grid map constant energy

map proteinhx.AD1.tsolv.N.map  #filename of grid map
nbp_coeffs 0.00 0.00 12 6  #N-C_non_bond_parameters/Cn,Cm,n,m
nbp_coeffs 0.00 0.00 12 6  #N-N_non_bond_parameters/Cn,Cm,n,m
nbp_coeffs 0.00 0.00 12 6  #N-O_non_bond_parameters/Cn,Cm,n,m
nbp_coeffs 0.00 0.00 12 6  #N-S_non_bond_parameters/Cn,Cm,n,m
nbp_coeffs 0.00 0.00 12 6  #N-H_non_bond_parameters/Cn,Cm,n,m *
nbp_coeffs 0.00 0.00 12 6  #N-X_non_bond_parameters/Cn,Cm,n,m *
nbp_coeffs 0.00 0.00 12 6  #N-M_non_bond_parameters/Cn,Cm,n,m *
nbp_coeffs 0.00 0.00 12 6  #N-n_non_bond_parameters/Cn,Cm,n,m *
sol_par 9.00 17.40  #N atomic fragmental volume, solvation param.
constant 0.000  #N grid map constant energy

map proteinhx.AD1.tsolv.O.map  #filename of grid map
nbp_coeffs 0.00 0.00 12 6  #O-C_non_bond_parameters/Cn,Cm,n,m
nbp_coeffs 0.00 0.00 12 6  #O-N_non_bond_parameters/Cn,Cm,n,m
nbp_coeffs 0.00 0.00 12 6 #O-O_non_bond_parameters/Cn,Cm,n,m
nbp_coeffs 0.00 0.00 12 6 #O-S_non_bond_parameters/Cn,Cm,n,m
nbp_coeffs 0.00 0.00 12 6 #O-H_non_bond_parameters/Cn,Cm,n,m *
ng_coeffs 0.00 0.00 12 6 #O-X_non_bond_parameters/Cn,Cm,n,m *
ng_coeffs 0.00 0.00 12 6 #O-M_non_bond_parameters/Cn,Cm,n,m *
ng_coeffs 0.00 0.00 12 6 #O-n_non_bond_parameters/Cn,Cm,n,m *
sol_par -11.04 17.40 #O atomic fragmental volume, solvation param.
constant 0.000 #O grid map constant energy

map proteinhx.AD1.tsolv.H.map #filename of grid map
nbp_coeffs 0.00 0.00 12 6 #H-C_non_bond_parameters/Cn,Cm,n,m
nbp_coeffs 0.00 0.00 12 6 #H-N_non_bond_parameters/Cn,Cm,n,m *
ng_coeffs 0.00 0.00 12 6 #H-O_non_bond_parameters/Cn,Cm,n,m *
ng_coeffs 0.00 0.00 12 6 #H-H_non_bond_parameters/Cn,Cm,n,m
nbp_coeffs 0.00 0.00 12 6 #H-X_non_bond_parameters/Cn,Cm,n,m *
ng_coeffs 0.00 0.00 12 6 #H-M_non_bond_parameters/Cn,Cm,n,m *
ng_coeffs 0.00 0.00 12 6 #H-n_non_bond_parameters/Cn,Cm,n,m *
sol_par 0.00 0.0000 #H atomic fragmental volume, solvation param.
constant 0.000 #H grid map constant energy

map proteinhx.AD1.tsolv.X.map
nbp_coeffs 0.00 0.00 12 6 #X-C_non_bond_parameters/Cn,Cm,n,m *
ng_coeffs 0.00 0.00 12 6 #X-N_non_bond_parameters/Cn,Cm,n,m *
ng_coeffs 0.00 0.00 12 6 #X-O_non_bond_parameters/Cn,Cm,n,m *
ng_coeffs 0.00 0.00 12 6 #X-H_non_bond_parameters/Cn,Cm,n,m *
ng_coeffs 0.00 0.00 12 6 #X-X_non_bond_parameters/Cn,Cm,n,m *
ng_coeffs 0.00 0.00 12 6 #X-M_non_bond_parameters/Cn,Cm,n,m *
ng_coeffs 0.00 0.00 12 6 #X-n_non_bond_parameters/Cn,Cm,n,m *
sol_par 0.00 0.0000 #X atomic fragmental volume, solvation param.
constant 0.000 #X grid map constant energy

map proteinhx.AD1.tsolv.S.map
nbp_coeffs 0.00 0.00 12 6 #S-C_non_bond_parameters/Cn,Cm,n,m *
ng_coeffs 0.00 0.00 12 6 #S-N_non_bond_parameters/Cn,Cm,n,m *
ng_coeffs 0.00 0.00 12 6 #S-O_non_bond_parameters/Cn,Cm,n,m *
ng_coeffs 0.00 0.00 12 6 #S-H_non_bond_parameters/Cn,Cm,n,m *
ng_coeffs 0.00 0.00 12 6 #S-X_non_bond_parameters/Cn,Cm,n,m *
ng_coeffs 0.00 0.00 12 6 #S-M_non_bond_parameters/Cn,Cm,n,m *
ng_coeffs 0.00 0.00 12 6 #S-n_non_bond_parameters/Cn,Cm,n,m *
sol_par -16.39 6.40 #S atomic fragmental volume, solvation param.
constant 0.000 #S grid map constant energy

