Local structure based method for prediction of the biochemical function of proteins:

Applications to glycoside hydrolases

Ramya Parasuram^{1†}, Caitlyn L. Mills[†], Zhouxi Wang², Saroja Somasundaram, Penny J. Beuning, Mary Jo Ondrechen^{*}

*Department of Chemistry and Chemical Biology, Northeastern University, Boston, MA 02115
USA*

*Corresponding author at: 360 Huntington Ave., Boston, MA 02115, USA.

E-mail addresses: mjo@neu.edu (M.J. Ondrechen)

[†]These authors contributed equally to the manuscript.

¹ Present address: Dana-Farber Cancer Institute, 450 Brookline Ave., Boston, MA 02115

² Present address: National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bldg. 38 A, Room 5S508, 8600 Rockville Pike, Bethesda, MD 20894

ABSTRACT

Thousands of protein structures of unknown or uncertain function have been reported as a result of high-throughput structure determination techniques developed by Structural Genomics (SG) projects. However, many of the putative functional assignments of these SG proteins in the Protein Data Bank (PDB) are incorrect. While high-throughput biochemical screening techniques have provided valuable functional information for limited sets of SG proteins, the biochemical functions for most SG proteins are still unknown or uncertain. Therefore, computational methods for the reliable prediction of protein function from structure can add tremendous value to the existing SG data. In this article, we show how computational methods may be used to predict the function of SG proteins, using examples from the six-hairpin glycosidase (6-HG) and the concanavalin A-like lectins/glucanases (CAL/G) superfamilies. Using a set of predicted functional residues, obtained from computed electrostatic and chemical properties for each protein structure, it is shown that these superfamilies may be sorted into functional families according to biochemical function. Within these superfamilies, a total of 18 SG proteins were analyzed according to their predicted, local functional sites: 13 from the 6-HG superfamily, five from the CAL/G superfamily. Within the 6-HG superfamily, an uncharacterized protein bacova_03626 from *Bacteroides ovatus* (PDB 3ON6) and a hypothetical protein BT3781 from *Bacteroides thetaiotaomicron* (PDB 2P0V) are shown to have very strong active site matches with exo- α -1,6-mannosidases, thus likely possessing this function. Also in this superfamily, it is shown that protein BH0842, a putative glycoside hydrolase from *Bacteroides halodurans* (PDB 2RDY), has a predicted active site that matches well with a known α -L-galactosidase. In the CAL/G superfamily, an uncharacterized glycosyl hydrolase family 16 protein from *Mycobacterium smegmatis* (PDB 3RQ0) is shown to have local structural similarity at the

predicted active site with the known members of the GH16 family, with the closest match to the endoglucanase subfamily. The method discussed herein can predict whether an SG protein is correctly or incorrectly annotated and can sometimes provide a reliable functional annotation. Examples of application of the method across folds, comparing active sites between two proteins of different structural folds, are also given.

KEYWORDS

Protein function prediction, Structurally Aligned Local Sites of Activity (SALSA), glycoside hydrolases, six-hairpin glycosidases, concanavalin A-like lectins/glucanases

ABBREVIATIONS

Concanavalin A-like lectins/glucanases: CAL/G

Glycoside Hydrolases: GHs

Partial Order Optimum Likelihood: POOL

Six-hairpin glycosidase: 6-HG

Structural Genomics: SG

Structurally Aligned Local Sites of Activity: SALSA

THEoretical Microscopic Anomalous TItration Curve Shapes: THEMATICS

1. INTRODUCTION

Currently there are over 13,500 Structural Genomics (SG) protein structures deposited in the Protein Data Bank (PDB) [1], in large part as a result of the Protein Structure Initiative (PSI). However, most of these structures are of unknown or uncertain function since the determination of biochemical function from three-dimensional (3D) structure alone has been far more challenging than originally anticipated. Furthermore, functional annotation error rates can be as high as 80% [2] within databases and these errors can propagate as more proteins are annotated by automated methods. Reliable computational approaches to the prediction of biochemical function from the 3D structure of a protein are therefore important and timely.

Many current structure-based methods, outlined in recent reviews [3-8], analyze the overall geometry of a given protein. Methods such as Dali [9, 10], Combinatorial Extension (CE) [11], or Topofit [12], are used for comparing the global 3D structure of a protein to another protein, or a protein set. However, global structural similarity does not necessarily imply functional similarity and assignment of function on this basis has proved to be problematic. Indeed, proteins with different functions and different catalytic residues can have the same global fold, while proteins with the same function and same catalytic residues can have different global folds. The use of local 3D structural motifs is now emerging as a more promising path for correct functional annotation from structure [8, 13-18]. We have proposed that local structural signatures, obtained from predicted sets of biochemically active residues based on computed electrostatic and chemical properties, lead to more reliable functional assignments [18, 19] than overall 3D structural comparison.

In this article, we review previously published methods and apply these methods to analyze two superfamilies: the six-hairpin glycosidase (6-HG) (SCOP 48208) and the

concanavalin A-like lectins/glucanases (CAL/G) (SCOP 49899) superfamilies. Starting with a set of 3D structures of proteins of known biochemical function, each residue in each structure is analyzed according to its computed electrostatic and chemical properties. Partial Order Optimum Likelihood (POOL) (Figure 1, left) is used to predict the residues that participate most in the biochemical function [20, 21]. Common sets of functionally active, spatially aligned residues are then used to define a local chemical signature for each of the known functional families within the two superfamilies, using the Structurally Aligned Local Sites of Activity (SALSA) method [19] (Figure 1, right). Next, functionally active residues are predicted for the SG protein structures and these predictions are aligned and compared with the chemical signatures for each functional family. The quality of the alignment between the sets of residues for the SG protein with those of the chemical signatures of the known functional families can then provide functional information and sometimes predict the function of the SG proteins. Poor alignments can also be used to rule out unlikely functions and thus identify misannotations within databases.

In addition to predicting the function of SG proteins within one superfamily, we can also analyze SG proteins from other superfamilies. Once the chemical signatures have been established for given functional types, one can search for matches between the predicted sets of residues for an SG protein of any structure with the chemical signatures obtained for known functional types. This helps to increase knowledge and understanding of proteins of uncertain or unknown function and to obtain better functional annotations. For example, glycoside hydrolases (GHs) occur in a very diverse range of folds, including the $(\alpha/\alpha)_6$ barrel, $\alpha + \beta$ lysozyme, $(\beta/\alpha)_8$ “TIM” barrel, 5-bladed β propeller, 6-bladed β propeller, β sandwich, and $(\beta)_3$ solenoid [22]. The two superfamilies featured here, the 6-HG superfamily, a class of all α proteins, and the CAL/G superfamily, a class of all β proteins, have completely different structural folds. However,

members of the two superfamilies can catalyze the same types of reactions. For instance, similar endoglucanases occur in both superfamilies.

Our method stands apart from the others in many ways. Specifically, we are able to predict distal residues in addition to residues that are in contact with a bound ligand or with a reacting substrate molecule [23]; distal residues are defined as residues that do not interact directly with the bound ligand. Thus, our method utilizes a more complete set of residues than those available in the Catalytic Site Atlas (CSA) [24] or PDB SITE field. While SALSA analyzes the local active site similarities between two proteins, POOL uses a combination of electrostatics, sequence-based, and surface geometric-based methods, as illustrated in Figure 1. In a departure from purely informatics-based methods, computed chemical properties of the individual residues are incorporated into our method, to establish the local structural motifs.

2. METHODS

2.1. Identification of protein structures of known function in the superfamilies

Proteins containing an experimentally-determined function (simply referred to as a known biochemical function here) and a known structure were selected from a number of databases. First, PASS2 [25] was used to determine an initial set of proteins containing a characterized function and structure from both superfamilies (6-HG superfamily SCOP code: 48208, CAL/G superfamily SCOP code: 49899). Next, the Dali server [9, 10] was used to identify structurally similar proteins deposited in the PDB [1] using the initial set of proteins as a starting point. Finally, for functional families containing only one functionally characterized protein with a known structure, the UniProt [26] database was used to find protein sequences that have the same characterized function. Protein sequences with greater than 60% sequence identity were discarded; low sequence similarities between proteins within a functional family are

preferred to avoid any bias when analyzing spatially overlapping residues [27]. In the cases where a functional family contains only one or two available structures, an attempt to generate a homology model was made from protein sequences of known function. In these cases with fewer than three structures of known function, if there are no available sequences for a functional family with verified biochemical function, homology models could not be generated with confidence about their function. In such a case where the functional family has only one or two protein structures or models of known function, the corresponding SALSA chemical signatures are deemed to be less certain. These homology models were constructed using the YASARA/WHATIF suite of programs [28] and are listed in Table S1 along with additional information about model assessment. The sets of structures of 6-HG and CAL/G superfamily members of known function, used to obtain the sets of biochemically active residues characteristic of each functional family, are given in Table 1 and Table 2, respectively. The CAZy classification numbers [29] and Pfam family numbers [30] accompany each functional family where applicable. The CAZy database categorizes proteins into families based on their sequence homology. Therefore, each CAZy family can be multifunctional. For example, in the 6-HG superfamily, CAZy family GH8 contains an endoglucanase, xylanase, chitosanase, and endo-xylanase. In Table 3, we observe a few aligned spatial positions that contain different POOL-predicted residue types within a functional family. For instance, for the endoglucanases, xylanases, chitosanases (GH8), different POOL-predicted residue types occur in spatial positions one, two, and three. Similarly, for the GH16 functional family in the CAL/G superfamily, spatial positions nine, 14, 25, and 27 have different POOL-predicted residue types (Table 4). These differences may be signatures of different specificities within a functional family.

Table 1. Structures of 6-HG superfamily members of known biochemical function used to obtain local chemical signatures of catalytically active residues for each functional family and sorted into 13

functional families according to biochemical function. *Indicates a homology model was built from sequences for which experimental data are available to support the functional assignment.

