PFIT and PFRIT: Bioinformatic algorithms for detecting glycosidase function from structure and sequence

GARY KLEIGER,[1] EKATERINA M. PANINA,[1] PARAG MALLICK,[1,2] AND DAVID EISENBERG[1]

[1]Howard Hughes Medical Institute, University of California, Los Angeles-Department Of Energy (UCLA-DOE) Institute of Genomics and Proteomics, UCLA, Los Angeles, California 90095, USA

(RECEIVED June 20, 2003; FINAL REVISION August 28, 2003; ACCEPTED September 17, 2003)

Abstract

The identification of the enzymes involved in the metabolism of simple and complex carbohydrates presents one bioinformatic challenge in the post-genomic era. Here, we present the PFIT and PFRIT algorithms for identifying those proteins adopting the α/β barrel fold that function as glycosidases. These algorithms are based on the observation that proteins adopting the α/β barrel fold share positions in their tertiary structures having equivalent sets of atomic interactions. These are conserved tertiary interaction positions, which have been implicated in both structure and function. Glycosidases adopting the α/β barrel fold share more conserved tertiary interactions than α/β barrel proteins having other functions. The enrichment pattern of conserved tertiary interactions in the glycosidases is the information that PFIT and PFRIT use to predict whether any given α/β barrel will function as a glycosidase or not. Using as a test set a database of 19 glycosidase and 45 nonglycosidase α/β barrel proteins with low sequence similarity, PFIT and PFRIT can correctly predict glycosidase function for 84% of the proteins known to function as glycosidases. PFIT and PFRIT incorrectly predict glycosidase function for 25% of the nonglycosidases. The program PSI-BLAST can also correctly identify 84% of the 19 glycosidases, however, it incorrectly predicts glycosidase function for 50% of the nonglycosidases (twofold greater than PFIT and PFRIT). Overall, we demonstrate that the structure-based PFIT and PFRIT algorithms are both more selective and sensitive for predicting glycosidase function than the sequence-based PSI-BLAST algorithm.

Keywords: glycosidase; tertiary interaction; bioinformatics; structure; α/β barrel; fold

Protein sequence and structural data are being generated by genomic sequencing and structural genomics projects at such a tremendous rate that immediate biochemical characterization of the functions of these proteins is impossible (Kanehisa and Bork 2003). Therefore, one goal of functional genomics is to identify the function of a newly identified protein through computational methods (Eisenberg et al. 2000).

Sequence or structural similarity between two proteins is evidence that they share related functions; however, use of homology-based methods often yields ambiguous or negative results. For example, only a minor fraction (20%–30%) of proteins identified from genomic sequencing projects share significant sequence similarity with proteins of known function (Mallick et al. 2000). In addition, many folds have proven highly adaptable at accommodating several different functions, such that two proteins sharing the same fold may have different functions. For example, at least 64 different enzymatic functions have been recorded for enzymes adopt-

Reprint requests to: David Eisenberg, Howard Hughes Medical Institute, UCLA-DOE Institute of Genomics and Proteomics, UCLA, Box 951570, Los Angeles, CA 90095, USA; e-mail: david@mbi.ucla.edu; fax: (310) 206-3914.

[2]Present address: Institute for Systems Biology, Seattle, WA 98703, USA

Abbreviations: PFIT, prediction of function through tertiary interactions; PFRIT, prediction of function through residues at tertiary interactions; 3D, three-dimensional; EC, enzyme commission; ROC, receiver-operator characteristic; BPI, bactericidal/permeability-increasing protein.

Article and publication are at http://www.proteinscience.org/cgi/doi/10.1110/ps.03274104.

ing the α/β barrel fold (Nagano et al. 2002). Therefore, our goal is to find properties that are conserved among functionally related proteins, and to use this property to assign function to newly identified proteins.

Tertiary atomic interactions, in the form of a hydrogen bond or a salt-bridge between a pair of residues, are often conserved at structurally equivalent positions in pairs of proteins adopting the same fold (Browne et al. 1969). These conserved tertiary interactions have been shown to be important for both protein structure and function (Kleiger et al. 2000, 2001). Conserved tertiary interactions were first identified in the two similarly folded domains of the human bactericidal/permeability-increasing protein (BPI; Kleiger et al. 2000), which share only 13% sequence identity (Beamer et al. 1997). Four conserved tertiary interactions were found between the two domains of BPI and are thought to be important determinants of the BPI-domain fold.

Another example of a conserved tertiary interaction was discovered in the x-ray structures of the E1β subunits from human, *Pseudomonas putida*, and *Pyrobaculum aerophilum* α-keto acid dehydrogenase complexes (Kleiger et al. 2001). This conserved tertiary interaction is found in all three E1β structures and is important for proper dehydrogenase assembly. Furthermore, a multiple sequence alignment of 19 E1β protein sequences shows that the residues predicted to occupy structurally equivalent positions at the conserved tertiary interaction are themselves conserved. The existence of conserved tertiary interactions in both BPI and E1β led us to the following two questions: could we find conserved tertiary interactions in a fold with many representatives in the protein databank, and could we use this information to predict protein function?

