PROGRESS IN PREDICTING PROTEIN FUNCTION FROM STRUCTURE: UNIQUE FEATURES OF O-GLYCOSIDASES

E.W. STAWISKI*, Y. MANDEL-GUTFREUND*, A.C. LOWENTHAL, L. M. GREGORET
Departments of Chemistry and Biochemistry and Molecular, Cell and Developmental Biology
University of California
Santa Cruz, CA 95060

The Structural Genomics Initiative promises to deliver between 10,000 and 20,000 new protein structures within the next ten years. One challenge will be to predict the functions of these proteins from their structures. Since the newly solved structures will be enriched in proteins with little sequence identity to those whose structures are known, new methods for predicting function will be required. Here we describe the unique structural characteristics of O-glycosidases, enzymes which hydrolyze O-glycosidic bonds between carbohydrates. O-glycosidase function has evolved independently many times and enzymes that carry out this function are represented by a large number of different folds. We show that O-glycosidases none-the-less have characteristic structural features that cross sequence and fold families. The electrostatic surfaces of this class of enzymes are particularly distinctive. We also demonstrate that accurate prediction of O-glycosidase function from structure is possible.

1 Introduction

1.1 Structural genomics

The completion of the sequencing of entire genomes has tremendously increased our knowledge of biology. It has also revolutionized our thinking about the scale at which is possible to make inquiries and has created new challenges. One of these challenges is the determination of the three dimensional structures of a large, representative set of proteins. This endeavor, called *structural genomics*, aims to 1) understand the basis of disease at the atomic resolution level of detail, 2) provide homology modeling scaffolds for all proteins, 3) obtain structures for the targets of structure-based drug discovery, and 4) map out protein fold space to better understand sequence-structure relationships[1]. Within the next ten years, between 10,000 and 20,000 new structures are expected to be solved as a result of the structural genomics initiative[2].

It is well established that the proteins that are similar in sequence are likely to have evolved from a common ancestor and thus retain similar functions. However, in order to sample protein structure space broadly, the structures selected as targets for structural genomics will primarily be those with little or no sequence identity to

*These authors contributed equally to this work

proteins whose structures have already been solved. Many of these proteins are expected to have unknown functions.

1.2 Predicting function from structure

In the absence of sequence identity, how is it possible to predict function from structure? It has been noted that proteins that adopt similar folds often have similar functions[3]. However, there are many well known examples, such as the SH3 fold and the TIM barrel[4], which are used for many different functions. It is also often possible to predict the functions of enzymes based on the spatial arrangement of catalytic residues[5]. However, some functions (e.g. proteolysis) have evolved multiple times and can not be accounted for by any single set of residues.

To account for the likelihood that entirely new folds and catalytic mechanisms will be discovered through structural genomics, it will be necessary to develop new methods of function prediction that do not rely on existing examples of known sequences, folds or catalytic residue arrangements. We have demonstrated recently that proteases as a group (including those with very different folds and catalytic mechanisms) share common structural features[6]. These features include smaller surface areas, higher packing densities, and less helical structure. We speculate that these features arose independently as mechanisms to avoid autolysis. More importantly, we are able to use these identifying features to train a neural network to predict protease function with high accuracy even in the absence of related structures being present in the training set[6]. More recently, we have tackled the problem of nucleic acid binding protein function prediction[7]. In this case, discrimination of the nucleic acid binding proteins (again at high accuracy) relied on a novel method using positive electrostatic patch analysis.

1.3 The O-glycosidases

We now turn our attention to identifying the characteristic features of O-glycosidases. These enzymes hydrolyze linkages between carbohydrate molecules. We chose this class of enzymes for several reasons. First, oligosaccharides play critical roles in a variety of biological processes including viral invasion and cell signaling events[8]. Thus O-glycosidases represent an important class of enzymes for drug discovery, especially with regard to antiviral and anticancer agents. Second, the O-glycosidases are structurally diverse and include members of at least six different folds ranging from all alpha to all beta[9, 10]. This enzyme family is therefore an excellent test case for fold independent function prediction methods. O-glycosidases are also well represented in the existing Protein Data Bank[11], allowing us to build a substantial representative data set. Finally, O-glycosidases were frequent false positives in our protease prediction effort[6]. We speculated that this is because they are also hydrolases, and may have been subjected to similar

anti-autolysis evolutionary pressures. Identification of the unique features of O-glycosidases should improve prediction on both classes of enzymes.