Functional Family	E.C. #	PDB ID	CAZy	Pfam	Protein Name	Source
1,4- α -Glucan Glucohydrolase	3.2.1.3	1GAI	GH15	PF00723	Glucosylase	<i>Aspergillus awamori</i>
		1AYX			Glucosylase	<i>Saccharomyces fibuligera</i>
		1LF9			Bacterial Glucosylase	<i>Thermoanaerobacterium thermosaccharolyticum</i>
	3.2.1.70	1UG9			Glucodextranase	<i>Arthrobacter globiformis</i>
Exo- α -1,6-Mannosidase	3.2.1.-	3QT9	GH125	PF06824	Exo- α -1,6-mannosidase	<i>Clostridium perfringens</i>
		3QSP			Exo- α -1,6-mannosidase	<i>Streptococcus pneumoniae</i>
Endoglucanase/Xylanase/ Chitosanase	3.2.1.4	1CEM	GH8	PF01270	Endoglucanase CelA	<i>Clostridium thermocellum</i>
	3.2.1.156	1WU4			Xylose-releasing exo- oligoxyranase	<i>Bacillus halodurans</i> C-125
	3.2.1.132	1V5C			Chitosanase	<i>Bacillus</i> sp. K17
	3.2.1.8	1H12			Endo-1,4- β -xylanase	<i>Pseudoalteromonas haloplanktis</i>
Cellulases	3.2.1.4	1CLC	GH9	PF00759	Endoglucanase CelD	<i>Clostridium thermocellum</i>
		1KFG			Endoglucanase Cel9G	<i>Clostridium cellulolyticum</i>
		1KSC			Endo- β -1,4-glucanase	<i>Nasutitermes takasagoensis</i>
		1IA6			Cellulase Cel9M	<i>Clostridium cellulolyticum</i>
Unsaturated Glucuronyl Hydrolase	3.2.1.-	2D5J	GH88	PF07470	Unsaturated glucuronyl hydrolase	<i>Bacillus</i> sp. gl1
		2ZZR			Unsaturated glucuronyl hydrolase	<i>Streptococcus agalactiae</i>
α -L-Rhamnosidase	3.2.1.40	2OKX	GH78	PF05592	α -L-Rhamnosidase	<i>Bacillus</i> sp. gl1
		3W5M			α -L-Rhamnosidase	<i>Streptomyces avermitilis</i>
		ALR1*			α -L-Rhamnosidase	<i>Thermomicrobia bacterium</i> PRI-1686
1,2- α -L-Fucosidase and α -L-galactosidase	3.2.1.-	4UFC	GH95	PF14498	α -L-galactosidase	<i>Bacteroides ovatus</i>
	3.2.1.63	2EAC			1,2- α -L-Fucosidase	<i>Bifidobacterium bifidum</i>
		ALF1*			1,2- α -L-Fucosidase	<i>Arabidopsis thaliana</i>
		ALF2*			1,2- α -L-Fucosidase	<i>Lilium longiflorum</i>
Trehalase	3.2.1.28	2JF4	GH37	PF01204	Trehalase	<i>Escherichia coli</i>
		TRE1*			Trehalase	<i>Apis mellifera</i>
Unsaturated Rhamnogalacturonyl Hydrolase	3.2.1.172	2D8L	GH105	PF07470	Unsaturated rhamnogalacturonyl hydrolase	<i>Bacillus subtilis</i>
α -Amylase	3.2.1.-	3REN	GH126	PF01270	α -Amylase	<i>Clostridium perfringens</i>
Phosphorylase I	2.4.1.-	1V7X	GH94	PF06165	Chitobiose phosphorylase	<i>Vibrio proteolyticus</i>
	2.4.1.20	2CQS			Cellobiose phosphorylase	<i>Cellobacterium gilvum</i>
	2.4.1.49	CDP1*			Cellobiose phosphorylase	<i>Clostridium stercorarium</i>
Phosphorylase II	2.4.1.8	1H54	GH65	PF03632	Maltose Phosphorylase	<i>Lactobacillus brevis</i>
	2.4.1.279	NGP1*			Nigerose Phosphorylase	<i>Clostridium phytofermentans</i>
N-Acylglucosamine-2- Epimerase	5.1.3.8	1FP3	N/A	PF07221	N-acylglucosamine-2- epimerase	<i>Sus scrofa</i>
		2GZ6			N-acylglucosamine-2- epimerase	<i>Anabaena</i> sp. CH1

Table 2. Structures of CAL/G superfamily members of known biochemical function used to obtain local chemical signatures of catalytically active residues for each functional family, and sorted into six functional families according to biochemical function.

Functional Family	E.C. #	PDB ID	CAZy	Pfam	Protein Name	Source
Xylanases	3.2.1.8	1M4W	GH 11	PF00457	β -1,4 Xylanase	<i>Thermopolyspora flexuosa</i>
		1H4G			Xylanase	<i>Bacillus agradae</i>
		1BCX			Xylanase	<i>Bacillus circulans</i>
Endoglucanases	3.2.1.4	1UU4	GH 12	PF01670	Endo- β -1,4-glucanase Cel12A	<i>Humicola grisea</i>
		1H8V			Endo- β -1,4-glucanase	<i>Trichoderma reesei</i>
		2NLR			Endoglucanase CelB2	<i>Streptomyces lividans</i>

Cellobiohydrolases	3.2.1.91	1Z3T	GH 7	PF00840	Cellobiohydrolase Cel7D	<i>Phanerochaete chrysosporium</i>
		1DY4			Cellobiohydrolase Cel7A	<i>Trichoderma reesei</i>
		2RFW			Cellobiohydrolase Cel7B	<i>Melanocarpus albomyces</i>
GH Family 16	3.2.1.-	2AYH	GH 16	PF00722	1,3-1,4- β -D-glucan glucanohydrolase	Hybrid <i>Bacillus</i>
		1DYP			κ -carrageenase	<i>Pseudoalteromonas carrageenovora</i>
		3ILF			Porphyranase A	<i>Zobellia galactanivorans</i>
		2VY0			Endo- β -1,3-glucanase	<i>Pyrococcus furiosus</i>
		1MVE			1,3-1,4- β -D-glucanase	<i>Fibrobacter succinogenes</i>
Lyases	4.2.2.-	1UAI	PL 7/18	PF08787	Aliginatase lyase	<i>Corymenacterium</i> sp.
		1J1T			Aliginatase lyase	<i>Alteromonas</i> sp. 272
		1VAV			Aliginatase lyase	<i>Pseudomonas aeruginosa</i>
Peptidases	3.4.23.-	2IFR	N/A	PF01828	Scytalidoglutamic peptidase	<i>Scytalidium lignicola</i>
		1Y43			Aspergilloglutamic peptidase	<i>Aspergillus niger</i>

2.2. Prediction of functional residues

Functionally important residues were predicted for proteins of known function as well as SG proteins using the POOL method [20, 21], with electrostatic and chemical properties from THEMATICS [31, 32], phylogenetic tree information from INTREPID [33, 34], and geometric features from ConCavity [35] as input (Figure 1, left).

2.3. Generation of SALSA tables

The protein structure comparison server PDBeFold [27, 36] was used to perform a multiple structure alignment (MSA) of all proteins of known function in the superfamily (Figure 1, right). For the members of the superfamily with previously characterized functions (Tables 1 and 2), the aligned spatial positions containing the top 12% of POOL-predicted residues were selected.

Each aligned position that contains residues predicted to be biochemically important for catalysis for the members of at least one functional family is labeled and the remaining aligned positions are discarded. These members of the superfamily are then sorted into groups according to the spatial arrangement of the local POOL-predicted residues and listed in a tabular format, known as SALSA tables (Tables 3 and 4). From these tables, a unique consensus signature,

defined as a set of predicted residues of similar amino acid type at aligned spatial positions for more than half of all members of a given functional family, can then be identified for each functional family. (For families with only two members, the consensus signature is considered to be the identical or similar aligned amino acids common to both members.) Then, a smaller table can be generated for each functional family, where the columns of each of these functional family tables represent only the aligned spatial positions predicted to be important for that functional family (see Supplementary Material).

2.4. Analysis of SG proteins of unknown or uncertain function

The predicted sets of functionally active residues for each SG protein are then aligned and compared with the consensus signatures for the different functional families, using the functional family SALSA tables shown in the Supplementary Material (and illustrated in Figure 1, right). The protein structure comparison server PDBeFold was also used here to perform a multiple structure alignment (MSA) of the SG proteins with the proteins of known function in the superfamily.

Fig. 1. Illustration of the POOL (left) and SALSA (right) methods. POOL is integrated into the SALSA method, as shown by the arrows between the two methods. Exo- α -1,6-mannosidase from *Clostridium perfringens* (PDB ID 3QT9), shown in light grey with its active site residues shown in red, is used to illustrate the POOL method on the left. In the SALSA method on the right, the following structures are

used to illustrate the method: glucoamylase from *Aspergillus awamori* (PDB ID 1GAI), shown in red; unsaturated glucuronyl hydrolase from *Bacillus sp. g11* (PDB ID 2DJ5), shown in purple; endoglucanase CelA from *Clostridium thermocellum* (PDB ID 1CEM), shown in teal; Exo- α -1,6-mannosidase from *Clostridium perfringens* (PDB ID 3QT9), shown in orange; N-acylglucosamine-2-epimerase from *Sus scrofa* (PDB ID 1FP3), shown in green; and an uncharacterized protein from *Bacteroides ovatus* (PDB ID 3ON6), shown in dark grey. (MSA: multiple structure alignment)

2.5. Scoring

In parallel with comparing the predicted residues for the SG proteins with the SALSA consensus signature tables, a scoring function is used to measure the quality of each local alignment. The similarity score (S_{ab}) for protein structures “a” and “b” is defined as:

$$S_{ab} = \frac{\sum_{x=1}^n M_{ax,bx}}{n} \quad (1)$$

where “x” is the column number in the SALSA table, representing an aligned spatial position, “n” is the number of columns, “ $M_{ax,bx}$ ” is the scoring matrix element, in this case from the BLOSUM62 matrix [37], for the two amino acids in proteins “a” and “b” at position “x”. All the S values are normalized to the best possible match (protein “a” aligned with itself) to give the normalized similarity (NS) score NS_{ab} , as:

$$NS_{ab} = \frac{S_{ab}}{S_{aa}} \quad (2)$$

The NS score serves as a metric for the degree of similarity between the active sites, defined as the set of aligned spatial positions of the top POOL-ranked residues of two proteins, or between the predicted active site of a query SG protein and the consensus signature residues of a particular functional family. NS scores close to or equal to 1 indicate a strong local active site similarity; smaller positive values indicate some similarity; zero or negative scores indicate a poor match. The spatial positions for the 6-HG superfamily (Table 1) were scored using Equations 1 and 2 and the scores are shown in Table S40. To establish consensus signatures for the different functional types, alignments are scored based on all spatial positions that have been

identified as important, as in Table S40. Once consensus signatures have been established, predicted active residues for proteins of unknown function are scored against each consensus signature, using only the spatial positions that are a part of the smaller table that includes only the functionally important positions for each individual functional family separately.