The α/β barrel, first observed in the x-ray structure of the enzyme triosephosphate isomerase, is an ideal fold for an-

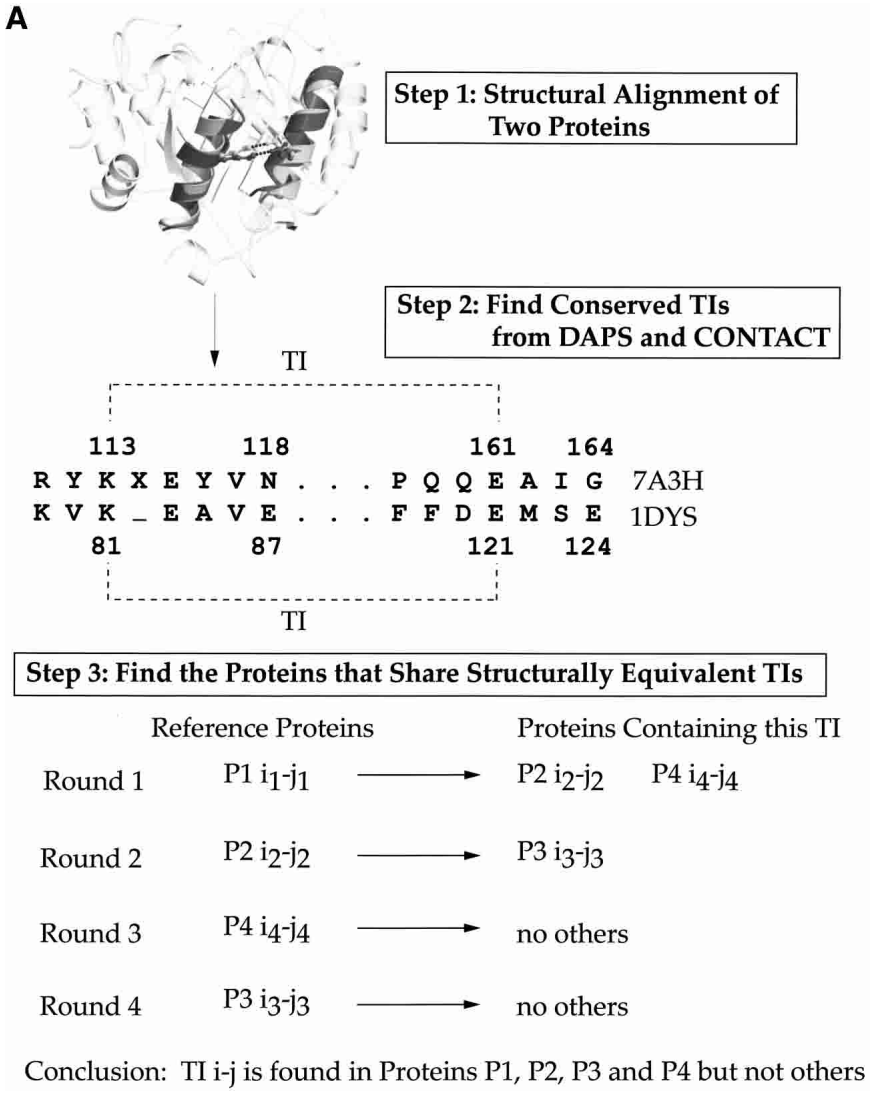

Figure 1. (Continued on next page)