map proteinhx.AD1.tsolv.P.map
nbp_coeffs 0.00 0.00 12 6 #P-C_non_bond_parameters/Cn,Cm,n,m *
ng_coeffs 0.00 0.00 12 6 #P-N_non_bond_parameters/Cn,Cm,n,m *
ng_coeffs 0.00 0.00 12 6 #P-O_non_bond_parameters/Cn,Cm,n,m *
ng_coeffs 0.00 0.00 12 6 #P-H_non_bond_parameters/Cn,Cm,n,m *
ng_coeffs 0.00 0.00 12 6 #P-X_non_bond_parameters/Cn,Cm,n,m *
ng_coeffs 0.00 0.00 12 6 #P-M_non_bond_parameters/Cn,Cm,n,m *
ng_coeffs 0.00 0.00 12 6 #P-n_non_bond_parameters/Cn,Cm,n,m *
sol_par -6.00 4.00 #D atomic fragmental volume, solvation param.
constant 0.000 #D grid map constant energy

elecmap proteinhx.AD1.tsolv.e.map #electrostatic potential map
dielectric -1.0 #<0,distance-dep.diel; >0,constant
#fmap proteinhx.AD1.solv.f.map #floating grid
--- End proteinhxn.tsolv.gpf -------

--- Begin proteinhxn.tsolv.dpf -------
seed time pid # for random number generator
types CNOPSHX # atom type names

fld proteinhx.AD1.tsolv.fld # grid data file
map proteinhx.AD1.tsolv.C.map # C-atomic affinity map file
map proteinhx.AD1.tsolv.N.map # N-atomic affinity map file
map proteinhx.AD1.tsolv.O.map # O-atomic affinity map file
map proteinhx.AD1.tsolv.P.map # P-atomic affinity map file
map proteinhx.AD1.tsolv.S.map # S-atomic affinity map file
map proteinhx.AD1.tsolv.H.map # H-atomic affinity map file
map proteinhx.AD1.tsolv.X.map # X-atomic affinity map file
map proteinhx.AD1.tsolv.e.map # electrostatics map file
--- End proteinhxn.tsolv.dpf -------
Appendix F: Complete Family 1 Docking Energies

Binding energies of all dockings resulting in a catalytically viable configuration are listed in the tables below. The lowest energy for particular enzyme is shown in bold, with a teal background; all energies within one standard deviation of the lowest energy are shown with a green background. Energies are in kcal/mol.