We have gleaned a set of identifying features of O-glycosidases. Like with nucleic acid binding proteins, we again used electrostatic patch analysis, this time concentrating on the negatively-charged surface patches to help characterize O-glycosidases. Using an ensemble of features, we were able to train a neural network to predict O-glycosidase function with over 87% accuracy. We are also now able to better discriminate between O-glycosidases and proteases.

2 Methods

2.1 Data Set Construction

Representative data sets of both O-glycosidase and non-O-glycosidase protein structures from the Protein Data Bank (PDB)[11] were constructed. These data sets consisted of proteins that were solved by X-ray crystallography and had an atomic resolution of better than 2.5 Å. Within the sets, sequence identity cutoffs were used such that no two members in any one data set contained more than 25% sequence identity within the non-O-glycosidase data set and 35% sequence identity for the O-glycosidases. The O-glycosidase set was constructed by mining the PDB for the Enzyme Commission (EC) number 3.2.1.x (x stands for substrate specificity) and consisted of 39 proteins. The non-O-glycosidase data set consisted of 258 monomeric proteins and was constructed from Hobohm's and Sander's "pdb select" list of proteins[12]. A description of the glycosidase data set along with the PDB codes for the non-glycosidase data set can be found at : http://www.chemistry.ucsc.edu/gregoret/PSB_supp.html.

2.2 Electrostatic Patch Analysis

The UHBD[13] program was used to derive a continuum electrostatic description for each protein, using the Poisson-Boltzmann equation. For all UHBD calculations the grid dimensions were set to 65x65x65 with 2.0 Å distance from each grid point to another. Dielectric constants of 2.0 and 80.0 was used for the protein and the solvent, respectively. Other parameters for UHBD were set to their default values. Patches were constructed based on UHBD output with an in-house program, PATCHFINDER. Continuum negative electrostatic patches, mapped to the protein surface, were constructed by assembling adjacent surface points, which possessed a negative potential less than or equal to -5 *kT*. The patch size was defined as the number of surface points within a continuous cluster of points. The largest negative patch was used to extract sequence and structural information.

2.3 Sequence Conservation

For each protein in the data set, a multiple sequence alignment (MSA) was constructed using PSI-BLAST version 2.1.1[14] to search the non-redundant NCBI data base for similar sequences that were significant (E-value <0.001). Since we wanted to include only sequences that are likely to be structurally related to the representative sequence we eliminated sequences with <35% identity. In addition, to reduce redundancy from very close homologous sequences, only sequences with <90% identity were included in the MSAs.

The conservation of specific residues within the negative electrostatic patch was analyzed, including residues which were occupied by Glu, Asp and Asn. In addition to the simple amino acid conservation we also calculated the conservation of aromatic residues as a group. A residue position was considered to be conserved when ≥75% of the sequences in the MSA contained the same amino acid (for amino acid conservation) or property (for property conservation) as in the representative sequence. For each of the amino acids above, the normalized frequency of conserved residues in the electrostatic patch was calculated.

2.4 Cleft Detection

The program SURFNET[15] was used to analyze protein clefts. The residues identified within the largest two clefts of each protein were examined and the number of residues that overlapped with the largest negative patch were calculated. The cleft identified as having the largest overlap with the patch was further examined to confirm whether it had potential residues that could participate in a glycolysis reaction. Only Asp and Glu were considered for donation of the acid or base atoms for the reaction. All distances between carboxylate groups of ASP and GLU residues were calculated. The two carboxylate atoms with a distance closest to either 5 or 9.5 Å of each other, were identified as possible catalytic residues.