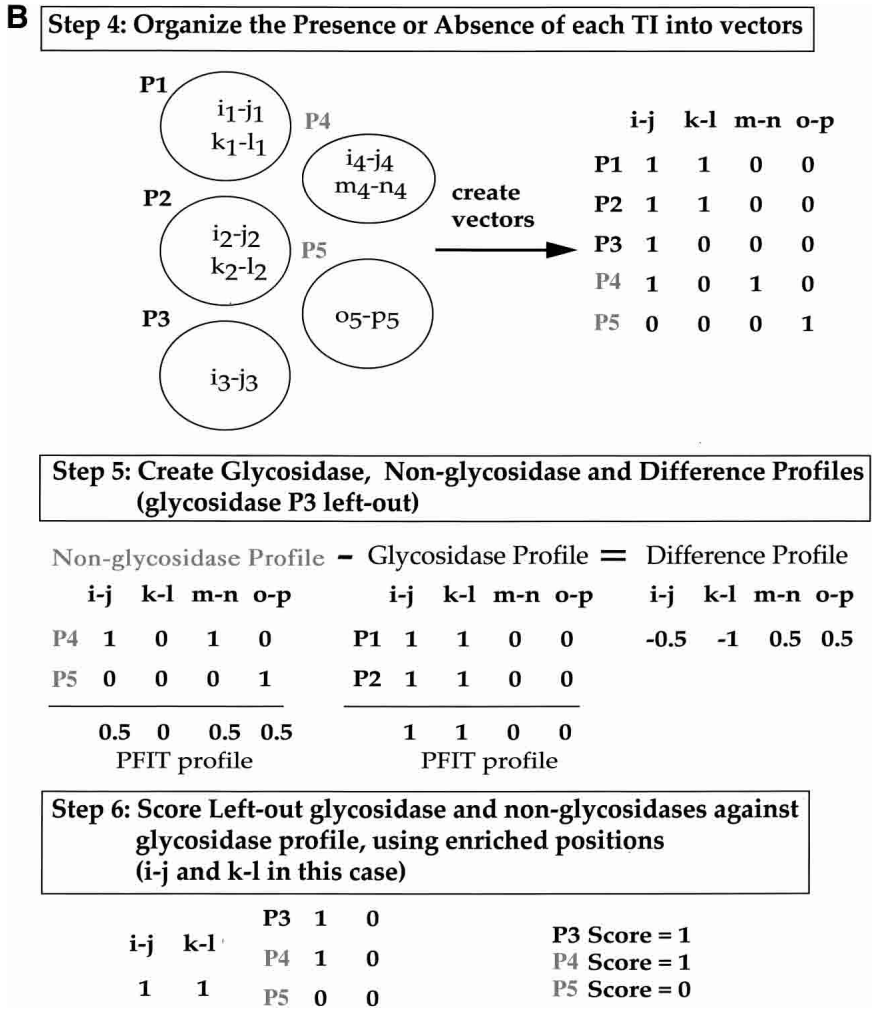

Figure 1. A simplified flow diagram for the PFIT algorithm. The first step in PFIT is to generate pairwise structural alignments between all 64 α/β barrel proteins. Results are placed in the α/β barrel DAPS database. The second step in PFIT is to find all conserved tertiary interactions (TI). As an example of a conserved tertiary interaction, consider the alignment between protein structures 7A3H (Davies et al. 1998) and 1DYS (Davies et al. 2000). The alignments between the two proteins shown in Steps 1 and 2 are the structural alignment and the structure-derived sequence alignment, respectively. The broken lines represent tertiary interactions between linked residues. Notice that Lys 113 forms a tertiary interaction with Glu 161 in 7A3H, and Lys 81 also forms a tertiary interaction with Glu 121 in 1DYS. Also notice that both the pairs of residues Lys 113 and Lys 81 and Glu 161 and Glu 121 are located at structurally equivalent positions. The locations of these residues on the 3D structures are also shown (light and dark gray ball-and-stick models). Once all of the conserved tertiary interactions have been found, Step 3 of PFIT is to count the number of proteins that contain any given tertiary interaction. In our example, there are five proteins, and a tertiary interaction between residues located at positions i and j in the structures is under consideration. Because many of the α/β barrel structures are misaligned during the alignment procedure, this process must be iterated over numerous rounds, exhaustively using all possible proteins and corresponding i-j as a reference in the next iteration. If only one iteration is used, many α/β barrel structures containing the tertiary interaction would not be correctly identified (data not shown). Step 4 of PFIT is to arrange the presence or absence of the tertiary interactions in an individual protein as a bit vector, in which 1 indicates that a tertiary interaction is present. Step 5 of PFIT involves generating two profiles, one for the glycosidase α/β barrels and another for the nonglycosidase ones. Each component of the PFIT profile is the mean for the values at identical positions in the vectors. Therefore, a component in the PFIT profile can range from 0–1. Notice that the vector for one glycosidase is left out of the profile calculation. This glycosidase will be used to test the performance of the algorithm. The glycosidase and nonglycosidase profiles are then subtracted to generate a difference profile. Components having negative values in the difference profile represent those tertiary interactions that are more conserved in the glycosidases than in the nonglycosidases, and are therefore used in the final profile. Step 6 of PFIT involves generating the glycosidase profile, using the positions identified by the difference profile, as well as eliminating the previously left-out glycosidase from the profile that will later be used for testing. The left-out glycosidase is then matched against the profile, generating the profile match score for a glycosidase hit. The same is done for the 45 nonglycosidases, generating the profile match scores for nonglycosidase hits. This procedure is repeated for all 19 glycosidases, resulting in 19 glycosidase scores and 855 nonglycosidase scores.

swering these two questions. At the time of preparation of this report, 529 α/β barrel protein structures had been deposited in the PDB (Orengo et al. 1997). The α/β barrel fold accommodates tremendous functional diversity, in which the largest family of enzymes adopting the α/β barrel fold is the well-characterized and numerously represented glycosidases (Enzyme Classification 3.2.1.x, where x represents substrate specificity; note that glycosidase proteins adopt many folds other than the α/β barrel; Henrissat and Davies 1997; Bourne and Henrissat 2001; Nagano et al. 2002). Previous studies have identified at least six distinct functional categories, on the basis of both the identities of residues in the enzyme's active site as well as the enzyme's catalytic mechanism, for all glycosidases adopting the α/β barrel fold (Henrissat and Davies 1997; Nagano et al. 2001). The sequence identity between any two members of the glycosidases may be as low as 5%, suggesting that a sequence-based algorithm such as PSI-BLAST may not detect the functional relationship between at least some of the glycosidase proteins.

Toward the goal of identifying conserved structural features in functionally related proteins, we ask whether the presence or absence of conserved tertiary interactions is more correlated among α/β barrel proteins that function as glycosidases, than among α/β barrel proteins with different functions. This is challenging due to the diversity in both sequence and structure for members of the glycosidases. Here, we show that by using the pattern of the presence or absence of conserved tertiary interactions in a protein (the PFIT algorithm), or the residue identities at these positions (the PFRIT algorithm), we are able to identify correctly some five of the six functional classes of glycosidases among all other α/β barrel proteins.

Results

Identification of conserved tertiary interactions in the α/β barrel fold

The first two steps of the PFIT algorithm are generating pairwise structural alignments between all possible α/β barrel protein pairs and identifying the conserved tertiary interactions within those alignments (Fig. 1; Materials and Methods). A tertiary interaction is defined as any hydrogen bond or salt-bridge between two residues separated by at least 10 positions from each other in the protein sequence. A tertiary interaction is considered conserved when both protein structures in the alignment contain a tertiary interaction at structurally equivalent positions. We analyzed 1932 pairwise structural alignments in the α/β barrel DAPS database and found a total of 5140 conserved tertiary interactions.

At least some of the conserved tertiary interactions found in the α/β barrel proteins are important for protein structure or catalytic function, as we observe a twofold increase in sequence identity for residues that occupy positions of conserved tertiary interactions compared with all positions. When considering the pairwise structural alignments of α/β barrel proteins, the average sequence identity between pairs of aligned residues located at positions of conserved tertiary interactions is 24%. The average sequence identity between pairs of aligned residues at all positions (including conserved tertiary interactions) is 10%. This enrichment is partly due to certain pairs of aligned proteins being functionally related, such as members of the glycosidases. Yet, even when eliminating pairs of functionally related proteins from the alignments (no two proteins share EC numbers past the second digit), the average sequence identity for residues at positions of conserved tertiary interactions is still 20%. This result supports the hypothesis that at least some of the tertiary interactions in α/β barrel protein structures are important for protein structure or catalytic function.

The next step in the PFIT algorithm is to find out how many of the 64 α/β barrel proteins conserve any one tertiary interaction. We found that the most conserved tertiary interactions are found in almost 90% of the 64 α/β barrels in our structural database. At the other extreme, there are tertiary interactions unique to one α/β barrel structure and, therefore, are not conserved.

As an example of the tertiary interactions from an α/β barrel protein, consider the structure of *Bacillus agaradhaerens* (BA) β-glycosidase (pdb 7A3H; Davies et al. 1998). This protein is one of the 19 glycosidases in the database of α/β barrel structures. BA β-glycosidase contains 72 tertiary interactions, 14 of which are conserved at structurally equivalent positions in at least 80% of the 64 α/β barrels (Table 1). A total of 30 of the tertiary interactions in BA β-glycosidase are conserved in at least 20% of the α/β barrels in our database. The locations of the 30 most highly conserved tertiary interactions on the 3-D structure of BA β-glycosidase are shown in Figure 2.