<table>
<thead>
<tr>
<th>PDB</th>
<th>Species</th>
<th>Enzyme</th>
<th>Substrates</th>
<th>Binding Energy (kcal/mol)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1bgg</td>
<td><em>Paenibacillus polymyxa</em></td>
<td>beta-glucosidase</td>
<td>all gal glc man 6-P-gal 6-P-glc</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>1-2</td>
<td>-165.81 gal -173.11 glc -169.75 man -101.03 6-P-gal -113.74 6-P-glc</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>1-3</td>
<td>-174.32 gal -194.77 glc -187.42 man -105.90 6-P-gal -85.52 6-P-glc</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>1-4</td>
<td>-163.45 gal -180.94 glc -170.76 man -162.50 6-P-gal -114.11 6-P-glc</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>1-6</td>
<td>-176.32 gal -170.27 glc -174.87 man -113.74 6-P-gal -113.74 6-P-glc</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1-1-thio</td>
<td>-166.94 gal -180.44 glc -190.41 man -165.18 6-P-gal -113.74 6-P-glc</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1cbg</td>
<td><em>Trifolium repens</em></td>
<td>cyanogenic beta-glucosidase</td>
<td>all gal glc man 6-P-gal 6-P-glc</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>1-2</td>
<td>-141.08 gal -153.13 glc -173.94 man -168.43 6-P-gal -133.74 6-P-glc</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>1-3</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1-4</td>
<td>-160.78 gal -179.26 glc -182.33 man -151.37 6-P-gal -81.36 6-P-glc</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>1-6</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td></td>
<td>1-1-thio</td>
<td>-167.85 gal -179.53 glc -179.68 man -165.18 6-P-gal -116.09 6-P-glc</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1ef</td>
<td><em>Zea mays</em></td>
<td>beta-glucosidase</td>
<td>all gal glc man 6-P-gal 6-P-glc</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>1-2</td>
<td>-189.33 gal -185.00 glc -203.25 man -189.34 6-P-gal -196.07 6-P-glc</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>1-3</td>
<td>-190.24 gal -204.57 glc -201.15 man -190.21 6-P-gal -189.97 6-P-glc</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>1-4</td>
<td>-176.18 gal -195.09 glc -194.55 man -177.30 6-P-gal -150.11 6-P-glc</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>1-6</td>
<td>-202.64 gal -222.99 glc -218.28 man -208.99 6-P-gal -209.08 6-P-glc</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1-1-thio</td>
<td>-178.97 gal -203.59 glc -209.79 man -184.48 6-P-gal -139.96 6-P-glc</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1ex</td>
<td><em>Sinapis alba</em></td>
<td>myrosinase</td>
<td>all gal glc man 6-P-gal 6-P-glc</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>1-2</td>
<td>-143.16 gal -140.29 glc -154.10 man -129.90 6-P-gal -147.98 6-P-glc</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>1-3</td>
<td>-160.33 gal -163.37 glc -155.37 man -160.51 6-P-gal -127.31 6-P-glc</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>1-4</td>
<td>-149.45 gal -168.52 glc -141.99 man -143.08 6-P-gal -133.60 6-P-glc</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>1-6</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td></td>
<td>1-1-thio</td>
<td>-160.71 gal -168.66 glc -188.37 man -154.20 6-P-gal -139.96 6-P-glc</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-------</td>
<td>-------------------------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
</tr>
<tr>
<td><strong>1gnx</strong> Streptomyces sp. beta-glucosidase</td>
<td><strong>all</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>1-2</strong></td>
<td>163.46</td>
<td>-176.98</td>
<td>-182.84</td>
<td>-164.72</td>
</tr>
<tr>
<td><strong>1-3</strong></td>
<td>177.26</td>
<td><strong>-191.77</strong></td>
<td>--</td>
<td>-185.87</td>
</tr>
<tr>
<td><strong>1-4</strong></td>
<td>168.37</td>
<td>-178.36</td>
<td>-174.27</td>
<td>-159.28</td>
</tr>
<tr>
<td><strong>1-6</strong></td>
<td>171.07</td>
<td>-183.20</td>
<td>-178.53</td>
<td>-177.92</td>
</tr>
<tr>
<td><strong>1-1-thio</strong></td>
<td>159.12</td>
<td>-165.53</td>
<td>-187.38</td>
<td>-159.74</td>
</tr>
<tr>
<td><strong>1gow</strong> Sulfolobus solfataricus beta-glycosidase</td>
<td><strong>all</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>1-2</strong></td>
<td>--</td>
<td>-164.59</td>
<td>-170.19</td>
<td>-145.32</td>
</tr>
<tr>
<td><strong>1-3</strong></td>
<td>169.45</td>
<td><strong>-194.75</strong></td>
<td>-176.10</td>
<td>-189.06</td>
</tr>
<tr>
<td><strong>1-4</strong></td>
<td>154.09</td>
<td>-165.36</td>
<td>-160.12</td>
<td>-159.46</td>
</tr>
<tr>
<td><strong>1-6</strong></td>
<td>169.41</td>
<td>-179.75</td>
<td>-172.51</td>
<td>-179.22</td>
</tr>
<tr>
<td><strong>1-1-thio</strong></td>
<td>-170.53</td>
<td>-182.89</td>
<td>-181.80</td>
<td>-156.84</td>
</tr>
<tr>
<td><strong>1hxj</strong> Zea mays beta-glucosidase</td>
<td><strong>all</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>1-2</strong></td>
<td>179.93</td>
<td>-181.52</td>
<td>-206.74</td>
<td>-169.02</td>
</tr>
<tr>
<td><strong>1-3</strong></td>
<td>193.13</td>
<td>-204.62</td>
<td>-199.92</td>
<td>-192.67</td>
</tr>
<tr>
<td><strong>1-4</strong></td>
<td>188.29</td>
<td>-194.37</td>
<td>-194.21</td>
<td>-185.02</td>
</tr>
<tr>
<td><strong>1-6</strong></td>
<td>-207.17</td>
<td>--</td>
<td><strong>-217.18</strong></td>
<td>--</td>
</tr>
<tr>
<td><strong>1-1-thio</strong></td>
<td>-177.58</td>
<td>-190.13</td>
<td>-203.62</td>
<td>-175.60</td>
</tr>
<tr>
<td><strong>1np2</strong> Thermus nonproteolyticus beta-glycosidase</td>
<td><strong>all</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>1-2</strong></td>
<td>176.05</td>
<td>-176.70</td>
<td>-176.87</td>
<td>-170.00</td>
</tr>
<tr>
<td><strong>1-3</strong></td>
<td>163.98</td>
<td>-179.94</td>
<td>-174.43</td>
<td>--</td>
</tr>
<tr>
<td><strong>1-4</strong></td>
<td>161.77</td>
<td>-172.00</td>
<td>-166.69</td>
<td>-164.93</td>
</tr>
<tr>
<td><strong>1-6</strong></td>
<td>168.53</td>
<td>--</td>
<td>-176.83</td>
<td>-177.35</td>
</tr>
<tr>
<td><strong>1-1-thio</strong></td>
<td>-166.36</td>
<td>-173.32</td>
<td><strong>-189.96</strong></td>
<td>-187.08</td>
</tr>
<tr>
<td><strong>1od0</strong> Thermotoga maritima beta-glucosidase</td>
<td><strong>all</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>1-2</strong></td>
<td>168.40</td>
<td>-181.49</td>
<td>-194.28</td>
<td>-179.87</td>
</tr>
<tr>
<td><strong>1-3</strong></td>
<td>177.81</td>
<td>--</td>
<td>-190.26</td>
<td>-184.42</td>
</tr>
<tr>
<td><strong>1-4</strong></td>
<td>173.32</td>
<td>-180.96</td>
<td>-182.42</td>
<td>-178.77</td>
</tr>
<tr>
<td><strong>1-6</strong></td>