2.5 Calculation of Other Structural Features

A protein's solvent accessible surface area was calculated by using Lee and Richards method[16] as implemented in the program CALC-SURFACE[17] with a default probe radius of 1.4 Å. The program DMS under the UCSF MidasPlus software package[18] was used to calculate molecular surface. The roughness, or fractal dimension, D, of the surface[19] was calculated using equation 1:

1. $$D = 2 - \frac{d\log A_S}{d\log R}$$

where R is the probe radius and A_S is the molecular surface area. In this case, radii of 1.25,1.5, 1.75 and 2.0 Å were used. A perfectly smooth surface will not depend on the probe size, and will thus have a fractal dimension of two.

2.6 Machine learning

For function prediction, we applied the Nevprop4 neural network package[20]. The neural network consisted of a single hidden layer with 3 nodes and a single output node. All training was performed with a standard feedforward, error backpropagation algorithm. The cross validation scheme used was to train on all but one member of the data set, which was withheld from training and subsequently tested. This was done for each of the members of the O-glycosidase data set. For the non-glycosidase set, in each training session a random sample of 10% of the data set was withheld and subsequently tested. For the false set the average performance over all runs was reported.

3 Results and Discussion

3.1 Identifying Characteristics of O-glycosidases

Based on general properties of the proteins structures, we wanted to identify unique features that could be used to distinguish O-glycosidase proteins from other classes of proteins. To do this, two representative data sets of crystallographically-determined three-dimensional protein structures were constructed as described in Methods. The O-glycosidase data set consisted of 39 proteins and the non-O-glycosidase proteins (including a full spectrum of different proteins except those with glycosidase function) had 258 members. A list of the proteins in the glycosidase data set and their structural classification (based on SCOP classification) is given at: http://www.chemistry.ucsc.edu/gregoret/PSB_supp.html. For the two data sets we calculated both global and local structural features that could potentially distinguish between them. Although most of the features analyzed did not show statistically significant differences between the O-glycosidases and other group of proteins, when combined, these inputs were successfully used in predicting glycosidase function.

3.1.1 Surface features

We have previously shown that electrostatic patch analysis, a combination of structural and sequence features extracted from a distinct region on the protein surface defined by electrostatic potential, can help to distinguish nucleic acid binding proteins from other proteins[7]. Surface electrostatics have previously been used to help indicate potential protein functions based on their structure[21]. O-glycosidases mostly use Glu and Asp residues for catalysis, and as a result usually have a negatively-charged surface associated with their active sites[9]. To characterize

the O-glycosidase protein family, we performed our analysis on negatively-charged electrostatic patches on the protein surface (similar to the analysis performed for nucleic acid binding proteins[7]). As a first step the continuum electrostatic potential was calculated for the whole protein using UHBD[13] software package, negative surface electrostatic patches were then assembled using an in house program PATCHFINDER (as described in methods) and the largest negative patch of every protein structure was analyzed.

O-glycosidases were found to have, on average, larger negative surface patches than non-O-glycosidases. The average size of the negative patch is 229 ± 163 surface points (compared to 86 ± 110 surface points for the non-O-glycosidases). In O-glycosidase enzymatic hydrolysis, one negative residue typically acts as a base and another as an acid[22]. We found that in 33 out of 39 proteins (85%), the largest negative patch contains at least one of the two residues (base or acid), which are known to be involved in the O-glycosidic reaction. In some cases where the patch did not contain either the acid or the base, these patches were found to be in close proximity to other functionally important residues. The high overlap between the surface patch and the O-glycosidases active site allows us to conclude that the region we are analyzing is directly involved in the function of these proteins.

As shown in Figure 1, the amino acid distribution in the negatively-charged patches of O-glycosidase proteins also differs from that of other proteins. The differences observed are mostly found in the aromatic amino acids Trp and Tyr and in Asn. All three amino acids show higher frequency in the O-glycosidase proteins than in all other proteins. Aromatic amino acids are known to commonly act as docking sites for the non-polar inner portion of cyclic carbohydrate molecules[23]**.** Since our negative patches are derived from Asp and Glu residues on the protein surface, the normalized percentage of these two amino acids was roughly the same in all proteins.