2.6. Active site similarities across different structural folds

When analyzing two proteins of different structural folds, of course it is not possible to obtain a global structural alignment of the two proteins. Therefore, the alignments of the residues in the two active sites are aligned manually using YASARA [28]. Then, we analyze the similarities and differences between the active sites of the two proteins of different global folds. This involves looking at not only the residue types to see which ones match, but looking at their local positions within the active site.

Table 3. Consensus signatures for the 13 functional families in the 6-HG superfamily. POOL-predicted residues are shown in uppercase font. The residues that have been shown to be catalytically important in the literature are underlined. The consensus signature residues are in **boldface**. At the top of each column, the location of that spatial position in the fold is indicated; helices are labeled H.

Functional Family	PDB ID	loop			H2			loop	H4	loop	H6		loop	H8		loop	H10		Loop			H12			
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
1,4- α -Glucan Glucosylhydrolase	1GAI	a39	Y48	W52	R54	D55	L58	q124	D126	W178	E179	F187	T188	a235	-	n247	Y311	W317	L319	E326	E400	I410	I415	W417	A421
	1AYX	-	Y63	W67	R69	D70	I73	q143	D145	W209	E210	F218	T219	E274	-	a290	Y351	W362	I364	Q371	E456	t466	I471	W473	A477
	1LF9	-	Y337	W341	R343	D344	H347	Q394	D396	W437	E438	A446	t447	r504	-	s535	Y581	W599	L601	m608	E636	T646	L652	W654	E658
	1UG9	-	Y326	W330	R332	D333	Q336	Q384	D386	W429	E430	S438	t439	R496	-	g527	Y573	W591	L593	E600	E628	G648	L653	W655	q659
Exo- α -1,6- Mannosidase	3QT9	g56	I58	W62	R64	D65	s68	E133	D135	S219	D220	p230	a231	-	a301	N302	H350	W357	i359	L366	E393	R405	F408	F408	W410
	3QSP	g54	I56	W60	R62	D63	a66	E132	D134	S217	D218	p228	s229	-	p301	N302	H350	W357	i359	L366	E393	R405	F408	F408	W410
Endoglucanase/ Xylanase/ Chitosanase	1CEM	R84	T88	V93	E95	g96	Y99	A150	D152	-	-	S214	Y215	p258	Y277	D278	-	n333	S335	a342	-	y368	G370	Y372	I376
	1WU4	D61	N64	R68	E70	g71	Y74	A126	D128	-	-	S197	Y198	p242	s262	D263	-	p320	g322	M329	-	g355	r358	Y360	L364
	1V5C	G106	T109	T120	E122	G123	Y126	A181	D183	-	-	S246	D247	s291	y318	N319	-	t377	v379	a386	-	-	R411	F413	Y417
	1H12	A66	d73	R76	E78	G79	W82	A142	D144	-	-	S202	Y203	t253	f280	D281	-	g341	g343	v350	-	g376	r379	Y381	I385
Cellulases	1KFG	D55	D58	F62	I64	P65	Y68	c153	a154	-	Y205	D212	D213	W254	g255	W258	W311	s313	R315	f322	-	E420	D424	Y425	g428
	1CLC	D198	D201	Y205	v207	N208	i211	t299	a300	-	Y354	d361	D362	-	d397	W400	W457	C459	g461	m468	-	E555	N559	W560	a563
	1KSC	D54	D57	F61	f63	P64	Y67	a152	g153	-	Y206	D213	E214	-	G250	W253	W301	t303	R305	f312	-	E412	D416	Y417	g420
	1IA6	D56	D59	F63	I65	p66	Y69	I155	s156	-	Y204	D211	D212	-	t253	W256	W304	v306	R308	m315	-	E410	D414	Y415	g418
Unsaturated Glucuronyl Hydrolase	2D5J	H26	v35	D41	W42	T43	f46	H86	D88	-	-	D149	C150	H193	R221	W225	-	d278	s280	C287	r335	d348	t350	w352	Y356
	2ZZR	P54	k62	E68	W69	T70	f73	H113	D115	-	-	D175	C176	H219	R247	W251	-	d304	s306	C313	h361	e374	n376	W378	Y382
α -L-Rhamnosidase	2OKX	D567	E572	W576	g578	d579	n582	I626	W629	D683	W684	h697	q698	d750	q765	v769	-	s800	f802	E809	E841	R858	H862	W864	a867
	3WSM	D630	E636	W640	G642	D643	F647	v692	W695	D746	W747	i759	a760	-	s813	I817	-	F851	P859	E895	n912	H916	a918	s921	
	ALR1*	D570	E575	W579	g581	D582	n585	i629	W632	D686	W687	h700	q701	d752	q767	v771	-	s802	f804	E811	E843	R855	H859	W861	a864
1,2- α -L-Fucosidase and α -L-galactosidase	4UFC	L353	w365	T370	N372	N374	M377	w441	s443	E501	H502	M511	D512	E560	H574	Y582	-	t611	W613	n620	H664	p666	Q668	D670	g674
	2EAC	L396	W414	H419	N421	N423	m426	s503	a505	E566	q567	y577	E578	E651	H678	F686	-	t720	W722	n729	H760	p762	Q764	D766	g770
	ALF1*	I414	w426	H431	n433	N435	M438	w503	m505	E563	h564	m582	d583	E632	H646	F654	-	w685	t687	A692	H734	p736	Q738	D740	G744
	ALF2*	I424	w436	H441	n443	N445	M448	p514	m515	E573	H574	m592	D593	e642	H656	F664	-	-	s696	s702	H744	P746	Q748	D750	g754
Trehalase	2JF4	p149	R152	Y157	W159	D160	f163	Q207	p209	W311	D312	d336	I337	-	-	Y410	Q446	W453	p455	E462	E496	q515	F518	W520	v524
	TRE1*	p178	r181	Y186	w188	D189	w192	Q236	p238	W334	D335	d363	I364	-	-	w437	Q479	w486	p488	m495	E532	q551	F554	W556	v560
Unsaturated Rhamnogalacturonyl Hydrolase	2D8L	-	-	H40	Y41	H42	V45	E86	D88	-	-	D143	G144	H189	R213	w217	r270	E276	s278	Y285	d328	s347	D350	H352	a356
α -Amylase	3REN	-	s70	L82	E84	S85	L88	A133	D136	n183	y184	C193	Y194	k235	L253	L254	t298	S303	S305	L312	-	-	Y355	F357	N361
Phosphorylase I	1V7X	R333	R343	g347	R349	D350	Q353	C424	D426	n491	D492	m503	v504	r556	-	E578	K636	F642	H644	v651	Q690	N704	T709	G710	w714
	2CQS	R351	R362	g365	R367	D368	Q371	F420	D422	N489	D490	f520	i521	r571	-	E593	K658	F664	H666	I673	Q712	k726	T731	G732	W736
	CDP1*	R331	g341	G345	R347	D348	Q351	Y405	D407	N472	D473	t486	t487	r537	-	n559	K617	F623	H625	i632	Q671	R685	T690	G691	W695

Phosphorylase II	1H54 NGP1*	P344 A334	Y352 Y342	Y357 F347	W358 W348	D359 D349	F363 F353	-	H427 H430	E487 E490	Y488 Y491	D497 Y500	t498 t501	-	Q593 Q605	W601 V613	-	S627 s639	L629 l641	H634 y646	N663 l681	D665 y683	H671 H688	T673 A690	g677 g694
N-Acylglucosamine- 2-Epimerase	1FP3 2GZ6	c41 c38	d50 d47	- -	W56 W53	l57 L54	R60 R57	-	F122 f115	-	-	A180 A172	V181 v173	E227 E218	G247 G238	H248 H239	-	W313 W303	W314 W304	E318 E308	Y361 Y351	l370 l360	g374 g364	K378 K368	H382 H372

Table 4. Consensus signatures for the six functional families in the CAL/G superfamily. POOL-predicted residues are shown in uppercase font. The residues that have been shown to be catalytically important in the literature are underlined. The consensus signature residues are in **boldface**. At the top of each column, an integer is assigned to each spatial position.