<td>179.70</td>
<td><strong>-198.88</strong></td>
<td>-181.28</td>
<td>-178.57</td>
</tr>
<tr>
<td><strong>1-1-thio</strong></td>
<td>-171.14</td>
<td>-175.55</td>
<td>-183.31</td>
<td>-159.36</td>
</tr>
<tr>
<td><strong>1pbg</strong> Lactococcus lactis 6-P-beta-galactosidase</td>
<td><strong>all</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>1-2</strong></td>
<td>-128.19</td>
<td>-135.36</td>
<td>-142.31</td>
<td>-150.86</td>
</tr>
<tr>
<td><strong>1-3</strong></td>
<td>-135.95</td>
<td>-149.35</td>
<td>--</td>
<td>-140.83</td>
</tr>
<tr>
<td><strong>1-4</strong></td>
<td>--</td>
<td>-137.57</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td><strong>1-6</strong></td>
<td>-139.46</td>
<td>-152.57</td>
<td>-147.43</td>
<td>-147.52</td>
</tr>
<tr>
<td><strong>1-1-thio</strong></td>
<td>-130.41</td>
<td>-149.13</td>
<td><strong>-173.88</strong></td>
<td>-123.83</td>
</tr>
<tr>
<td>1qox</td>
<td>Bacillus circulans beta-glucosidase</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td></td>
<td>all</td>
<td>gal</td>
<td>glc</td>
<td>man</td>
</tr>
<tr>
<td>1-2</td>
<td>--</td>
<td>--</td>
<td>-141.86</td>
<td>-139.92</td>
</tr>
<tr>
<td>1-3</td>
<td>--</td>
<td>--</td>
<td>-141.12</td>
<td>-168.12</td>
</tr>
<tr>
<td>1-4</td>
<td>-134.21</td>
<td>-129.69</td>
<td>-148.24</td>
<td>-143.79</td>
</tr>
<tr>
<td>1-6</td>
<td>-162.68</td>
<td>-164.20</td>
<td>-152.51</td>
<td>-159.64</td>
</tr>
<tr>
<td>1-1-thio</td>
<td>-149.75</td>
<td>-140.15</td>
<td>-179.47</td>
<td>-126.30</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>1qvβ</th>
<th>Thermosphaera aggregans beta-glycosidase</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>all</td>
<td>gal</td>
<td>glc</td>
<td>man</td>
<td>6-P-gal</td>
<td>6-P-glc</td>
</tr>
<tr>
<td>1-2</td>
<td>-148.30</td>
<td>-176.29</td>
<td>--</td>
<td>-154.88</td>
<td>--</td>
<td>-86.34</td>
</tr>
<tr>
<td>1-3</td>
<td>-194.84</td>
<td>-208.26</td>
<td>-201.14</td>
<td>-197.48</td>
<td>-128.28</td>
<td>-129.82</td>
</tr>
<tr>
<td>1-4</td>
<td>-170.38</td>
<td>-187.11</td>
<td>-185.43</td>
<td>-167.27</td>
<td>--</td>
<td>-115.32</td>
</tr>
<tr>
<td>1-6</td>
<td>-186.23</td>
<td>-209.75</td>
<td>-200.13</td>
<td>-191.26</td>
<td>-133.78</td>
<td>-114.23</td>
</tr>
<tr>
<td>1-1-thio</td>
<td>-170.47</td>
<td>-183.98</td>
<td>-194.00</td>
<td>-171.05</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>1ug6</th>
<th>Thermus thermophilus beta-glycosidase</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>all</td>
<td>gal</td>
<td>glc</td>
<td>man</td>
<td>6-P-gal</td>
<td>6-P-glc</td>
</tr>
<tr>
<td>1-2</td>
<td>-172.84</td>
<td>-191.30</td>
<td>--</td>
<td>-177.76</td>
<td>--</td>
<td>-121.88</td>
</tr>
<tr>
<td>1-3</td>
<td>-189.92</td>
<td>-194.56</td>
<td>-195.24</td>
<td>-184.56</td>
<td>-98.02</td>
<td>-167.30</td>
</tr>
<tr>
<td>1-4</td>
<td>-171.30</td>
<td>-185.25</td>
<td>-177.61</td>
<td>-160.92</td>
<td>--</td>
<td>-114.01</td>
</tr>
<tr>
<td>1-6</td>
<td>-178.40</td>
<td>-189.19</td>
<td>-180.85</td>
<td>-179.69</td>
<td>-135.12</td>
<td>-149.73</td>
</tr>
<tr>
<td>1-1-thio</td>
<td>-175.52</td>
<td>-182.94</td>
<td>-189.34</td>
<td>-171.14</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>1v02</th>
<th>Sorghum bicolor dhurrinase</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>all</td>
<td>gal</td>
<td>glc</td>
<td>man</td>
<td>6-P-gal</td>
<td>6-P-glc</td>
</tr>
<tr>
<td>1-2</td>
<td>-127.15</td>
<td>-140.06</td>
<td>-169.87</td>
<td>-117.01</td>
<td>-133.58</td>
<td>--</td>
</tr>
<tr>
<td>1-3</td>
<td>-180.98</td>
<td>-185.76</td>
<td>-185.90</td>
<td>-184.77</td>
<td>-116.89</td>
<td>-89.19</td>
</tr>
<tr>
<td>1-4</td>
<td>-169.05</td>
<td>-184.96</td>
<td>-175.48</td>
<td>-173.28</td>
<td>-138.69</td>
<td>--</td>
</tr>
<tr>
<td>1-6</td>
<td>-165.91</td>
<td>-196.92</td>
<td>-175.26</td>
<td>-175.32</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>1-1-thio</td>
<td>-163.37</td>
<td>-163.15</td>
<td>-197.55</td>
<td>-152.14</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>1vff</th>
<th>Pyrococcus horikoshii beta-glycosidase</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>all</td>
<td>gal</td>
<td>glc</td>
<td>man</td>
<td>6-P-gal</td>
<td>6-P-glc</td>
</tr>
<tr>
<td>1-2</td>
<td>-155.85</td>
<td>-168.70</td>
<td>-178.03</td>
<td>-160.60</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>1-3</td>
<td>-168.77</td>
<td>-183.86</td>
<td>-173.46</td>
<td>-173.21</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>1-4</td>
<td>-231.97</td>
<td>-239.87</td>
<td>-240.46</td>
<td>-240.46</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>1-6</td>
<td>-182.23</td>
<td>-188.59</td>
<td>-180.08</td>
<td>-176.95</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>1-1-thio</td>
<td>-222.39</td>
<td>-231.59</td>
<td>-181.91</td>
<td>-226.39</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>1wgc</th>
<th>Brevicoryne brassicae  myrosinase</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>all</td>
<td>gal</td>
<td>glc</td>
<td>man</td>
<td>6-P-gal</td>
<td>6-P-glc</td>
</tr>
<tr>
<td>1-2</td>
<td>-164.97</td>
<td>-171.88</td>
<td>-177.68</td>
<td>-174.18</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>1-3</td>
<td>-176.98</td>
<td>-181.91</td>
<td>-179.52</td>
<td>-180.30</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>1-4</td>
<td>-217.89</td>
<td>-226.66</td>
<td>-216.89</td>
<td>-222.83</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>1-6</td>
<td>-181.29</td>
<td>-197.41</td>
<td>-184.68</td>
<td>-184.07</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>1-1-thio</td>
<td>-226.43</td>
<td>-237.86</td>
<td>-178.22</td>
<td>-244.93</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>all</td>
<td>gal</td>
<td>glc</td>
<td>man</td>
<td>6-P-gal</td>
<td>6-P-glc</td>
</tr>
<tr>
<td>----------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
<td>---------</td>
<td>---------</td>
</tr>
<tr>
<td>1-2</td>
<td>-181.11</td>
<td>-178.50</td>
<td>-189.15</td>
<td>-169.43</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>1-3</td>
<td>-185.36</td>
<td>-185.08</td>
<td>-187.60</td>
<td>-184.57</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>1-4</td>
<td>-226.23</td>
<td>-243.87</td>
<td>-241.32</td>
<td>-227.60</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>1-6</td>
<td>-193.90</td>
<td><strong>211.12</strong></td>
<td>-200.48</td>
<td>-194.74</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>1-1-thio</td>
<td>-222.30</td>
<td>-235.37</td>
<td>-188.70</td>
<td>-237.30</td>
<td>--</td>
<td>--</td>
</tr>
</tbody>
</table>
Appendix G: Complete Family 1 ANOVA Analysis