In some cases, there are tertiary interactions enriched in a functionally related subgroup, such as the glycosidases. These tertiary interactions serve functional roles that are likely to be specific to the glycosidases. Therefore, any α/β barrel proteins containing these tertiary interactions likely also function as glycosidases. This is the basis of how PFIT works.

Glycosidase detection using vectors of conserved tertiary interactions

With PFIT, α/β barrel proteins with glycosidase function can be identified from a library of α/β barrel proteins having many functions. PFIT first finds the tertiary interactions that are more conserved in the glycosidase proteins than in nonglycosidase proteins (Fig. 1; Materials and Methods).

Table 1. *Statistics for the tertiary interactions found in the BA β-glycosidase x-ray structure*

Tertiary interaction	%	Tertiary interaction	%
GLN 28 O ASN 59 ND2	70%	ARG 62 NE GLU 135 OE1	81%
ILE 94 O ASN 131 ND2	81%	ARG 62 NE ASP 99 OD2	61%
ARG 163 NE ASP 192 OD2	83%	TRP 229 O ASN 261 ND2	70%
ARG 62 NE SER 227 O	23%	SER 34 OG ALA 64 N	69%
LYS 112 NZ GLU 157 OE2	84%	SER 123 OG ASN 165 ND2	80%
TYR 129 O ASN 169 ND2	72%	TYR 202 OH GLU 228 OE2	80%
ARG 53 NH1 LEU 92 O	83%	LYS 116 NZ GLU 157 OE1	53%
HIS 200 NE2 SER 227 OG	73%	LYS 152 NZ ASN 188 OD1	81%
LYS 83 NZ GLU 121 OE2	47%	ARG 62 NH2 ASN 138 OD1	80%
ASP 120 OD1 ASN 165 ND2	89%	SER 33 OG ARG 62 NH1	81%
GLU 87 OE2 TYR 126 OH	66%	ALA 198 O THR 175 OG1	80%
THR 67 OG1 HIS 101 O	73%	ASN 138 OD1 GLU 228 OE1	84%
LYS 79 NZ GLU 121 OE1	50%	GLU 139 OE2 HIS 200 ND1	78%
ASP 215 OD1 ARG 255 NE	86%	LYS 30 N ASN 59 OD1	78%
LEU 103 O ASN 141 ND2	33%	ILE 102 O ASN 138 ND2	69%

The 30 tertiary interactions that are conserved at structurally equivalent positions in at least 20 percent of the 64 α/β barrels in our database are shown. The residues and atoms forming each tertiary interaction and their numerical positions in the structure are given. The percentage (%) of structures that conserve the respective interaction is also given. All tertiary interactions are hydrogen bonds, except LYS 116–GLU 157, which is a salt-bridge.

PFIT organizes the presence or absence of each tertiary interaction in a given protein as a component of a bit vector. Leaving one glycosidase aside for testing the performance of PFIT, vectors for the remaining 18 glycosidases were averaged into a profile (PFIT profile). The remaining glycosidase vector as well as the 45 nonglycosidase vectors are then aligned to the PFIT profile to obtain profile match scores. This leave-one-out procedure is repeated for all 19 glycosidases, yielding 19 glycosidase scores and 855 nonglycosidase scores (45 total nonglycosidase scores for each iteration of PFIT). The profile match scores were used to generate a Receiver-Operator Characteristic (ROC) curve (Fig 3).

PFIT correctly identifies 68% of the glycosidases with an overall error rate of 14% (PFIT profile match score cutoff of 1.08; Fig. 3; Table 2). Notably, at least one member from five of the six functional subclasses for the α/β barrel glycosidases, as defined by Nagano and Thornton (Nagano et al. 2001), are included in the correctly identified glycosidases (F1, F2, F4/F5, and F6 subclasses; CAZy families GH1, GH2, GH3, GH5, GH10, GH13, GH18, and GH20). Using a more liberal cutoff of 0.75, PFIT correctly identifies 84% of the glycosidases with an error rate of 25%. Additional members of the F1 and F2 functional subclasses are now included (CAZy families GH6 and GH17). The analysis presented in Table 2 omits CAZy families GH26, GH27, GH42, GH53, GH56, and GH67 for the following reasons. Structural representatives for these families were not available in CATH during the time of our analysis, or the proteins belonging to these families shared significant sequence similarity to a protein already in our database.

Glycosidase detection using the residues located at positions of conserved tertiary interactions: PFRIT

The residues that occupy the positions of the tertiary interactions enriched in the glycosidases are also somewhat conserved, and their identities can be used to predict glycosidase function in α/β barrel structures. This algorithm, PFRIT, is a modification of PFIT, in which the identities of the residues that form the tertiary interactions are now used to create a vector of residues. Averaging of the glycosidase vectors now results in a sequence profile. Using the same leave-one-out procedure outlined in the previous section, PFRIT correctly identifies 84% of the glycosidases with an overall error rate of 29% (PFRIT profile match score cutoff of 0.27; Fig. 3; Table 2). In addition, PFRIT has one advantage over PFIT, a conservative cutoff value of 3 for the PFRIT profile matching score correctly identifies 37% of the glycosidases with an overall error rate of 0. All of the glycosidases detected at this cutoff value correspond to the F2 functional subgroup (CAZy families GH1, GH2, and GH5; Table 2).