Functional Family	PDB ID	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	
Xylanases	1M4W	W20	d22	-	N46	V48	a49	Y74	I77	Y78	W80	-	-	-	E87	y88	Y89	i90	-	w94	-	-	-	-	-	r98	p99	t100	R121	P125	S126	-	f133	Q135	W137	-	s146	t150	-	i172	E176	y178	-
	1H4G	W19	d21	-	N45	L47	I48	Y81	v84	Y85	W87	-	-	-	E94	Y95	Y96	i97	-	w101	-	-	-	-	-	r105	p106	p107	R129	P133	S134	-	F141	Q143	W145	-	A180	s158	-	a180	E184	Y186	-
	1BCX	W9	d11	-	N35	V37	v38	Y65	I68	y69	w71	-	-	-	E78	y79	Y80	v81	-	w85	-	-	-	-	-	-	r89	p90	t91	R112	P116	s117	-	f125	Q127	W129	-	v168	t145	-	v168	E172	y174
Endo-glucanases	1U04	N22	W24	-	k62	Y64	p65	n99	y102	d103	F105	-	-	-	E120	I121	M122	i121	-	r127	-	-	-	-	-	y132	p133	i134	R115	-	-	-	M158	v160	s162	-	d172	d176	-	Q201	E202	F207	-
	1H8V	N20	W22	-	k58	Y60	q61	n95	Y98	d99	F101	-	-	-	E116	I117	M118	i119	-	k123	-	-	-	-	-	g128	p129	i130	N151	-	-	-	M154	v156	s158	-	n167	d171	-	Q196	E200	p201	-
	2NLR	N22	w24	-	k55	Y57	p58	n100	y103	d104	W106	-	-	-	E120	i121	M122	i123	-	r127	-	-	-	-	-	q132	p133	i134	N151	-	-	-	n158	v160	s162	-	g171	d175	-	Q199	E203	p204	-
Cellulobio-hydrolases	1Z3T	g23	c25	-	s103	R104	v105	g139	Y142	I143	a145	Y168	D170	S171	Q172	E207	m208	D209	i210	w211	E212	n215	-	-	a217	a219	t221	H223	-	c238	R240	-	d246	D248	D251	R256	t276	q280	v358	S362	W364	D366	
	1DY4	g23	c25	-	a106	R107	I108	g142	Y145	I146	s148	Y171	D173	S174	Q175	E212	m213	D214	i215	w216	E217	s220	-	-	s222	a224	t226	H228	-	c243	r251	-	D257	D259	D262	R267	v290	q293	v361	w367	D369		
	2RFW	g23	c25	-	a106	R107	I108	g142	Y145	I146	s148	Y171	D173	a174	Q175	E212	m213	D214	t215	w216	E217	a220	-	-	a222	a224	t226	H228	-	c243	R251	-	D257	D259	D262	R267	v288	r291	v360	S364	W366	D368	
GH Family 16	2AYH	e8	d22	N26	c61	E63	y64	S90	F92	Y94	i95	-	-	-	W103	d104	E105	i106	D107	I108	E109	g112	-	-	k117	Q119	N121	Y123	-	-	-	v127	g128	H130	E131	I147	w151	K178	i179	n182	w184		
	1DYP	d47	t61	-	e62	a114	a115	a142	W144	Y146	s147	-	-	-	Y161	s162	E163	i164	D165	-	E168	g171	-	K172	e177	D179	D181	H183	-	-	-	m195	R196	n204	h205	Y212	N252	I253	s256	g258			
	3ILF	d37	d51	W56	y52	a102	d103	f130	W131	R133	v134	-	-	-	I137	s138	E139	i140	D141	-	E144	g147	-	-	-	q161	H163	n165	H167	-	-	-	I175	Q176	I178	g179	Y195	w199	r228	m229	D232		
Lyase	2VY0	d52	k68	g71	s120	R122	I123	A152	W154	I156	g157	-	-	-	C168	g169	E170	i171	D172	-	E175	g178	-	-	-	T183	H185	T187	H189	-	-	-	i199	t200	a202	t204	f219	w223	Y259	i260	n263	a265	
	1MVE	-	-	-	g2	E42	I45	s31	F44	Y42	q53	-	-	-	W54	v55	E56	v57	D58	-	E60	g63	-	-	-	s68	Q70	N72	I74	-	-	-	Q81	t83	s84	E85	y101	w103	t105	I136	N139		
Peptidases	1UAI	t19	-	T65	R72	S73	E74	v114	Q117	H119	d120	-	-	-	D123	D124	v125	t126	v127	R129	t133	-	-	-	d141	t143	H145	-	-	-	-	-	-	-	-	E159	Y189	-	K191	Y195	Q197		
	1J1T	e21	n24	T45	R52	H53	E54	i87	Q90	H92	A93	-	-	-	g97	T98	i99	S100	K101	y103	f132	-	-	-	g133	e136	k137	-	-	-	-	-	-	-	n150	Y180	-	k182	Y186	Q188			
	1VAV	t10	p12	T52	R59	S60	E61	i99	Q102	H104	S105	-	-	-	p115	a114	P118	I116	v117	K118	i120	r124	r129	-	-	-	g140	A145	-	-	-	-	-	-	-	s159	Y193	-	K195	Y199	Q201		
Peptidases	2IFR	s4	-	-	g7	-	-	T37	w39	G41	d43	-	-	-	-	t49	a50	i57	-	I52	Q53	G55	D57	D65	W67	E69	Y71	P72	-	-	-	E136	D137	f138	-	-	-	-	-	-	-		
	1Y43	-	-	-	-	-	-	S8	W10	G12	D14	-	-	-	-	t20	-	-	-	-	Q24	G26	D28	D39	W41	E43	v45	P46	-	-	-	E110	D111	F112	-	-	-	-	-	-	-		

3. RESULTS AND DISCUSSION

3.1. Functional families in the 6-HG superfamily

Members of this superfamily are all- α proteins that share a common $(\alpha/\alpha)_6$ barrel structure. These proteins catalyze the hydrolysis of glycosidic bonds in poly- or oligosaccharides through a general acid-base mechanism; this catalysis occurs with the inversion of the anomeric carbon configuration [38, 39]. These enzymes have many industrial uses, such as in the food industry, in the paper industry, and for breaking down plant material for ethanol production; they are important components of biomass degradation pathways with potential applications for the production of biofuels.

Using the scores from the local structural alignments, the 6-HG superfamily was classified into 13 diverse functional families based on the 24 local active site positions. Each functional family was then realigned and rescored to obtain a more specific set of consensus signature residues. These predicted biochemically active residues, as well as detailed descriptions of the reactions catalyzed and the consensus signatures for each of the 13 functional families, are given in the Supplementary Material. Two functional families are highlighted below to serve as examples of our analysis on this superfamily. For the SG proteins in these functional families, misannotations are identified and functional assignments are made when possible.

3.1.1. *Exo- α -1,6-mannosidase functional family*

Members of this functional family belong to the GH125 family, process the terminal mannose molecule in N-linked glycans, and are specific to the α -1,6 glycosidic linkage [40]. The two known *exo- α -1,6-mannosidases* from *Clostridium perfringens* (PDB ID 3QT9) and *Streptococcus pneumoniae* (PDB ID 3QSP) were used to determine the 15 spatial positions that contain functionally important residues for this functional family (Table S7).

The NS score between these two proteins of known function is 1 (Table 5). Thirteen SG proteins were analyzed and compared to this functional family. Of the 13, only two showed strong matching at the local active site: an uncharacterized protein BACOVA_03626 from *Bacteroides ovatus* (PDB ID 3ON6) and BT3781 from *Bacteroides thetaiotaomicron* (PDB ID 2P0V). The sequences of these two SG proteins were submitted to the CombFunc [41] server and the structures to the ProFunc [16] server. While CombFunc was only able to reveal that the uncharacterized protein possessed “catalytic activity,” ProFunc was able to show that the closest structural matches were other uncharacterized proteins as well as probable matches to mannosidases. ProFunc also reported that there is a low probability hit against the library of enzyme active site templates to a glucoamylase (1,4- α -glucan glucohydrolase from *Aspergillus awamori*) [42] and a possible ligand binding template hit to a glucoamylase. Furthermore, strong matches by reverse templates, i.e. templates generated from the query structure [16], were with proteins of unknown function and with exo- α -1,6-mannosidases. As for protein BT3781 (PDB ID 2P0V), the ProFunc server revealed mostly matches to uncharacterized proteins. However, there were strong matches by reverse templates to known exo- α -1,6-mannosidases, similar to the uncharacterized SG protein. In the CAZy database, both SG proteins are listed in the GH125 family; this family was first reported in 2011 [40] and exo- α -1,6-mannosidase (E.C. 3.2.1.-) is the only known activity in the family to date. For the structure annotations in CAZy, no E.C. number is given for the uncharacterized protein BACOVA_03626 (PDB ID 3ON6). On the other hand, BT3781 (PDB ID 2P0V) is listed as exo- α -1,6-mannosidase and the E.C. number is given as E.C. 3.2.1.-.

Table 5. NS scores for known exo- α -1,6-mannosidases (PDB IDs 3QT9 and 3QSP) and two SG proteins of unknown function (PDB IDs 3ON6 and 2P0V) using the consensus signature of the exo- α -1,6-mannosidase functional family (Table S7). PDB IDs for the SG proteins are shown in italics.

	3QT9	3QSP	<i>3ON6</i>	<i>2P0V</i>
3QT9	1.00	0.91	0.91	0.91

3QSP	0.91	1.00	0.87	0.87
3ON6	0.91	0.87	1.00	1.00
2P0V	0.91	0.87	1.00	1.00

Fig. 2. Structurally aligned residues in key spatial positions for the two α -1,6-mannosidase of known function from *Clostridium perfringens* (PDB ID 3QT9, red) and from *Streptococcus pneumoniae* (PDB ID 3QSP, orange) with an SG protein of unknown function from *Bacteroides ovatus* (PDB ID 3ON6, blue). Note: the SG protein from *Bacteroides thetaiotaomicron* (PDB ID 2P0V, not pictured) has identical residue types and numbers as the SG protein shown here (see Table S41).

The two SG proteins have a NS score of 0.87 or 0.91 with the consensus signature for the α -1,6-mannosidase functional family as shown in Table 5. When compared to the other 12 functional families, both SG protein have a NS score around 0.30 with the glucoamylases, around 0.20 with the trehalases, and below 0.10 with all other functional families (Table S42). Based on the similarity scores and the excellent alignment at the local site of biochemical activity, as observed in Figure 2, we assign α -1,6-mannosidase function to these two SG proteins.

3.1.2. 1,2- α -L-fucosidase and α -L-galactosidase family

Members of this functional family belong to the GH95 family and hydrolyze glycosidic linkages of α -1,2-linked fucose and galactose [43]. Since there are only structures for two proteins of known function in this functional family, two homology models were made from

sequences of known 1,2- α -L-fucosidase function, using the known *Bifidobacterium bifidum* structure (PDB ID 2EAC) as the main template. The two model structures are for 1,2- α -L-fucosidase from *Arabidopsis thaliana* and from *Lilium longiflorum* and are labeled as ALF1 [44] and ALF2 [45] respectively, as shown in Table 1.