**Table 1.** Second-level interactions in ANOVA. The residue numbering is based upon *Zea mays* β-glucosidase unless the residue is not present there, in which case the PDB ID of the residue source is shown. “Gly.bond” is the probability of the observed interaction between the glycon and bond configuration happening randomly, “aa.bond” the same between the amino acid and bond configuration, and “aa.gly” between the amino acid and the glycon. Interactions significant at $\alpha = 0.10$ are shown in red.

<table>
<thead>
<tr>
<th>Residue</th>
<th>Residue substitutions</th>
<th>gly.bond</th>
<th>aa.bond</th>
<th>aa.gly</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ser35</td>
<td>Ser, Ala</td>
<td>0.981</td>
<td>0.641</td>
<td>0.682</td>
</tr>
<tr>
<td>Gln38</td>
<td>Gln</td>
<td>0.933</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Trp51 (1E6X)</td>
<td>Trp, (del)</td>
<td>0.146</td>
<td>0.473</td>
<td>0.457</td>
</tr>
<tr>
<td>Arg96</td>
<td>Arg, (del)</td>
<td>0.964</td>
<td>0.666</td>
<td>0.743</td>
</tr>
<tr>
<td>His142</td>
<td>His</td>
<td>0.933</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Trp143</td>
<td>Trp, Phe</td>
<td>0.933</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Asn190</td>
<td>Asn</td>
<td>0.933</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Glu191</td>
<td>Glu</td>
<td>0.937</td>
<td>&lt; 0.001</td>
<td>0.074</td>
</tr>
<tr>
<td>Gln193</td>
<td>Trp, Gln, Tyr, Asn, Gly, Glu</td>
<td>0.902</td>
<td>&lt; 0.001</td>
<td>0.454</td>
</tr>
<tr>
<td>Thr194</td>
<td>Gly, Thr, Ser, Cys, Val, Pro</td>
<td>0.948</td>
<td>&lt; 0.001</td>
<td>0.099</td>
</tr>
<tr>
<td>Phe195</td>
<td>Val, Phe, Ile, (del)</td>
<td>0.968</td>
<td>0.803</td>
<td>0.920</td>
</tr>
<tr>
<td>Asp166 (1PBG)</td>
<td>Met, Asp, Glu, (del)</td>
<td>0.966</td>
<td>0.942</td>
<td>0.896</td>
</tr>
<tr>
<td>Phe198</td>
<td>Asn, Phe, Arg, Leu, Val, Gly, Gln</td>
<td>0.872</td>
<td>&lt; 0.001</td>
<td>0.177</td>
</tr>
<tr>
<td>Val204</td>
<td>Thr, Val, Ser, Glu, Lys, Gly, (del)</td>
<td>0.909</td>
<td>&lt; 0.001</td>
<td>0.276</td>
</tr>
<tr>
<td>Phe205</td>
<td>Phe, Asp, His, Leu</td>
<td>0.905</td>
<td>&lt; 0.001</td>
<td>0.091</td>
</tr>
<tr>
<td>Phe260</td>
<td>Leu, Phe, Met, Tyr, His, Pro</td>
<td>0.870</td>
<td>&lt; 0.001</td>
<td>0.118</td>
</tr>
<tr>
<td>Asp261</td>
<td>Val, Asp, Ile, Asn, Ala</td>
<td>0.883</td>
<td>&lt; 0.001</td>
<td>0.105</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
</tr>
<tr>
<td>Val262</td>
<td>Ser, Val, Thr, Ile, Asn, Phe, Leu</td>
<td>0.989</td>
<td>0.760</td>
<td>0.992</td>
</tr>
<tr>
<td>Met263</td>
<td>His, Met, Arg, Ser, Ala, Gly, Pro, Gln, Phe</td>
<td>0.872</td>
<td>&lt; 0.001</td>
<td>0.500</td>
</tr>
<tr>
<td>Gly264</td>
<td>Trp, Gly, Ser, Pro, Tyr, Thr, (del)</td>
<td>0.975</td>
<td>0.919</td>
<td>0.989</td>
</tr>
<tr>
<td>Arg265</td>
<td>Arg, Phe, Lys, (del)</td>
<td>0.949</td>
<td>0.927</td>
<td>0.987</td>
</tr>
<tr>
<td>Asn285</td>
<td>Met, Asn, Phe, His, Ser, Cys, (del)</td>
<td>0.726</td>
<td>&lt; 0.001</td>
<td>0.020</td>
</tr>
<tr>
<td>Asn243 (1PBG)</td>
<td>Asn, (del)</td>
<td>0.938</td>
<td>0.908</td>
<td>0.986</td>
</tr>
<tr>
<td>Asn331</td>
<td>Asn</td>
<td>0.933</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Tyr332</td>
<td>Tyr</td>
<td>0.933</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Tyr333</td>
<td>Tyr</td>
<td>0.933</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Thr334</td>
<td>Ser, Thr, Phe, Ala, Met, Ser</td>
<td>0.912</td>
<td>&lt; 0.001</td>
<td>0.124</td>
</tr>
<tr>
<td>Ser335</td>
<td>Ser, Thr, Pro, Arg, (del)</td>
<td>0.937</td>
<td>&lt; 0.001</td>
<td>0.248</td>
</tr>
<tr>
<td>Arg336</td>
<td>Tyr333, Arg, Gln, Thr, Val, His, Asp, Leu, (del)</td>
<td>0.656</td>
<td>&lt; 0.001</td>
<td>0.093</td>
</tr>
<tr>
<td>Gly341 (1GOW)</td>
<td>Tyr, Gly, Phe, (del)</td>
<td>0.943</td>
<td>&lt; 0.001</td>
<td>0.418</td>
</tr>
<tr>
<td>His342 (1GOW)</td>
<td>His, (del)</td>
<td>0.942</td>
<td>0.735</td>
<td>0.998</td>
</tr>
<tr>
<td>Asn376</td>
<td>Ser, Asn, Glu, Met, Phe, Arg, Ile</td>
<td>0.864</td>
<td>&lt; 0.001</td>
<td>0.041</td>
</tr>
<tr>
<td>Trp378</td>
<td>Trp, Ile</td>
<td>0.937</td>
<td>&lt; 0.001</td>
<td>0.074</td>
</tr>
<tr>
<td>Ile379</td>
<td>Leu, Ile, Phe, (del)</td>
<td>0.953</td>
<td>&lt; 0.001</td>
<td>0.141</td>
</tr>
<tr>
<td>Asn278 (1V02)</td>
<td>Asn, (del)</td>
<td>0.922</td>
<td>0.372</td>
<td>0.915</td>
</tr>
<tr>
<td>Glu406</td>
<td>Glu</td>
<td>0.933</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Gly408</td>
<td>Gly, (del)</td>
<td>0.933</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Ile390 (1GOW)</td>
<td>Ile, Met, (del)</td>
<td>0.933</td>
<td>0.605</td>
<td>0.997</td>
</tr>
<tr>
<td>Gly410</td>
<td>Asn, Gly, Ser, Ala, Cys, (del)</td>
<td>0.932</td>
<td>0.526</td>
<td>0.508</td>
</tr>
<tr>
<td>Tyr401 (1GOW)</td>
<td>Tyr, (del)</td>
<td>0.931</td>
<td>0.894</td>
<td>0.114</td>
</tr>
<tr>
<td>Trp457</td>
<td>Trp</td>
<td>0.933</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Ser458</td>
<td>Ser, (del)</td>
<td>0.925</td>
<td>0.911</td>
<td>0.924</td>
</tr>
<tr>
<td>Asn462</td>
<td>Asn, Val</td>
<td>0.933</td>
<td>—</td>
<td>—</td>
</tr>
</tbody>
</table>
Table 2. Tukey’s LSQ analysis on the amino acids that significantly interact with the glycon.
Numbering is based on *Zea mays* β-glucosidase