Glycosidase detection using PSI-BLAST

The sequence-based program PSI-BLAST can also be used to identify the α/β barrel proteins with glycosidase function from all 64 α/β barrel proteins in our database, however, PFIT and PFRIT are both more selective and sensitive than PSI-BLAST. Using PSI-BLAST and one of the 19 glycosidase proteins as a query, we searched our database of α/β barrel proteins. The E-values for the matches between the query protein and any given protein from the database were

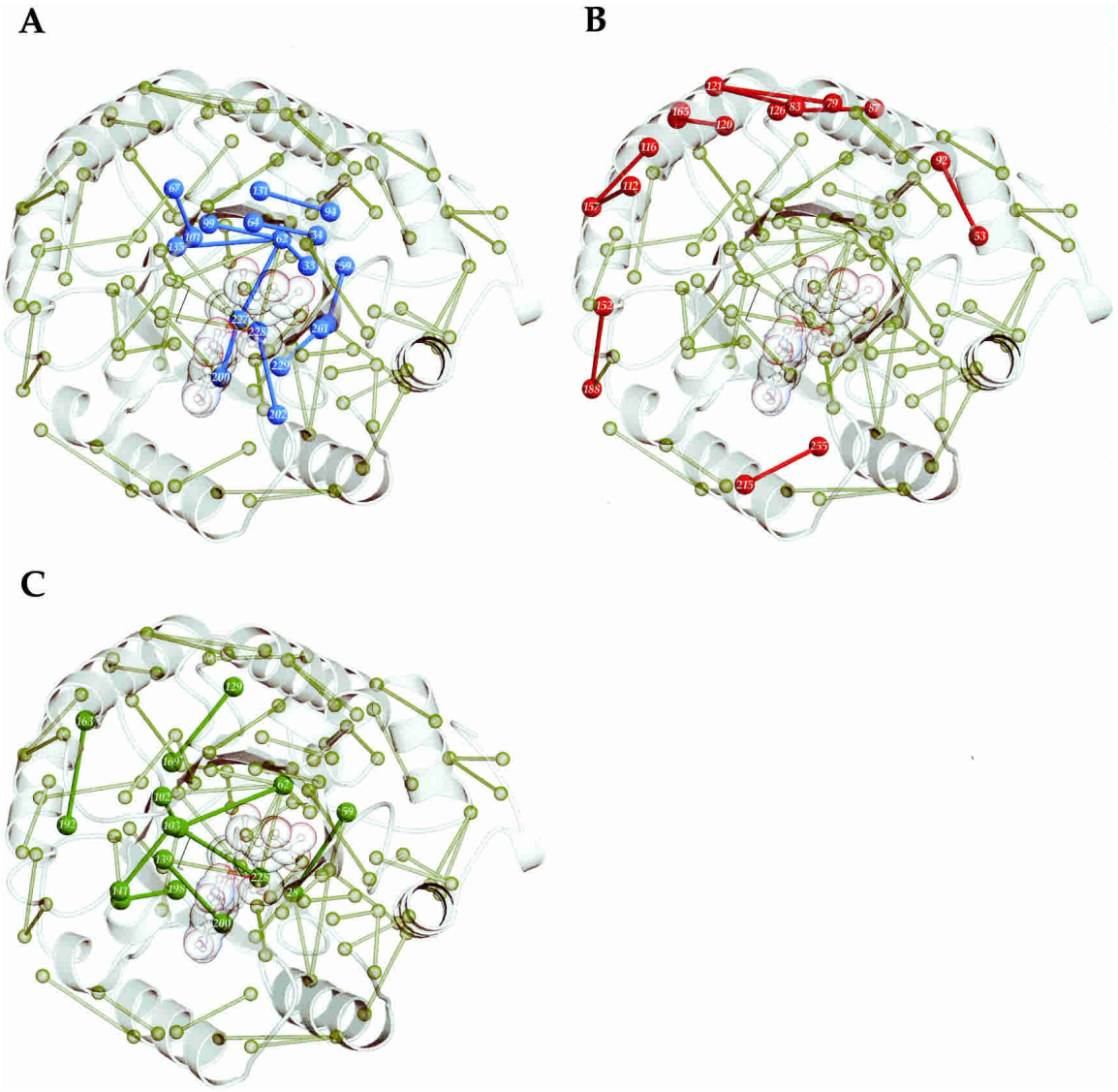

Figure 2. The 30 tertiary interactions conserved in at least 20% of the α/β barrel proteins. Ribbon diagram of β-glycosidase from *Bacillus agaradhaerens* (pdbcode 7A3H). The locations of the Cα atoms of the interacting residues are shown as spheres, connected by a rod representing the tertiary interaction. The substrate cellobioside is represented as gray and red space-filling atoms. The 42 tertiary interactions not conserved in at least 20% of the structures are light brown. (*A*) β-strand/β-strand conserved tertiary interactions in blue; (*B*) α-helix/α-helix conserved tertiary interactions in red; (*C*) β-strand/loop, α-helix/loop, or loop/loop conserved tertiary interactions in green. Notice that the majority of conserved tertiary interactions cluster near the active site in the α/β barrel structure.

used to generate a ROC curve. Notice that this curve falls under the ROC curve for both the PFIT and PFRIT algorithms, demonstrating that these algorithms have greater predictive power for glycosidase detection than PSI-BLAST. We also attempted to combine information from PFIT and PFRIT with PSI-BLAST, however, we did not observe any improvement in either selectivity or sensitivity.

Discussion

We show that conserved tertiary interactions and the residues that occupy these positions are useful for identifying α/β barrel proteins that function as glycosidases from a set of α/β barrel proteins of any function. We present two independent algorithms that are complementary to each other as follows: (1) PFIT, which is based on the pattern of the presence or absence of tertiary interactions in protein structures; and (2) PFRIT, which is based on the identity of the residues found at conserved tertiary interaction positions. Note that PFIT and PFRIT are not dependent on a multiple sequence alignment, which is often difficult to generate from the sequences of distantly related proteins. Using the PFIT and PFRIT algorithms, we are able to identify