Because of the high correlation between the active site and the negative patch, we also expected the residues within the patch to be functionally important and thus more conserved[7, 24]. To examine that we analyzed the frequency of conserved residues within the negative patch. Specifically we were interested in the conservation of Asn, Glu, Asp, Tyr and Trp, which were highly frequent in the O-glycosidase negatively-charged patches. In addition to analyzing the conservation of the individual residues we also looked at the conservation of the aromatic amino acids when grouped together. As summarized in Table 1, we found that Asp, Glu, Asn, Tyr and Trp are on average more conserved (though not significant) within the negative patches of the O-glycosidase family than in the non-O-glycosidases. These differences were more obvious when we grouped the aromatic residues together. O-glycosidases have on average more conserved aromatics (6.9±4.6 residues) than do the non-O-glycosidases (1.3±2.0 residues).

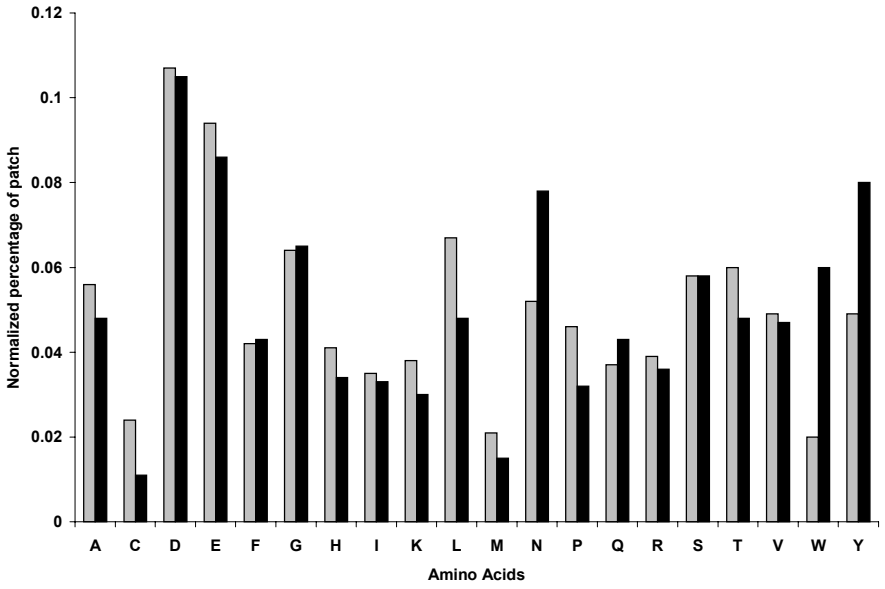

Figure 1. The normalized frequency of the 20 amino acids within the largest negative patch for O-glycosidases (black) and non-O-glycosidases (gray).

Table 1. Frequency* and conservation of specific amino acids in the negative surface patches of O-glycosidase and non-O-glycosidase proteins

	O-glycosidase		Non-O-glycosidase	
	Total number of residues in patch	**Number of conserved residues in patch**	**Total number of residues in patch**	**Number of conserved residues in patch**
GLU	**4.5 (2.9)**	**2.3 (1.6)**	**2.1 (2.5)**	**0.7 (1.2)**
ASP	**4.8 (3.9)**	**2.5 (2.3)**	**2.2 (3.1)**	**1.0 (1.5)**
ASN	**3.4 (2.0)**	**1.7 (1.3)**	**1.1 (1.9)**	**0.4 (0.9)**
TRP	**3.4 (2.0)**	**1.7 (1.1)**	**1.1 (1.8)**	**0.4 (0.9)**
TYR	**3.7 (2.8)**	**2.0 (1.5)**	**1.1 (1.8)**	**0.5 (0.9)**

*Frequency is averaged over all proteins, the average number and standard deviation are shown.

3.1.2 Structural features

Laskowski et al., have previously shown that amongst enzymes, a protein's largest and second largest clefts are bound to the ligand 84% and 9% of the time, respectively[25]. To see if information on protein clefts could help us to further discriminate O-glycosidases, we first identified the clefts belonging to each protein in both our datasets, using the program SURFNET[15]. For each protein structure, we identified the two largest clefts. We then extracted the residues associated with each cleft, and calculated how many of the residues within the cleft were also in the largest negative patch of the protein. Similar to the correlation between the cleft and the active site found by Laskowski et al., we found that 82% of O-glycosidases proteins contained the negative patch in their largest cleft, and 10% of the proteins showed an overlap between the patch and the second largest cleft. The overlap between the patch and the largest cleft is much less frequent in non-O-glycosidases (58% showed overlap between the patch and the largest cleft and 21% showed overlap between the patch and the second largest cleft).