Alignment of the four structures revealed 11 spatial signature positions represented in Table S22. The four proteins of known function have high NS scores against each other (Table 6). The following spatial position designations refer to those in Table 3, and the residue numbering refers to PDB ID 2EAE (which is the same protein as PDB ID 2EAC, but with different ligands). In this functional family, the glutamic acid residue in spatial position nine (E566) acts as the proton donor, or the catalytic acid. As shown in Figure 3, the hydrogen bonds between the active site residues of fucosidase (shown in green) with fucose (shown in orange) and those of the galactosidase (shown in pink) with galactose (shown in light blue) are very similar. The major difference is seen between the residues in position three of Table 3. The threonine in the galactosidase (T370) interacts with the O6 atom of the galactose. Since the fucose does not have an O6 atom, a bulky residue sidechain, such as histidine (H419), can fit into this spatial position to help stabilize and bind the fucose [43, 46].

Fig. 3. Close up view of the active site pocket of α -L-galactosidase from *Bacteroides ovatus*, shown in pink (PDB ID 4UFC) with bound galactose in light blue, and 1,2- α -L-fucosidase from *Bifidobacterium bifidum*, shown in green (PDB ID 2EAE) with bound fucose in orange. The predicted hydrogen bonds

between galactose and the residues of α -L-galactosidase are shown in black, while the hydrogen bonds between fucose and the residues of 1,2- α -L-fucosidase are shown in grey.

Of the 13 SG proteins analyzed against this functional family, one showed a significant match at the local active site: putative glycoside hydrolase from *Bacillus halodurans* (BH0842, PDB ID 2RDY). Figure 4 shows the alignment of the predicted active site residues of the known α -L-galactosidase from *Bacteroides ovatus* (PDB ID 4UFC [46]) and α -L-fucosidase from *Bifidobacterium bifidum* (PDB ID 2EAC) with this SG protein. Note that the SG protein BH0842 has 100% matching of the predicted active site residues when compared to the known galactosidase. The tyrosine in the SG protein (Y595) aligns with a tyrosine in the known galactosidase (Y582). However, the fucosidase has a phenylalanine in this position. This residue is considered a distal residue [23], but our method predicts it to be important. Interesting to note, submission of the BH0842 sequence to the CombFunc server [41] predicts fucosidase activity with relatively high certainty (0.933 SVM probability); submission of the BH0842 structure (PDB ID 2RDY) to the ProFunc server [16] also predicts fucosidase activity. In the CAZy database, this SG protein is listed in the GH95 family, which lists three different known activities within the family: α -L-fucosidase (E.C. 3.2.1.51), α -1,2-L-fucosidase (E.C. 3.2.1.63), and α -L-galactosidase (E.C. 3.2.1.-). Under the structure tab, BH0842 does not have an accompanying E.C. number listed; thus the specific activity within the family is not defined in CAZy for this SG protein. This is one example where our method provides a more detailed prediction, including substrate specificity, than the more conventional bioinformatics techniques.

The NS score for the SG protein BH0842 with the consensus signature of the members of this family is 0.76 or 0.81 with the fucosidases and 1.00 with the galactosidase, as shown in Table 6; its NS score is zero or negative against the other 12 functional families (Table S42). These values suggest probable α -L-galactosidase activity and therefore we tentatively assign this

SG protein this function. For this case our method not only has assigned function to an SG protein, but also has predicted substrate specificity.

Table 6. NS scores for known 1,2- α -L-fucosidases (PDB ID 2EAC and models ALF1 and ALF2) and a putative glycoside hydrolase from SG (PDB ID 2RDY) using the consensus signature for the α -L-fucosidase functional family (Table S22).

	4UFC	2EAC	ALF1	ALF2	2RDY
4UFC	1.00	0.77	0.83	0.83	1.00
2EAC	0.77	1.00	0.96	0.96	0.76
ALF1	0.83	0.96	1.00	1.00	0.81
ALF2	0.83	0.96	1.00	1.00	0.81
2RDY	1.00	0.76	0.81	0.81	1.00

Fig. 4. Structurally aligned residues in key spatial positions for the known α -L-galactosidase from *Bacteroides ovatus* (PDB ID 4UFC, red) and known 1,2- α -L-fucosidase from *Bifidobacterium bifidum* (PDB ID 2EAC, orange) with those of a putative glycoside hydrolase BH0842 from *Bacillus halodurans* (PDB 2RDY, blue). POOL-predicted residues are in uppercase; residues not predicted by POOL are in lowercase.

3.1.3. SG proteins that could not be reliably assigned

Three SG proteins from *Bacteroides thetaiotaomicron* (PDB ID 3K11), *Klebsiella pneumoniae* (PDB ID 3PMM), and *Salmonella enterica* (PDB ID 3QWT) show small, positive NS scores with both the unsaturated glucuronyl hydrolase and the unsaturated rhamnogluconyl hydrolase functional families (Table 7) and negative or zero scores with the other 11 functional

families (Table S42). Two of these SG proteins (PDB IDs 3PMM and 3QWT) are listed as possible GH105 family proteins, while the other (PDB ID 3K11) is listed simply as a putative glycosyl hydrolase.

Table 7. NS scores for the two unsaturated glucuronyl hydrolases (PDB IDs 2D5J and 2ZZR), a known unsaturated rhamnogalacturonyl hydrolase (PDB ID 2D8L), and three SG proteins (PDB IDs 3K11, 3PMM, and 3QWT) using the consensus signature for both of the unsaturated hydrolase functional families (Tables S16 and S28).

	2D5J	2ZZR	2D8L	3K11	3PMM	3QWT
2D5J	1.00	0.92	0.22	0.22	0.40	0.40
2ZZR	0.92	1.00	0.24	0.20	0.42	0.42
2D8L	0.22	0.24	1.00	0.59	0.41	0.41
3K11	0.22	0.20	0.59	1.00	0.45	0.45
3PMM	0.40	0.42	0.41	0.45	1.00	1.00
3QWT	0.40	0.42	0.41	0.45	1.00	1.00

Table 7 shows that the two possible GH105 family proteins (PDB IDs 3PMM and 3QWT) could be assigned either glucuronyl hydrolase or rhamnogalacturonyl hydrolase function. However, these scores suggest that rhamnogalacturonyl hydrolase function is more probable than glucuronyl hydrolase for the putative glycosyl hydrolase (PDB ID 3K11). Since there is only one structure available for the unsaturated rhamnogalacturonyl hydrolase functional family, we regard these functional assignments as tentative at best.

Putative α -rhamnosidase from *Bacteroides thetaiotaomicron* (PDB ID 3CIH) has an NS score of 0.62 or 0.83 with each of the three proteins of known function in the α -L-rhamnosidase functional family (Table 8) and negative NS scores with the other functional families (Table S42). Similarly, two other SG proteins, putative NAG isomerase from *Salmonella typhimurium* (PDB ID 2AFA) and an N-acetylglucosamine-2-epimerase family protein from *Xylella fastidiosa* (PDB ID 3GT5), have an NS score of 0.54 or 0.56 with the proteins of known function in the 2-acetylglucosamine-2-epimerase functional family (Table 9) and negative NS scores with the

other functional families (Table S42). In these instances, more information is needed to determine the function of these SG proteins. However, it is possible they are correctly annotated since they show mediocre to good matches at the local active site compared to the functional families to which they are putatively assigned (Tables 8 and 9).

Table 8. NS scores for three known α -L-rhamnosidases (PDB IDs 2OKX, 3W5M, and ALR1) and putative α -rhamnosidase (PDB ID 3CIH) using the consensus signature for the α -L-rhamnosidase functional family (Table S19).

	2OKX	3W5M	ALR1	3CIH
2OKX	1.00	0.72	1.00	0.83
3W5M	0.72	1.00	0.76	0.62
ALR1	1.00	0.76	1.00	0.83
3CIH	0.83	0.62	0.83	1.00

Table 9. NS scores for two known the N-acetylglucosamine-2-epimerases (PDB IDs 1FP3 and 2GZ6) and two SG proteins: putative NAG isomerase (PDB ID 2AFA) and N-acetylglucosamine-2-epimerase family protein (PDB ID 3GT5) using the consensus signature for the N-acetylglucosamine-2-epimerase functional family (Table S38).

	1FP3	2GZ6	2AFA	3GT5
1FP3	1.00	1.00	0.56	0.56
2GZ6	1.00	1.00	0.56	0.56
2AFA	0.56	0.56	1.00	0.54
3GT5	0.56	0.56	0.54	1.00

Three other SG proteins, mannosidase from *Bacteroides thetaiotaomicron* (PDB ID 4C1S), putative glycosyl hydrolase from *Bacteroides thetaiotaomicron* (PDB ID 4MU9), and Lin0763 protein from *Listeria innocua* (PDB ID 3K7X) have poor NS scores (negative or zero) with all previously characterized functional families in this superfamily (Table S42). However, they do have positive NS scores in the local structural alignment with each other (Table S42) and have been classified into the GH76 family [47]; this may indicate that they share some similarity of function and might constitute a new functional family. In addition, the final two SG proteins, putative glucoamylases from *Bacteroides fragilis* (PDB ID 3EU8) [18] and *Bacteroides uniformis* (PDB ID 4GL3) have negative NS scores with the glucoamylase functional family, indicating that the putative annotations are incorrect. Negative or nearly zero scores are also

obtained with the other 12 functional families. However, they have a positive NS score with each other. This could indicate another new functional family addition to the superfamily.