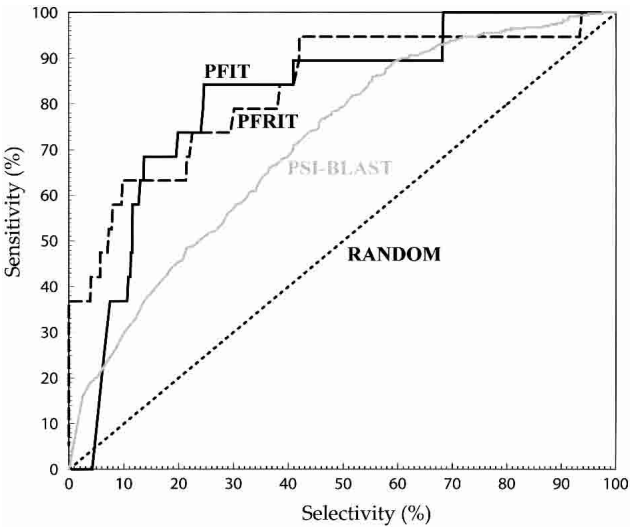

Figure 3. PFIT correctly identifies a high percentage of the glycosidases in our database with a low error rate. The profile match scores for both the 19 glycosidases and the 45 nonglycosidases were used to generate a Receiver-Operator Characteristic (ROC) plot (also commonly referred to as a selectivity vs. sensitivity plot). The ROC curve for PFIT is the black line. The ROC curve for PFRIT is the black broken line. The ROC curve for PSI-BLAST is the solid gray line. Sensitivity, also referred to as coverage, is the percentage of glycosidases that had a profile match score higher than the ROC cutoff value. Selectivity, also referred to as the error rate, is the percentage of nonglycosidases that had a profile match score higher than the ROC cutoff value. Note that a higher ROC cutoff value will decrease the error rate but also decreases the coverage. A diagonal line (black dotted line) would be expected if one were to randomly guess the function of a given protein. Notice that the ROC curves for all three algorithms deviate significantly from the diagonal, indicating an increased performance over random.

correctly glycosidases from five of the six functional classes as defined by Nagano and Thornton (Nagano et al. 2001; the F4 and F5 groups are related at the sequence level, such that only representatives from the F5 group are in our culled database because of the BLAST restrictions). Note that the glycosidase-related protein concanavalin B (pdb 1CNV; Hennig et al. 1995), which has an unrelated function to glycoside hydrolysis, is not identified as a glycosidase by PFIT or PFRIT (data not shown).

Conserved tertiary interactions and the residues that occupy their positions are more conserved in the glycosidases than in α/β barrel proteins as a whole, because they serve specific roles in the function of those enzymes. For example, Arg 62 in BA β-glycosidase forms a network of tertiary interactions with many residues in the protein, such as Ser 33, Asp 99, Ser 227, Glu 135, and Asn 138 (Fig. 2). These interactions are important to stabilize the energetically unfavorable conformation of the guanidino group of Arg 62, which is oriented toward the catalytic nucleophile Glu 228. Arg 62 forms a charge-dipole interaction with Glu 228, consequently, making that residue more nucleophilic. However, none of the tertiary interactions with Arg 62 is strictly conserved in the glycosidase structures, and nonglycosidase structures may contain some of these interactions at structurally equivalent positions as well. This observation reflects the inherent adaptability, both at the structural and sequence level, that the glycosidases have while still performing the same function. Our method of using many conserved tertiary interactions in a vector to establish the function of the α/β barrel protein, rather than a single tertiary interaction, takes into account this adaptability.

An evolutionary relationship for members of the α/β barrel fold

Others have found evidence for divergent evolutionary relationships among the members of the α/β barrel glycosidases (Reardon and Farber 1995). Nagano and Thornton detected weak sequence similarity between members of the F1 and F2 groups (CAZy families GH1, GH2, GH3, GH5, GH6, GH10, GH13, GH17, and GH20) using the program PSI-BLAST (Nagano et al. 2001). Furthermore, they also were able to align the catalytic residues for members of the F4, F5, and F6 groups (CAZy families GH14, GH18, and GH20).

Our results also support the notion of a divergent evolutionary relationship for members of the α/β barrel glycosi-

Table 2. *Profile match scores for the 19 members of the glycosidase enzymes that are also found in the α/β barrel DAPS database*

PDBcode	EC	Functional group	CAZy	PFIT	PFRIT
7A3H	3.2.1.4	F2	GH5	1.62	4.18
1EDG	3.2.1.4	F2	GH5	1.61	4.27
1E56	3.2.1.21	F2	GH1	1.31	3.45
1CEO	3.2.1.4	F2	GH5	1.61	4.73
1QNR	3.2.1.78	F2	GH5	1.62	4.39
1DYS	3.2.1.4	F2	GH6	0.91	1.04
1BQC	3.2.1.78	F2	GH5	1.62	3.07
1TML	3.2.1.4	F2	GH6	0.17	0.28
1EX1	3.2.1.58	F2	GH3	1.33	0.54
1TAX	3.2.1.8	F2	GH10	1.08	1.83
1DP0	3.2.1.23	F2	GH2	1.62	3.89
1AQ0	3.2.1.73	F2	GH17	0.75	1.15
1SMD	3.2.1.1	F1	GH13	1.62	2.36
1QHP	3.2.1.1	F1	GH20	0.76	1.93
1BF2	3.2.1.68	F1	GH13	1.31	0.59
1EDT	3.2.1.96	F5	GH18	–0.45	–0.18
1EOK	3.2.1.96	F5	GH18	1.17	–0.64
1BYB	3.2.1.2	F3	GH14	–0.45	–0.05
1C7S	3.2.1.52	F6	GH20	1.36	1.49

The EC number represents the functional classification for each enzyme, where the last digit refers to the substrate specificity (International Union of Biochemistry and Molecular Biology [http://www.chem.qmul.ac.uk/iubmb]). The functional group is based on the classification of α/β barrel glycosidases as defined by Nagano and Thornton. The CAZy family classifications for glycosidases (Bourne and Henrissat 2001) are also given.

dases. In this investigation, the average sequence identity for aligned residues at the positions of conserved tertiary interactions is 32% when considering the pairwise structural alignments between members of the glycosidases. The residues that occupy positions of conserved tertiary interactions are more conserved, because at least some of the tertiary interactions are important for the function of the protein. Mutation of either residue contributing to the conserved interaction might disrupt the interaction and, therefore, the function of the protein. Therefore, we would expect that both residues forming the interaction would need to be mutated at the same time to preserve the interaction, which would be rare.