It is known that amongst O-glycosidases, there are two types of possible reactions, an inversion or retention reaction[22]. The difference between the reactions is whether they invert or retain the anomeric configuration of the sugar at the cleavage site. The acid and base residues involved with the enzymatic reaction, have been found to be predominantly Glu and Asp residues, although exceptions involving His and Ser have been found[26]. For a retention reaction, the average spacing between the two participating carboxyl groups is 5.5 Å, and for an inversion reaction, the spacing is 9.5 Å. Although there are exceptions to these reaction geometry distances[27], they serve as a general guideline for glycosidic hydrolysis.

We examined the cleft that had the highest patch overlap and identified the candidate residues that could be involved in a glycosidic reaction (see Methods). In general, the O-glycosidases have only slightly "better" putative catalytic residues (determined by the smallest difference in distance between two negative residues in either 5 or 9.5 Å) within their clefts (90% less than 0.24 Å) than other classes of proteins (59% less than 0.24 Å). The relatively large number of potential binding residues in non-O-glycosidase proteins are most likely due to the high frequency of Asp and Glu in the negative patches.

Lewis and Rees originally proposed that roughness may be associated with ligand binding, since a greater surface area allows more possibilities for van der Waals contacts[19]. Indeed, it has been found that functional sites in proteins (e.g. enzyme active sites) are rougher than other parts of the protein's surface[28]. To see if roughness could further discriminate the negative patches within our data set, we calculated the fractal dimension (D) for each protein patch. Although not significant, the O-glycosidase patches had slightly rougher surfaces (2.67 ± 0.25) than the non-O-glycosidase patches (2.52 ± 0.41). Interestingly the 6 O-glycosidases whose catalytic residues were not in the patch showed a lower degree

of a roughness 2.46 ± 0.05 which is comparable to the roughness of the non-O-glycosidases. This again strengthens the idea that the patches are associated with an O-glycosidase active site.

We have previously shown that other hydrolases (the proteases), have less accessible surface area per molecular weight as compared to other types of proteins[6]. Since a large number of the false positives in this experiment were O-glycosidases, we examined the accessible surface area per molecular weight amongst our glycosidase data set. Interestingly, O-glycosidases were also found to have strikingly less accessible surface area per molecular weight than non-O-glycosidases. As shown in Fig. 2, nearly 84% of the O-glycosidase data set falls below the line that represents the best fit for the non-O-glycosidase proteins. Moreover, the majority of the O-glycosidase proteins fall at the lower limit for accessible surface area per molecular weight. For proteases, we speculated that the lower solvent accessible surface area per molecular weight may have evolved to prevent self-cleavage. It has previously been proposed that different types hydrolases have evolved from a common ancestor[29]. O-glycans are also known to inhibit proteolysis[30]. O-glycosidases may therefore have to work in coordination with proteases to remove O-glycans from proteins in order to make them accessible to protein degradation. We again speculate that O-glycosidases may be under similar types of evolutionary pressure and hence show similar structural properties to proteases.

3.2 Prediction of O-Glycosidases vs. Non-O-Glycosidases

Although most of the features characterized above were not individually sufficient to discriminate O-glycosidases from other non-O-glycosidase proteins, we wanted to see if we could use them collectively to infer function. To do this we created a feature vector consisting of the following 10 inputs: surface area per molecular weight, negative patch size, patch-cleft overlap booleans (1 for true) for the largest and second largest cleft, sequence conservation within the patch (aromatics, Asn, Glu, and Asp), roughness of patch, and the best distance for a putative active site. We used this feature vector to train a neural network (see Methods), to distinguish O-glycosidase proteins from non-O-glycosidase proteins.

Using the cross validation scheme as described in Methods, we were able to predict O-glycosidase proteins with 87% accuracy and non-O-glycosidase proteins with 93% accuracy. When we examined the relevance of the inputs, defined as the sum of the square weights for a given input group, divided by the sum of all input groups, the most relevant inputs were the frequency of conserved aromatics, the SA/MW, patch size, conservation of Glu, and geometry of the putative active site.