3.2. Functional families in the CAL/G superfamily

Members of this superfamily are all- β proteins consisting of 12-14 anti-parallel β -strands arranged in two curved sheets. These enzymes play an important role in many biological processes, including biogenesis, cellular development and localization, and macromolecule metabolic processes [48, 49]. In similar fashion to the 6-GH superfamily, the CAL/G superfamily was sorted into six diverse functional families based on 41 local active site positions. Within each respective functional family, members were realigned and rescored to obtain a more specific set of residues important for catalysis. More information on the functional families can be found in the Supplementary Material. For illustrative example, some SG proteins from this superfamily are discussed below.

3.2.1. GH Family 16

Members of the GH Family 16 generally cleave the β -1,3 or β -1,4 glycosidic bonds in various types of glucans. There are many substrate specificities found in this functional family, including but not limited to κ -carrageenases, xyloglucanases, and various endoglucanases. As seen in Table 4, some spatial positions (nine, 14, 25, and 27) contain different residue types across the functional family, but most are predicted by POOL to be important for catalysis. However, the main catalytic residues, two glutamic acids in spatial positions 16 and 20 and an aspartic acid in spatial position 18, are present in all members of the superfamily. These residues participate in the hydrolysis of polycarbohydrates with similar catalytic mechanisms. The lack of substrate specificity may account for the lack of conservation of residues in positions associated with substrate recognition across this functional family. The variability of POOL-predicted

residues in a given aligned position may be indicative of differing substrate specificities. This difference illustrates how the SALSA method is able to predict not only the catalytic residues, but the residues involved in substrate binding as well.

Five SG proteins were analyzed against this functional family GH family 16 and the other functional families in this superfamily as well. One SG protein, MSMEG_5345, was originally annotated as a putative GH family 16 protein from *Mycobacterium smegmatis* (PDB ID 3RQ0). In the CAZy database, this SG protein is listed as a member of the GH16 family. However, there are numerous known activities associated with this family, including, but not limited to, endo-1,3- β -glucanase (E.C. 3.2.1.39), κ -carrageenase (E.C. 3.2.1.83), and xyloglucanase (E.C. 3.2.1.151); no specific functional assignment or E.C. number is listed for this protein. Therefore, while CAZy identifies the family to which this SG protein belongs, it does not provide any further information on the specific activity for this SG protein. When this protein was scored against the consensus signatures of the members of the GH family 16 functional family, the NS scores ranged from 0.13 to 0.49, as shown in Table 10. While these values may seem a bit low compared to the examples from the 6-HG superfamily, they are within range of the values seen across the known members of the functional family (0.18-0.86). This wide range is most likely due to the various substrate specificities of the members of this family as mentioned before. When this SG protein was scored against the other functional families, NS scores were negative, indicating a poor match at the local active site with those functional families. Figure 5 shows the alignment of the SG protein to two representative members of the GH Family 16 functional family. Based on the alignment, it is apparent that the known catalytic residues (Glu-Asp-Glu) are present in the SG protein and align perfectly in the spatial positions across the functional family. In the SG protein, T167 aligns in a spatial position containing POOL-predicted charged

residues from the other proteins. Aromatic residues W142 and H169 are also shown in Figure 5 to align with the spatial positions containing other bulky and/or aromatic residues that are POOL-predicted. Therefore, the matching of key residues between the local active site of this functional family and SG protein lead to the prediction that this SG protein is plausibly a GH Family 16 protein and the putative annotation is likely correct.

Even though this annotation is likely correct, it only provides tentative information on the specific function of this SG protein. The best specific functional match from this functional family to this SG protein based on the consensus signatures (Table 10), and the residues across the aligned, predicted positions for the entire superfamily (Table S63), is to the endoglucanase function. The NS score between this SG protein and the endoglucanase (PDB ID 2VY0) using the consensus signature for this functional family is 0.49. There are two other NS scores similar to this one: κ -carrageenases (PDB ID 1DYP), 0.44; porphyranase A (PDB ID 3ILF), 0.47. While these NS scores lead us to believe that this SG protein can possess any of these three functions, Table S63, which compares active site matches across a larger set of key residues, indicates that the closest functional match based on NS scores is the endoglucanase: the SG protein scores a 0.41 with the endoglucanase and under 0.20 with the other proteins in this functional family. The expanded SALSA table takes into account residue positions involved in substrate specificity. We predict that this protein possesses endoglucanase activity, which is more specific than the classification currently available.

Table 10. NS scores for known 1,3-1,4- β -D-glucan glucanohydrolase (PDB ID 2AYH), κ -carrageenase (PDB ID 1DYP), porphyranase A (PDB ID 3ILF), endo- β -1,3-glucanase (PDB ID 2VY0), and 1,3-1,4- β -D-glucanase (PDB ID 1MVE) compared to an SG protein annotated as a putative GH16 protein (PDB 3RQ0) using the consensus signature for the GH Family 16 functional family (Table S53).

	2AYH	1DYP	3ILF	2VY0	1MVE	3RQ0
2AYH	1.00	0.50	0.41	0.28	0.86	0.21
1DYP	0.50	1.00	0.59	0.49	0.39	0.44

3ILF	0.41	0.59	1.00	0.65	0.32	0.47
2VY0	0.28	0.49	0.65	1.00	0.18	0.49
1MVE	0.86	0.39	0.32	0.18	1.00	0.13
3RQ0	0.21	0.44	0.47	0.49	0.13	1.00

Fig. 5. Structurally aligned residues in key spatial positions for porphyranase A from *Zobellia galactanivorans* (PDB ID 3ILF, red), endo- β -1,3-glucanase from *Pyrococcus furiosus* (PDB ID 2VY0, orange), and SG protein putative GH Family 16 protein from *Mycobacterium smegmatis* (PDB ID 3RQ0, blue). The two known representatives shown correspond to the two enzymes from the GH Family 16 functional family that had the highest NS scores against the SG protein (see Table 8).

3.2.2. SG proteins that could not be reliably assigned

Four other SG proteins were analyzed against the functional families in the CAL/G superfamily as well: two putative glycosyl hydrolases from *Parabacteroides distasonis* and *Bacteroides thetaiotaomicron* (PDB IDs 3HBK and 3OSD, respectively) and two putative sugar hydrolases from *Parabacteroides distasonis* and *Bacteroides ovatus* (PDB IDs 3H3L and 3NMB, respectively). Sequence and structure similarity searches, using Blast and Dali, of the putative glycosyl hydrolases revealed that the top hits were SG proteins or proteins of unknown/uncertain function. Therefore, simple transfer of function based on sequence or structure only does not provide valuable functional information.

Next, these four SG proteins were aligned and scored against the functional families in the CAL/G superfamily. For the putative glycosyl hydrolase from *Parabacteroides distasonis* (PDB ID 3HBK), the predicted residues did not align well with any of the consensus signatures. It is interesting to note, however, when the positions of the POOL-predicted residues K229, Y235, and H254 are mirrored across both the x-axis and y-axis in Figure 6, these symmetry-related positions of the residues align with the catalytic residues R72, Y195, and H119 of an alginate lyase (PDB ID 1UAI). While alignment of three residues is not sufficient to predict a definitive function, it provides some information about a possible function for this SG protein.

Fig. 6. Analysis of three residues in putative glycosyl hydrolase from *Parabacteroides distasonis* (PDB ID 3HBK, green) and alginate lyase from *Corymenacterium* sp. (PDB ID 1UAI, pink). The x-axis and y-axis are shown only to illustrate the symmetry operations on the SG protein to align the residues.

In three of the SG proteins (PDB IDs 3OSD, 3NMB, and 3HBK), a pair of Glu-Gln residues are located ~ 4.5 Å apart: E164-Q273, E162-Q271, and E145-Q252, respectively. This is similar to the Glu-Gln catalytic dyad in the peptidase functional family. Although this is not sufficient to assign function confidently, this information might provide a clue for functional determination. The last SG protein (PDB ID 3H3L) possesses potential catalytic residues at key spatial positions as well. Residues E232, Y168, and D153 align with spatial positions 14 and 39 of the xylanases and the endoglucanases functional families, both of which contain aligned glutamic acid residues across their functional family. However, similar to the argument for

YP_001304206.1 [19], the lack of other aligned spatial positions makes the annotation of this SG protein as a GH11 or GH12 protein weak.

Unfortunately after scoring these four SG proteins against all other functional families, the NS scores were zero or negative. This suggests that these SG proteins do not possess the same function as any of the functional families in this superfamily. The information gathered on these SG proteins can be helpful in determining their function in that it can rule out unlikely functions, but cannot be used to assign a likely function.

3.3. Similar active sites across different structural folds

Often, a superfamily is defined as a set of proteins with a similar structural fold and function. However, there are cases where structural folds are found to include a variety of functional types, as in the TIM barrel and Rossmann folds [50, 51]. In addition, enzymes of similar function can have different structures. For example, two glycosidases with the same E.C. numbers but different structural folds are both analyzed at the local active site using the SALSA method. Both Lam55A β -1,3-glucanase from *Phanerochaete chrysosporium* (PDB ID 3EQN) and exo- β -1,3-glucanase from the human pathogen *Candida albicans* (PDB ID 1CZ1) have an E.C. number of 3.2.1.58. However, the structure of the Lam55A β -1,3-glucanase from GH family 55 is a right-handed β -helix fold, while the structure of the exo- β -1,3-glucanase from GH family 5 is a $(\beta/\alpha)_8$ TIM barrel fold, shown in Figure 7A. These two folds are drastically different, but five of the residues at each local active site align perfectly with each other. This illustrates how enzymes that catalyze the same types of reactions, but have different folds, can still possess the same residue types in the same spatial positions within their respective active sites. The residues D575, H576, E610, H611, and E633 in the Lam55A β -1,3-glucanase active site align with the residues D251, H252, E192, H226, and E292, respectively, in the exo- β -1,3-

glucanase active site. These residues in Lam55A are in the second of two domains; the three acidic residues (two glutamic acids and one aspartic acid) are highly conserved among GH55 family members. E633 acts as the catalytic acid, and these three conserved acidic residues form hydrogen bonds with the ligand [52]. Similarly, the E192 residue acts as the proton donor and E292 acts as the nucleophile [53]. In both instances, the other remaining residues shown in Figure 7B could arguably be involved in stabilizing the active site pocket itself or in stabilizing a ligand in the active site pocket.