Conserved tertiary interactions and PFIT profiles in the context of functional genomics

As structural genomics projects continue to produce more protein structures, certainly many of these structures will be of proteins that adopt the α/β barrel fold. Past experience tells us that some of these proteins will not share significant sequence identity with other α/β barrel proteins of known function. Here, we show two algorithms in which these α/β barrel proteins can be tested for glycosidase function.

One goal in the future will be to apply our method to both α/β barrel proteins with different functions, as well as to proteins that adopt other folds. In fact, conserved tertiary interactions are found in other folds, and the residues occupying those positions are more conserved than the residues at other positions (data not shown). The limiting factor for applying PFIT and PFRIT to other folds is obtaining structural alignments of high quality, an essential requirement for PFIT and PFRIT to work. α/β barrel proteins are more easily aligning than other proteins because of the pseudo-eightfold symmetry of the barrel. In our experience, structures belonging to other folds are more difficult to align for a variety of reasons. In the future, structural alignment tools will have to be developed that focus on members of only one fold, so that the alignment algorithm can exploit the features that are unique to that fold. Nevertheless, structural genomics promises to deliver more data, opening the door for algorithms such as PFIT and PFRIT, as well as others, to contribute to the important goal of identifying the functions of proteins from structures.

Materials and methods

Databases

CATH version 2.4 contains some 529 structural representatives of the α/β barrel fold (Orengo et al. 1997). Our goal is to detect distant relationships among these proteins beyond the capabilities of sequence similarity methods such as BLAST (Altschul et al. 1990). Therefore, from the α/β barrel proteins whose x-ray structures have been determined, we created a culled list of proteins with low sequence similarity. We used BLAST to obtain all pairwise sequence alignments (139,656) between the 529 proteins. For pairs of aligned proteins with an E-value of 10e-20 or lower, the protein structure with lower resolution was removed from the database. The final list contains 64 α/β barrel protein structures. These 64 structures were then used to generate a database of pairwise structural alignments, called the α/β barrel DAPS (Database of Aligned Protein Structures) database (Mallick et al. 2002). All pairwise combinations of these structures were aligned using the structural alignment program COMPARER (Sali and Blundell 1990). Some 99% of the pairwise structural alignments in our database are between proteins sharing <20 percent sequence identity, and 84.4% share <15% sequence identity. This database is publicly available at the following Web address: http://www.doe-mbi.ucla.edu/~kleiger/alphaBetaBarrels.html.

Tertiary interactions were found in the protein structures using the program CONTACT (CCP No. 4 1994). Two atoms were considered hydrogen bonded if the distance between the hydrogen bond donor and the hydrogen bond acceptor atoms is no shorter than 2.2 Å, or no greater than 3.5 Å. CONTACT also analyzes the angles between the hydrogen-bond donor, hydrogen-bond acceptor, and hydrogen atoms, rejecting those that do not fall within acceptable tolerances. Two atoms were considered to form a salt-bridge if they are oppositely charged and fall within 5.0 Å of each other.

The PFIT algorithm

The PFIT algorithm is described in Figure 1. Briefly, each α/β barrel protein is presented as a bit vector, in which 1 indicates the presence of a given tertiary interaction in the structure, and 0 indicates the absence. These vectors are then averaged into a PFIT profile. Each component of the PFIT profile is calculated by summing the values at identical positions in the vectors, and then dividing by the total number of vectors. The numerical value for each component in the PFIT profile can range from 0 to 1. One glycosidase is left out of the glycosidase profile to be used later for testing the algorithm. Therefore, the glycosidase profile is based on 18 proteins, and the nonglycosidase profile is based on 45 other α/β barrel proteins. The glycosidase profile is then subtracted from the nonglycosidase profile to create a difference profile. Components with values greater than −0.2 were removed from the final glycosidase profile.

Profile match scores between both the left-out glycosidase and the 45 nonglycosidase vectors is given by:

$$\mathrm{Pr}ofileMatchScore(PFIT) = \sum_{j=1}^{N} f(j)i \qquad (1)$$

in which $f(j)$ is the value at position j from the glycosidase profile, i is the identity of the bit at position j in the vector, and N is the number of components in the profile.

Profile match scores are converted to normalized profile match scores by the following equation:

$$normalized\ \mathrm{Pr}ofileMatchScore = \frac{X - \langle X \rangle}{\sigma} \qquad (2)$$

in which X is the profile match score, <X> is the average profile match score, and σ is the standard deviation of profile match scores. This procedure is repeated for all 19 glycosidases. The final

profile match scores were used to create the PFIT ROC curve (sensitivity versus selectivity) in Figure 3.

The PFRIT algorithm

The PFRIT algorithm is very similar to PFIT. Tertiary interactions that are specific for glycosidases are identified using the difference profile approach described for PFIT. For each glycosidase or nonglycosidase vector, if a tertiary interaction is present in that protein, the amino acid identities for the residues occupying the positions of the tertiary interaction are used in the vector. Therefore, rather than bit vectors, these vectors are now amino acid sequences. If the tertiary interaction is not present, *X* s are used instead. Profile match scores between both the left-out glycosidase and the 45 nonglycosidase vectors are now given by:

$$profileMatchScore(PFRIT) = \sum_{j=1}^{N} f_i(j)i \qquad (3)$$

in which $f_i(j)$ is the frequency of amino acid i at position j in the glycosidase profile, i is the identity of the amino acid in the protein sequence that is being matched against the profile, and N is the number of components in the glycosidase profile.

Profile match scores are normalized in an identical manner as for PFIT. This procedure is repeated for all 19 glycosidases, and the final profile match scores were used to generate the PFRIT ROC curve in Figure 3.