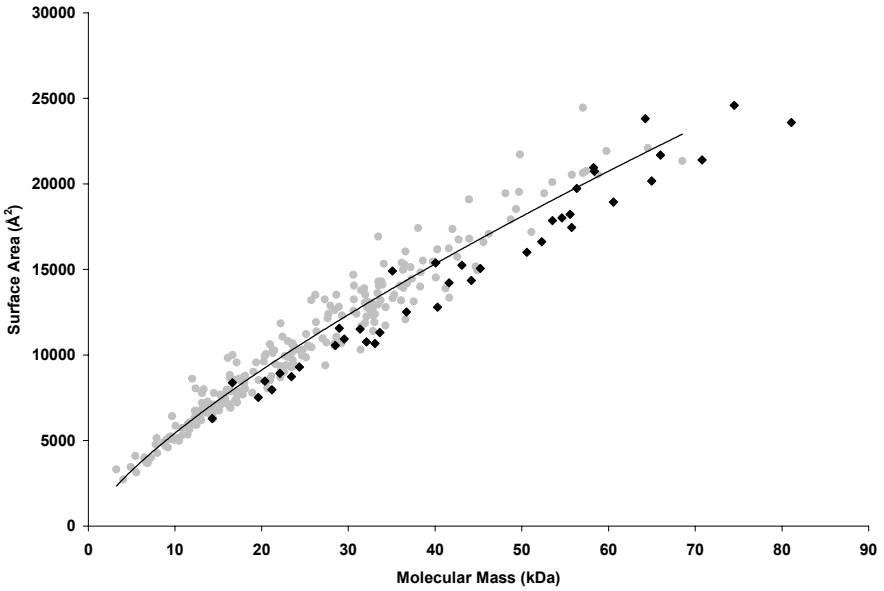

Figure 2. Molecular mass vs. surface area for O-glycosidases (♦) and non-O-glycosidases (●). The solid line is the best fit for the non-O-glycosidases. The average surface area to molecular mass ratio for O-glycosidases is 0.36 ± 0.04 and 0.46 ± 0.08 for the non-O-glycosidases.

3.3 Discriminating Glycosidases and Proteases

Since we found some similar structural properties to exist between proteases and O-glycosidases, we were interested to see if we are now able to distinguish between these two classes of proteins. From the output of the neural net discussed above, we were able to correctly identify over 80% of the proteases as being non-O-glycosidases. The biggest difference between the two data sets arises from the electrostatic patch analysis (data not shown). One feature that differs between the two groups, that was not addressed earlier, is the presence of the TIM fold. The TIM barrel is a common fold amongst O-glycosidases, but is absent amongst proteins with protease function. This could further help to differentiate between the two. We predict that similar approaches could be used for distinguishing between other protein classes which possess different functions, yet retain some similar structural characteristics.

4 Conclusions

We have shown that O-glycosidases possess some unique global and local properties that can be distinguished from other proteins with different functions. Like proteases, O-glycosidases have less accessible surface area per unit molecular weight. We speculate that as for the proteases, this structural property can be involved in preventing proteolysis. Beyond this we have shown that O-glycosidases have characteristic electrostatic features. Most O-glycosidases have large negative patches on their surface that are highly correlated with active sites. This could help to identify putative active sites, which could be targeted for molecular based drug design. We have also shown for the second time that electrostatic patch analysis, using a combination of structural and sequence analysis, can be used to characterize a class of proteins. We believe that electrostatic patch analysis should therefore be a common tool in the characterization of the relationship between a protein's structure and its function.

Using this combination of global and region-specific features, we were able to successfully train a neural network to predict O-glycosidase function with high accuracy. Since the O-glycosidase data set used in this study was comprised of many different folds and had little sequence homology amongst them, we propose that the neural net would potentially be able to identify a protein having an O-glycosidase function even if it possessed a novel fold. This could be particularly useful for products of the structural genomic initiative. Here, for the first time, we show that prediction of protein function from structure can be improved by the ability to identify a class of proteins that was a common false positive prediction of another protein family. We suggest that protein function prediction, may be improved by analyzing common themes amongst classes of proteins that are incorrectly identified. Multi-class automated protein function prediction based on structure, may therefore be close at hand.