Fig. 7. A. The overall global folds of Lam55A β-1,3-glucanase from *Phanerochaete chrysosporium* (PDB ID 3EQN, purple) and exo-β-1,3-glucanase from the human pathogen *Candida albicans* (PDB ID 1CZ1, teal) show the drastic differences between the two protein structures that catalyze the same reaction. **B.** The catalytic residues of these two proteins align perfectly in terms of both the residue type and their spatial positions within the active sites.

In another example, two endoglucanases possessing the E.C. number 3.2.1.4 are studied at the local active site. First, hyperthermophilic endoglucanase Cel5A from *Thermotoga maritima* (PDB ID 3AMC), a member of the GH5 family, is shown in Figure 8A as having a global structural fold similar to the TIM barrel fold. On the other hand, xyloglucanase from

Bacillus licheniformis (PDB ID 2JEN), a member of the GH12 family, also shown in Figure 8A, has a global structural fold of a “jelly roll,” or a mainly β -sheet sandwich with a single α -helix on the outer surface. As seen in Figure 8A, these two structures cannot be aligned. However, comparing the local active sites of the two proteins reveals a few key residues types that align perfectly, as shown in Figure 8B. First, the catalytic acid/base from each protein, E136 in Cel5A and E243 in the xyloglucanase, align spatially. Second, E132 from Cel5A, which is located just outside the active site, is aligned with catalytically important E155 from the xyloglucanase. Finally, the catalytic E253 from Cel5A aligns with D137 of the xyloglucanase; this residue is spatially located near the nucleophilic glutamic acid and participates in hydrogen bonding with the ligand. Additionally, two asparagine residues (N20 from Cel5A and N51 from the xyloglucanase) and two phenylalanine residues (F254 from Cel5A and F240 from the xyloglucanase) show excellent local spatial alignment within the two active sites. These residues in their respective active sites are involved in stabilizing, stacking, and hydrogen bonding with the oligosaccharide ligands.

Fig. 8. A. The overall global folds of endoglucanase Cel5A from *Thermotoga maritima* (PDB ID 3AMC, purple) and a xyloglucanase from *Bacillus licheniformis* (PDB ID 2JEN, teal) show the dramatic differences between the two protein structures that catalyze the same reaction. **B.** The key catalytic residues align well in terms of the residue type and their spatial positions within the active sites. In addition, other residues involved in stability of the active site also align well.

4. Conclusion

A strategy has been developed, based on local structure at the site of biochemical activity, to predict function and to identify misannotations. The chemical signatures, as arranged in SALSA tables, associated with protein function are used to sort the superfamily according to biochemical function and also to determine whether SG protein structure members are correctly or incorrectly annotated. Using this method, we have been able to successfully characterize the 6-HG superfamily and provide functional information for a few of the SG proteins within the superfamily.

The members of the 6-HG superfamily have been classified according to their biochemical function using the SALSA method, which analyzes the local-structure-based consensus signatures for each functional family. These consensus signatures were used to annotate SG protein members of the superfamily as well. From this analysis of 13 SG proteins, two proteins of unknown function, an uncharacterized protein from *Bacteroides ovatus* (PDB ID 3ON6) and BT3781 from *Bacteroides thetaiotaomicron* (PDB ID 2P0V), were predicted to possess exo- α -1,6-mannosidase function and one putative glycoside hydrolase BH0842 from *Bacillus halodurans* (PDB ID 2RDY) showed a good local structure match with the known α -L-galactosidase and is predicted to have that function. In the case of BH0842, we predict that the SG protein will be specific for galactose over fucose based on its local active site residues. Of the ten remaining SG proteins, three show positive NS scores with the glucuronyl hydrolase or

rhamnoglucuronyl hydrolase functional families, while the predicted local active sites of most of the others suggest possible new functional families within the 6-HG superfamily.

Members of the CAL/G superfamily have been classified in to six functional families using the SALSA method. The consensus signatures generated were used to annotate one SG protein (PDB ID 3RQ0) as a GH Family 16 protein, most likely possessing endoglucanase activity. Even though the NS scores and consensus signatures did not show a strong enough match to annotate the other four SG proteins, they were able to provide information about the local active sites that was not known before.

A key advantage of the method described herein over other local structure based methods is the ability of the POOL method [20, 21] to predict a wider set of functionally important residues that characterize the biochemical active site than is currently available in other databases. The SALSA consensus signature tables established herein for the functional families of the 6-HG superfamily consist of nine to 16 residue-positions per functional family, about three to four times the size of a typical set of residues listed in the Catalytic Site Atlas [24]. Most literature-reported residues that are directly involved in catalysis and ligand binding appear in the top percentile of POOL-predicted residues. However, it has also been demonstrated recently [23, 54-56] that residues in supporting roles, often not reported in the literature, are also important for both reactivity and specificity; knowledge of these residues enables better functional annotation.

When studying two proteins of different global structural folds, we have shown that analyzing the similarities between the two proteins' active sites can provide a great deal of information about their function. Two glycosidases with the same E.C. numbers but completely different global structures possess the same residue types within the active site. Similarly, two endoglucanases possess the same E.C. numbers, but they have distinctly different global

structural folds. However, a closer look at the catalytic residues and the residue types in the active site pocket show that a good local alignment of the active site residues exists. Using this SALSA method to analyze the local active sites of similarly functioning enzymes with vastly different global structural folds can provide insight into the function of other SG proteins. For example, if an SG protein with the TIM barrel fold does not align with the functional families in one superfamily, the local structure at the predicted active site can be compared to those of other structures in different superfamilies to determine if there are likely similarities in biochemical function.

In some cases, the method can be applied to protein sequences of unknown function, if the sequence is amenable to homology modeling. The input features to POOL have been shown to predict functional residues effectively for comparative model structures [57].

The quality of the SALSA predictions depends on the set of proteins of known function and on the availability of structural and biochemical data for these benchmark proteins. As these data sets expand, our ability to make more accurate and more specific functional predictions is improved.

The SALSA method has the ability not only to identify reaction types, but also distinguish between functional families based on substrate specificity within this class of proteins with increasingly important applications [58]. Local structures of active site residues, based on computed chemical and electrostatic properties, constitute a very powerful tool that can be used to uncover valuable functional information about newly discovered SG proteins. Information about the function of SG proteins greatly increases the value of Structural Genomics data.

Acknowledgements

This work is supported by the National Science Foundation under grant numbers MCB-1158176 and CHE-1305655 (PJB, MJO), the American Cancer Society under Research Scholar Grant RSG-12-161-01-DMC (PJB), MathWorks, and the PhRMA Foundation (fellowship to CLM). We also thank Dr. Srinivas Somarowthu and Dr. Pengcheng Yin for their kind assistance.

References

- [1] Westbrook, J., Z. Feng, L. Chen, H. Yang, and H.M. Berman, The Protein Data Bank and structural genomics, *Nucleic Acids Res.* 31 (2003) 489-91.
- [2] Schnoes, A.M., S.D. Brown, I. Dodevski, and P.C. Babbitt, Annotation error in public databases: misannotation of molecular function in enzyme superfamilies, *PLoS Comput Biol.* 5 (2009) e1000605.
- [3] Watson, J.D., R.A. Laskowski, and J.M. Thornton, Predicting protein function from sequence and structural data, *Curr Opin Struct Biol.* 15 (2005) 275-84.
- [4] Lee, D., O. Redfern, and C. Orengo, Predicting protein function from sequence and structure, *Nat Rev Mol Cell Biol.* 8 (2007) 995-1005.
- [5] Gherardini, P.F. and M. Helmer-Citterich, Structure-based function prediction: approaches and applications, *Brief Funct Genomic Proteomic.* 7 (2008) 291-302.
- [6] Sleator, R.D. and P. Walsh, An overview of in silico protein function prediction, *Arch Microbiol.* 192 (2010) 151-5.
- [7] Chi, X. and J. Hou, An iterative approach of protein function prediction, *BMC Bioinformatics.* 12 (2011) 437.
- [8] Mills, C.L., P.J. Beuning, and M.J. Ondrechen, Biochemical functional predictions for protein structures of unknown or uncertain function, *Comput Struct Biotechnol J.* 13 (2015) 182-91.
- [9] Holm, L., S. Kaariainen, C. Wilton, and D. Plewczynski, Using Dali for structural comparison of proteins, *Curr Protoc Bioinformatics.* Chapter 5 (2006) Unit 5 5.
- [10] Holm, L. and P. Rosenstrom, Dali server: conservation mapping in 3D, *Nucleic Acids Res.* 38 (2010) W545-9.
- [11] Shindyalov, I.N. and P.E. Bourne, Protein structure alignment by incremental combinatorial extension (CE) of the optimal path, *Protein Eng.* 11 (1998) 739-47.
- [12] Ilyin, V.A., A. Abyzov, and C.M. Leslin, Structural alignment of proteins by a novel TOPOFIT method, as a superimposition of common volumes at a topomax point, *Protein Sci.* 13 (2004) 1865-74.
- [13] Ivanisenko, V.A., S.S. Pintus, D.A. Grigorovich, and N.A. Kolchanov, PDBSiteScan: a program for searching for active, binding and posttranslational modification sites in the 3D structures of proteins, *Nucleic Acids Res.* 32 (2004) W549-54.
- [14] Meng, E.C., B.J. Polacco, and P.C. Babbitt, Superfamily active site templates, *Proteins.* 55 (2004) 962-76.
- [15] Binkowski, T.A., A. Joachimiak, and J. Liang, Protein surface analysis for function annotation in high-throughput structural genomics pipeline, *Protein Sci.* 14 (2005) 2972-81.
- [16] Laskowski, R.A., J.D. Watson, and J.M. Thornton, ProFunc: a server for predicting protein function from 3D structure, *Nucleic Acids Res.* 33 (2005) W89-93.
- [17] Shulman-Peleg, A., R. Nussinov, and H.J. Wolfson, SiteEngines: recognition and comparison of binding sites and protein-protein interfaces, *Nucleic Acids Res.* 33 (2005) W337-41.
- [18] Parasuram, R., J.S. Lee, P. Yin, S. Somarowthu, and M.J. Ondrechen, Functional classification of protein 3D structures from predicted local interaction sites, *J Bioinform Comput Biol.* 8 Suppl 1 (2010) 1-15.