<table>
<thead>
<tr>
<th>Thr194</th>
<th>All, Thr</th>
<th>G H I</th>
<th>0.97</th>
</tr>
</thead>
<tbody>
<tr>
<td>Man, Thr</td>
<td>H I</td>
<td>0.95</td>
<td></td>
</tr>
<tr>
<td>Gal, Cys</td>
<td>E F G H</td>
<td>1.02</td>
<td></td>
</tr>
<tr>
<td>Glc, Cys</td>
<td>E F G H I</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>All, Cys</td>
<td>F G H I</td>
<td>0.98</td>
<td></td>
</tr>
<tr>
<td>Man, Cys</td>
<td>G H I</td>
<td>0.97</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Phe205</th>
<th>All, Asp</th>
<th>A</th>
<th>1.28</th>
</tr>
</thead>
<tbody>
<tr>
<td>Man, Asp</td>
<td>B C D E</td>
<td>1.08</td>
<td></td>
</tr>
<tr>
<td>Gal, Phe</td>
<td>C</td>
<td>1.08</td>
<td></td>
</tr>
<tr>
<td>Glc, Phe</td>
<td>C D</td>
<td>1.06</td>
<td></td>
</tr>
<tr>
<td>All, Phe</td>
<td>E F G</td>
<td>0.98</td>
<td></td>
</tr>
<tr>
<td>Man, Phe</td>
<td>E F G</td>
<td>0.97</td>
<td></td>
</tr>
<tr>
<td>Gal, His</td>
<td>C D E F</td>
<td>1.03</td>
<td>All, His</td>
</tr>
<tr>
<td>-----------</td>
<td>---------</td>
<td>------</td>
<td>-----------</td>
</tr>
<tr>
<td>Glc, His</td>
<td>D E F G</td>
<td>1.01</td>
<td>Gal, (del)</td>
</tr>
<tr>
<td>All, His</td>
<td>E F G</td>
<td>0.97</td>
<td>Glc, (del)</td>
</tr>
<tr>
<td>Man, His</td>
<td>E F G</td>
<td>0.97</td>
<td>All, (del)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Man, (del)</td>
</tr>
<tr>
<td>Glc, Leu</td>
<td>C D E F G</td>
<td>0.99</td>
<td>Glc, Cys</td>
</tr>
<tr>
<td>Gal, Leu</td>
<td>C D E F G</td>
<td>0.98</td>
<td>Gal, Cys</td>
</tr>
<tr>
<td>All, Leu</td>
<td>F G</td>
<td>0.93</td>
<td>All, Cys</td>
</tr>
<tr>
<td>Man, Leu</td>
<td>G</td>
<td>0.92</td>
<td>Man, Cys</td>
</tr>
</tbody>
</table>