5 Acknowledgements

We would like to thank Janet Newman for helpful discussions. This work was supported by NIH grant GM52885 to LMG, an award from the University of California Cancer Research Coordinating Committee to LMG, and an American Cancer Society California Division postdoctoral fellowship to YMG.

References

1. A. Sali, *Nat Struct Biol* **5**, 1029 (1998); C. A. Orengo, A. E. Todd, J. M. Thornton, *Curr Opin Struct Biol* **9**, 374 (1999).
2. D. Vitkup, E. Melamud, J. Moult, C. Sander, *Nat Struct Biol* **8**, 559 (2001).
3. W. A. Koppensteiner, P. Lackner, M. Wiederstein, M. J. Sippl, *J Mol Biol* **296**, 1139 (2000).
4. J. A. Gerlt, P. C. Babbitt, *Annu Rev Biochem* **70**, 209 (2001).
5. A. C. Wallace, N. Borkakoti, J. M. Thornton, *Protein Sci* **6**, 2308 (1997); J. S. Fetrow, J. Skolnick, *J Mol Biol* **281**, 949 (1998).
6. E. W. Stawiski, A. E. Baucom, S. C. Lohr, L. M. Gregoret, *PNAS*, **97**, 3954 (2000).
7. E. W. Stawiski, Y. Mandel-Gutfreund, L. M. Gregoret, *Submitted*, (2001).
8. K. von Figura, *Curr Opin Cell Biol* **3**, 642 (1991); R. C. Wade, *Structure* **5**, 1139 (1997).
9. G. Davies, B. Henrissat, *Structure* **3**, 853 (1995).
10. H. Hegyi, M. Gerstein, *J Mol Biol* **288**, 147 (1999).
11. H. M. Berman *et al.*, *Nucleic Acids Res* **28**, 235 (2000).
12. U. Hobohm, C. Sander, *Protein Sci* **3**, 522 (1994).
13. M. E. Davis, Mandura, J.D., Luty, B.A., McCammon, J.A.,, *Comp. Phys. Comm.*,, 187 (1991).
14. S. F. Altschul *et al.*, *Nucleic Acids Res* **25**, 3389 (1997).
15. R. A. Laskowski, *J Mol Graph* **13**, 323 (1995).
16. B. Lee, F. M. Richards, *J Mol Biol* **55**, 379 (1971).
17. M. Gerstein, *Acta Crystallogr A* **48**, 271 (1992).
18. T. E. Ferrin, C. C. Huang, L. E. Jarvis, R. Langridge, *J Mol Graph* **6**, 13 (1988).
19. M. Lewis, D. C. Rees, *Science* **230**, 1163 (1985).
20. P. Goodman. (University of Nevada, Renvo, NV, 1996).
21. T. J. Boggon, W. S. Shan, S. Santagata, S. C. Myers, L. Shapiro, *Science* **286**, 2119 (1999); B. Honig, A. Nicholls, *Science* **268**, 1144 (1995).
22. M. L. Sinnott, *Chem Rev* **90**, 1171 (1990).
23. T. N. Petersen, S. Kauppinen, S. Larsen, *Structure* **5**, 533 (1997).
24. R. Landgraf, I. Xenarios, D. Eisenberg, *J Mol Biol* **307**, 1487 (2001).
25. R. A. Laskowski, N. M. Luscombe, M. B. Swindells, J. M. Thornton, *Protein Sci* **5**, 2438 (1996).
26. B. Henrissat, G. Davies, A. Bairoch. *http://afmb.cnrs-mrs.fr/~pedro/CAZY/ghf.html* (2001).
27. R. Pickersgill, D. Smith, K. Worboys, J. Jenkins, *J Biol Chem* **273**, 24660 (1998).
28. F. K. Pettit, J. U. Bowie, *J Mol Biol* **285**, 1377 (1999).
29. D. L. Ollis *et al.*, *Protein Eng* **5**, 197 (1992).
30. B. Garner *et al.*, *J Biol Chem* **276**, 22200 (2001).