PSI-BLAST and HMMER

The program PSI-BLAST (Altschul et al. 1997) was used to search our database of 64 α/β barrel proteins. The query protein was one of the 19 glycosidases. The PSI-BLAST procedure was iterated for two cycles. E-values were obtained for the matches between this query protein and all other 63 α/β barrel proteins. This procedure was then repeated for all 19 glycosidases.

Remote homology detection was also carried out by the program HMMER (Bateman et al. 1999). We found that the results from PSI-BLAST and HMMER were highly similar.

Acknowledgments

We thank Magdalena Ivanova and Tom Graeber for helpful discussion and review of this manuscript. We thank DOE, NIH, and HHMI for support.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

References

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. 1990. Basic local alignment search tool. *J. Mol. Biol.* **215:** 403–410.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. 1997. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* **25:** 3389–3402.

Bateman, A., Birney, E., Durbin, R., Eddy, S.R., Finn, R.D., and Sonnhammer, E.L. 1999. Pfam 3.1: 1313 multiple alignments and profile HMMs match the majority of proteins. *Nucleic Acids Res.* **27:** 260–262.

Beamer, L.J., Carroll, S.F., and Eisenberg, D. 1997. Crystal structure of human BPI and two bound phospholipids at 2.4 Å resolution. *Science* **276:** 1861–1864.

Bourne, Y. and Henrissat, B. 2001. Glycoside hydrolases and glycosyltransferases: Families and functional modules. *Curr. Opin. Struct. Biol.* **11:** 593–600.

Browne, W.J., North, A.C., Phillips, D.C., Brew, K., Vanaman, T.C., and Hill, R.L. 1969. A possible three-dimensional structure of bovine α-lactalbumin based on that of hen's egg-white lysozyme. *J. Mol. Biol.* **42:** 65–86

C.C.P., No. 4, 1994. The CCP4 suite: Programs for protein crystallography. *Acta Crystallogr.* **D50:** 760.

Davies, G.J., Mackenzie, L., Varrot, A., Dauter, M., Brzozowski, A.M., Schulein, M., and Withers, S.G. 1998. Snapshots along an enzymatic reaction coordinate: Analysis of a retaining β-glycoside hydrolase. *Biochemistry* **37:** 11707–11713.

Davies, G.J., Brzozowski, A.M., Dauter, M., Varrot, A., and Schulein, M. 2000. Structure and function of Humicola insolens family 6 cellulases: Structure of the endoglucanase, Cel6B, at 1.6 Å resolution. *Biochem. J.* **348 Pt. 1:** 201–207.

Eisenberg, D., Marcotte, E.M., Xenarios, I., and Yeates, T.O. 2000. Protein function in the post-genomic era. *Nature* **405:** 823–826.

Hennig, M., Jansonius, J.N., Terwisscha van Scheltinga, A.C., Dijkstra, B.W., and Schlesier, B. 1995. Crystal structure of concanavalin B at 1.65 Å resolution. An "inactivated" chitinase from seeds of *Canavalia ensiformis*. *J. Mol. Biol.* **254:** 237–246.

Henrissat, B. and Davies, G. 1997. Structural and sequence-based classification of glycoside hydrolases. *Curr. Opin. Struct. Biol.* **7:** 637–644.

International Union of Biochemistry and Molecular Biology. Nomenclature Committee., and Webb, E.C. 1992. *Enzyme nomenclature 1992: Recommendations of the nomenclature committee of the international union of biochemistry and molecular biology on the nomenclature and classification of enzymes*, pp. xiii, 862. Published for the International Union of Biochemistry and Molecular Biology, Academic Press, San Diego, CA.

Kanehisa, M. and Bork, P. 2003. Bioinformatics in the post-sequence era. *Nat. Genet.* **33:** 305–310.

Kleiger, G., Beamer, L.J., Grothe, R., Mallick, P., and Eisenberg, D. 2000. The 1.7 Å crystal structure of BPI: A study of how two dissimilar amino acid sequences can adopt the same fold. *J. Mol. Biol.* **299:** 1019–1034.

Kleiger, G., Perry, J., and Eisenberg, D. 2001. 3D structure and significance of the GPhiXXG helix packing motif in tetramers of the E1β subunit of pyruvate dehydrogenase from the archeon *Pyrobaculum aerophilum*. *Biochemistry* **40:** 14484–14492.

Mallick, P., Goodwill, K.E., Fitz-Gibbon, S., Miller, J.H., and Eisenberg, D. 2000. Selecting protein targets for structural genomics of *Pyrobaculum aerophilum*: Validating automated fold assignment methods by using binary hypothesis testing. *Proc. Natl. Acad. Sci.* **97:** 2450–2455.

Mallick, P., Weiss, R., and Eisenberg, D. 2002. The directional atomic solvation energy: An atom-based potential for the assignment of protein sequences to known folds. *Proc. Natl. Acad. Sci.* **99:** 16041–16046.

Nagano, N., Porter, C.T., and Thornton, J.M. 2001. The (βα)(8) glycosidases: Sequence and structure analyses suggest distant evolutionary relationships. *Protein Eng.* **14:** 845–855.

Nagano, N., Orengo, C.A., and Thornton, J.M. 2002. One fold with many functions: The evolutionary relationships between TIM barrel families based on their sequences, structures and functions. *J. Mol. Biol.* **321:** 741–765.

Orengo, C.A., Michie, A.D., Jones, S., Jones, D.T., Swindells, M.B., and Thornton, J.M. 1997. CATH—a hierarchic classification of protein domain structures. *Structure* **5:** 1093–1108.

Reardon, D. and Farber, G.K. 1995. The structure and evolution of α/β barrel proteins. *FASEB J.* **9:** 497–503.

Sali, A. and Blundell, T.L. 1990. Definition of general topological equivalence in protein structures. A procedure involving comparison of properties and relationships through simulated annealing and dynamic programming. *J. Mol. Biol.* **212:** 403–428.