**Asn285**

<table>
<thead>
<tr>
<th>Gal, Phe</th>
<th>A</th>
<th>1.28</th>
<th>All, Ser</th>
<th>D E F G H I J</th>
<th>0.97</th>
</tr>
</thead>
<tbody>
<tr>
<td>Glc, Phe</td>
<td>A</td>
<td>1.23</td>
<td>Man, Ser</td>
<td>E F G H I J</td>
<td>0.96</td>
</tr>
<tr>
<td>All, Phe</td>
<td>A B</td>
<td>1.21</td>
<td>Gal, Ser</td>
<td>E F G H I J</td>
<td>0.95</td>
</tr>
<tr>
<td>Man, Phe</td>
<td>B C D E F G</td>
<td>1.08</td>
<td>Glc, Ser</td>
<td>H</td>
<td>0.94</td>
</tr>
</tbody>
</table>

**Arg336**

<table>
<thead>
<tr>
<th>Gal, Gln</th>
<th>A</th>
<th>1.28</th>
<th>Gal, Tyr</th>
<th>A B C</th>
<th>1.18</th>
</tr>
</thead>
<tbody>
<tr>
<td>Glc, Gln</td>
<td>A</td>
<td>1.23</td>
<td>Glc, Tyr</td>
<td>B C D</td>
<td>1.10</td>
</tr>
<tr>
<td>All, Gln</td>
<td>A B</td>
<td>1.21</td>
<td>All, Tyr</td>
<td>B C D E F G H I J K</td>
<td>1.06</td>
</tr>
<tr>
<td>Man, Gln</td>
<td>B C D E F G</td>
<td>1.08</td>
<td>All, Arg</td>
<td>C D E F</td>
<td>1.06</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Glc, Arg</td>
<td>C D F</td>
<td>1.06</td>
</tr>
<tr>
<td></td>
<td>D</td>
<td>E</td>
<td>F</td>
<td>G</td>
<td>H</td>
</tr>
<tr>
<td>----------------</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>Man, Arg</td>
<td>D</td>
<td>E</td>
<td>F</td>
<td>G</td>
<td>H</td>
</tr>
<tr>
<td>All, Arg</td>
<td>D</td>
<td>E</td>
<td>F</td>
<td>G</td>
<td>H</td>
</tr>
<tr>
<td>Gal, Val</td>
<td>C</td>
<td>D</td>
<td>E</td>
<td>F</td>
<td>G</td>
</tr>
<tr>
<td>Glc, Val</td>
<td>D</td>
<td>E</td>
<td>F</td>
<td>G</td>
<td>H</td>
</tr>
<tr>
<td>Man, Val</td>
<td>D</td>
<td>E</td>
<td>F</td>
<td>G</td>
<td>H</td>
</tr>
<tr>
<td>All, Val</td>
<td>D</td>
<td>E</td>
<td>F</td>
<td>G</td>
<td>H</td>
</tr>
<tr>
<td>Gal, His</td>
<td>C</td>
<td>D</td>
<td>E</td>
<td>F</td>
<td>G</td>
</tr>
<tr>
<td>Glc, His</td>
<td>D</td>
<td>E</td>
<td>F</td>
<td>G</td>
<td>H</td>
</tr>
<tr>
<td>Man, His</td>
<td>E</td>
<td>F</td>
<td>G</td>
<td>H</td>
<td>I</td>
</tr>
<tr>
<td>All, His</td>
<td>E</td>
<td>F</td>
<td>G</td>
<td>H</td>
<td>I</td>
</tr>
<tr>
<td>Gal, Leu</td>
<td>D</td>
<td>E</td>
<td>F</td>
<td>G</td>
<td>H</td>
</tr>
<tr>
<td>Glc, Leu</td>
<td>C</td>
<td>D</td>
<td>E</td>
<td>F</td>
<td>G</td>
</tr>
<tr>
<td>All, Leu</td>
<td>H</td>
<td>I</td>
<td>J</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Man, Leu</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gal, Thr</td>
<td>D</td>
<td>E</td>
<td>F</td>
<td>G</td>
<td>H</td>
</tr>
<tr>
<td>Glc, Thr</td>
<td>D</td>
<td>E</td>
<td>F</td>
<td>G</td>
<td>H</td>
</tr>
<tr>
<td>All, Thr</td>
<td>G</td>
<td>H</td>
<td>I</td>
<td>J</td>
<td></td>
</tr>
<tr>
<td>Man, Thr</td>
<td>I</td>
<td>J</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gal, (del)</td>
<td>E</td>
<td>G</td>
<td>H</td>
<td>I</td>
<td>J</td>
</tr>
<tr>
<td>Glc, (del)</td>
<td>G</td>
<td>H</td>
<td>I</td>
<td>J</td>
<td>K</td>
</tr>
<tr>
<td>All, (del)</td>
<td>H</td>
<td>I</td>
<td>J</td>
<td>K</td>
<td></td>
</tr>
<tr>
<td>Man, (del)</td>
<td>J</td>
<td>K</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Asn376

<table>
<thead>
<tr>
<th></th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
<th>I</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gal, Glu</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>A</td>
<td>1.28</td>
</tr>
<tr>
<td>Glc, Glu</td>
<td>A</td>
<td>B</td>
<td></td>
<td></td>
<td></td>
<td>1.23</td>
</tr>
<tr>
<td>All, Glu</td>
<td>A</td>
<td>B</td>
<td>C</td>
<td></td>
<td></td>
<td>1.21</td>
</tr>
<tr>
<td>Man, Glu</td>
<td>C</td>
<td>D</td>
<td>E</td>
<td>F</td>
<td>G</td>
<td>1.08</td>
</tr>
<tr>
<td>Gal, Ser</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>A</td>
<td>1.18</td>
</tr>
<tr>
<td>Glc, Ser</td>
<td>B</td>
<td>C</td>
<td>D</td>
<td>E</td>
<td></td>
<td>1.10</td>
</tr>
<tr>
<td>All, Ser</td>
<td>C</td>
<td>D</td>
<td>E</td>
<td>F</td>
<td>G</td>
<td>H</td>
</tr>
<tr>
<td>Man, Ser</td>
<td>C</td>
<td>D</td>
<td>E</td>
<td>F</td>
<td>G</td>
<td>H</td>
</tr>
<tr>
<td>Gal, Phe</td>
<td></td>
<td>D</td>
<td>E</td>
<td>F</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Glc, Phe</td>
<td></td>
<td>E</td>
<td>F</td>
<td>G</td>
<td>H</td>
<td>I</td>
</tr>
<tr>
<td>All, Phe</td>
<td>F</td>
<td>G</td>
<td>H</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Man, Phe</td>
<td></td>
<td>H</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gal, Met</td>
<td></td>
<td></td>
<td></td>
<td>E</td>
<td>I</td>
<td>1.04</td>
</tr>
<tr>
<td>Glc, Met</td>
<td>E</td>
<td>F</td>
<td>G</td>
<td>H</td>
<td></td>
<td></td>
</tr>
<tr>
<td>All, Met</td>
<td>F</td>
<td>G</td>
<td>H</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Man, Met</td>
<td>F</td>
<td>G</td>
<td>H</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Glc, Asn</td>
<td></td>
<td>E</td>
<td>F</td>
<td></td>
<td>I</td>
<td></td>
</tr>
<tr>
<td>Gal, Asn</td>
<td>E</td>
<td>F</td>
<td>G</td>
<td>H</td>
<td></td>
<td></td>
</tr>
<tr>
<td>All, Asn</td>
<td>F</td>
<td>G</td>
<td>H</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Man, Asn</td>
<td>G</td>
<td>H</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>All, Ile</td>
<td>E</td>
<td>F</td>
<td>G</td>
<td>H</td>
<td>I</td>
<td></td>
</tr>
<tr>
<td>Man, Ile</td>
<td>E</td>
<td>F</td>
<td>G</td>
<td>H</td>
<td>I</td>
<td></td>
</tr>
<tr>
<td>Gal, Ile</td>
<td>E</td>
<td>F</td>
<td>G</td>
<td>H</td>
<td>I</td>
<td></td>
</tr>
<tr>
<td>Glc, Ile</td>
<td>F</td>
<td>G</td>
<td>H</td>
<td>I</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Trp378</td>
<td></td>
<td>Trp465</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>----------</td>
<td>--------</td>
<td>----------</td>
<td>--------</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gal, Ile</td>
<td>A</td>
<td>1.28</td>
<td>Gal, Phe</td>
<td>A</td>
<td>1.28</td>
<td></td>
</tr>
<tr>
<td>Glc, Ile</td>
<td>A</td>
<td>1.23</td>
<td>Glc, Phe</td>
<td>A</td>
<td>1.23</td>
<td></td>
</tr>
<tr>
<td>All, Ile</td>
<td>A B</td>
<td>1.21</td>
<td>All, Phe</td>
<td>A B</td>
<td>1.21</td>
<td></td>
</tr>
<tr>
<td>Man, Ile</td>
<td>B C D</td>
<td>1.08</td>
<td>Man, Phe</td>
<td>B C D</td>
<td>1.08</td>
<td></td>
</tr>
<tr>
<td>Gal, Trp</td>
<td>C</td>
<td>1.05</td>
<td>Gal, Trp</td>
<td>C</td>
<td>1.05</td>
<td></td>
</tr>
<tr>
<td>Glc, Trp</td>
<td>C</td>
<td>1.03</td>
<td>Glc, Trp</td>
<td>C</td>
<td>1.03</td>
<td></td>
</tr>
<tr>
<td>All, Trp</td>
<td>D E</td>
<td>0.97</td>
<td>All, Trp</td>
<td>D E</td>
<td>0.97</td>
<td></td>
</tr>
<tr>
<td>Man, Trp</td>
<td>E</td>
<td>0.96</td>
<td>Man, Trp</td>
<td>E</td>
<td>0.96</td>
<td></td>
</tr>
</tbody>
</table>