- [19] Wang, Z., P. Yin, J.S. Lee, R. Parasuram, S. Somarowthu, and M.J. Ondrechen, Protein function annotation with Structurally Aligned Local Sites of Activity (SALSAs), *BMC Bioinformatics*. 14 Suppl 3 (2013) S13.
- [20] Tong, W., Y. Wei, L.F. Murga, M.J. Ondrechen, and R.J. Williams, Partial order optimum likelihood (POOL): maximum likelihood prediction of protein active site residues using 3D Structure and sequence properties, *PLoS Comput Biol*. 5 (2009) e1000266.
- [21] Somarowthu, S., H. Yang, D.G. Hildebrand, and M.J. Ondrechen, High-performance prediction of functional residues in proteins with machine learning and computed input features, *Biopolymers*. 95 (2011) 390-400.
- [22] Naumoff, D.G., Hierarchical classification of glycoside hydrolases, *Biochemistry (Mosc)*. 76 (2011) 622-35.
- [23] Brodtkin, H.R., N.A. DeLateur, S. Somarowthu, C.L. Mills, W.R. Novak, P.J. Beuning, D. Ringe, and M.J. Ondrechen, Prediction of distal residue participation in enzyme catalysis, *Protein Sci*. 24 (2015) 762-78.
- [24] Furnham, N., G.L. Holliday, T.A. de Beer, J.O. Jacobsen, W.R. Pearson, and J.M. Thornton, The Catalytic Site Atlas 2.0: cataloging catalytic sites and residues identified in enzymes, *Nucleic Acids Res*. 42 (2014) D485-9.
- [25] Bhaduri, A., G. Pugalenth, and R. Sowdhamini, PASS2: an automated database of protein alignments organised as structural superfamilies, *BMC Bioinformatics*. 5 (2004) 35.
- [26] TheUniProtConsortium, Activities at the Universal Protein Resource (UniProt), *Nucleic Acids Res*. 42 (2014) D191-8.
- [27] Krissinel, E. and K. Henrick, Multiple Alignment of Protein Structures in Three Dimensions, *Computational Life Sciences*. 3695 (2005) 67-78.
- [28] Krieger, E., G. Koraimann, and G. Vriend, Increasing the precision of comparative models with YASARA NOVA--a self-parameterizing force field, *Proteins*. 47 (2002) 393-402.
- [29] Lombard, V., H. Golaconda Ramulu, E. Drula, P.M. Coutinho, and B. Henrissat, The carbohydrate-active enzymes database (CAZy) in 2013, *Nucleic Acids Res*. 42 (2014) D490-5.
- [30] Finn, R.D., A. Bateman, J. Clements, P. Coghill, R.Y. Eberhardt, S.R. Eddy, A. Heger, K. Hetherington, L. Holm, J. Mistry, E.L. Sonnhammer, J. Tate, and M. Punta, Pfam: the protein families database, *Nucleic Acids Res*. 42 (2014) D222-30.
- [31] Ondrechen, M.J., J.G. Clifton, and D. Ringe, THEMATICS: a simple computational predictor of enzyme function from structure, *Proc Natl Acad Sci U S A*. 98 (2001) 12473-8.
- [32] Wei, Y., J. Ko, L.F. Murga, and M.J. Ondrechen, Selective prediction of interaction sites in protein structures with THEMATICS, *BMC Bioinformatics*. 8 (2007) 119.
- [33] Sankararaman, S. and K. Sjolander, INTREPID--INformation-theoretic TREe traversal for Protein functional site IDentification, *Bioinformatics*. 24 (2008) 2445-52.
- [34] Sankararaman, S., B. Kolaczowski, and K. Sjolander, INTREPID: a web server for prediction of functionally important residues by evolutionary analysis, *Nucleic Acids Res*. 37 (2009) W390-5.
- [35] Capra, J.A., R.A. Laskowski, J.M. Thornton, M. Singh, and T.A. Funkhouser, Predicting protein ligand binding sites by combining evolutionary sequence conservation and 3D structure, *PLoS Comput Biol*. 5 (2009) e1000585.

- [36] Krissinel, E. and K. Henrick, Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions, *Acta Crystallogr D Biol Crystallogr.* 60 (2004) 2256-68.
- [37] Henikoff, S. and J.G. Henikoff, Amino acid substitution matrices from protein blocks, *Proc Natl Acad Sci U S A.* 89 (1992) 10915-9.
- [38] Aleshin, A.E., P.H. Feng, R.B. Honzatko, and P.J. Reilly, Crystal structure and evolution of a prokaryotic glucoamylase, *J Mol Biol.* 327 (2003) 61-73.
- [39] Sauer, J., B.W. Sigurskjold, U. Christensen, T.P. Frandsen, E. Mirgorodskaya, M. Harrison, P. Roepstorff, and B. Svensson, Glucoamylase: structure/function relationships, and protein engineering, *Biochim Biophys Acta.* 1543 (2000) 275-293.
- [40] Gregg, K.J., W.F. Zandberg, J.H. Hehemann, G.E. Whitworth, L. Deng, D.J. Vocadlo, and A.B. Boraston, Analysis of a new family of widely distributed metal-independent alpha-mannosidases provides unique insight into the processing of N-linked glycans, *J Biol Chem.* 286 (2011) 15586-96.
- [41] Wass, M.N., G. Barton, and M.J. Sternberg, CombFunc: predicting protein function using heterogeneous data sources, *Nucleic Acids Res.* 40 (2012) W466-70.
- [42] Aleshin, A.E., L.M. Firsov, and R.B. Honzatko, Refined Structure for the Complex of Acarbose with Glucoamylase from *Aspergillus awamori* Var. X100 to 2.4-Angstrom Resolution, *Journal of Biological Chemistry.* 269 (1994) 15631-15639.
- [43] Nage, M., A. Tsuchiya, T. Katayama, K. Yamamoto, S. Wakatsuki, and R. Kato, Structural basis of the catalytic reaction mechanism of novel 1,2-alpha-L-fucosidase from *Bifidobacterium bifidum*, *J Biol Chem.* 282 (2007) 18497-509.
- [44] Leonard, R., M. Pabst, J.S. Bondili, G. Chambat, C. Veit, R. Strasser, and F. Altmann, Identification of an Arabidopsis gene encoding a GH95 alpha1,2-fucosidase active on xyloglucan oligo- and polysaccharides, *Phytochemistry.* 69 (2008) 1983-8.
- [45] Ishimizu, T., C. Hashimoto, R. Takeda, K. Fujii, and S. Hase, A Novel alpha1,2-L-fucosidase acting on xyloglucan oligosaccharides is associated with endo-beta-mannosidase, *J Biochem.* 142 (2007) 721-9.
- [46] Rogowski, A., J.A. Briggs, J.C. Mortimer, T. Tryfona, N. Terrapon, E.C. Lowe, A. Basle, C. Morland, A.M. Day, H. Zheng, T.E. Rogers, P. Thompson, A.R. Hawkins, M.P. Yadav, B. Henrissat, E.C. Martens, P. Dupree, H.J. Gilbert, and D.N. Bolam, Glycan complexity dictates microbial resource allocation in the large intestine, *Nat Commun.* 6 (2015) 7481.
- [47] Cantarel, B.L., P.M. Coutinho, C. Rancurel, T. Bernard, V. Lombard, and B. Henrissat, The Carbohydrate-Active EnZymes database (CAZy): an expert resource for Glycogenomics, *Nucleic Acids Research.* 37 (2009) D233-D238.
- [48] Daiyasu, H., H. Saino, H. Tomoto, M. Mizutani, K. Sakata, and H. Toh, Computational and experimental analyses of furcain hydrolase for substrate specificity studies of disaccharide-specific glycosidases, *J Biochem.* 144 (2008) 467-75.
- [49] Hakulinen, N., O. Turunen, J. Janis, M. Leisola, and J. Rouvinen, Three-dimensional structures of thermophilic beta-1,4-xylanases from *Chaetomium thermophilum* and *Nonomuraea flexuosa*. Comparison of twelve xylanases in relation to their thermal stability, *Eur J Biochem.* 270 (2003) 1399-412.
- [50] Nagano, N., C.A. Orengo, and J.M. Thornton, One Fold with Many Functions: The Evolutionary Relationships between TIM Barrel Families Based on their Sequences, Structures and Functions, *J Mol Biol.* 321 (2002) 741-65.

- [51] Anantharaman, V., L. Aravind, and E.V. Koonin, Emergence of diverse biochemical activities in evolutionarily conserved structural scaffolds of proteins, *Curr Opin Chem Biol.* 7 (2003) 12-20.
- [52] Ishida, T., S. Fushinobu, R. Kawai, M. Kitaoka, K. Igarashi, and M. Samejima, Crystal structure of glycoside hydrolase family 55 {beta}-1,3-glucanase from the basidiomycete *Phanerochaete chrysosporium*, *J Biol Chem.* 284 (2009) 10100-9.
- [53] Cutfield, S.M., G.J. Davies, G. Murshudov, B.F. Anderson, P.C.E. Moody, P.A. Sullivan, and J.F. Cutfield, The structure of the exo-beta-(1,3)-glucanase from *Candida albicans* in native and bound forms: Relationship between a pocket and groove in family 5 glycosyl hydrolases, *Journal of Molecular Biology.* 294 (1999) 771-783.
- [54] Brodtkin, H.R., W.R.P. Novak, A.C. Milne, J.A. D'Aquino, N.M. Karabacak, I.G. Goldberg, J.N. Agar, M.S. Payne, G.A. Petsko, M.J. Ondrechen, and D. Ringe, Evidence of the Participation of Remote Residues in the Catalytic Activity of Co-Type Nitrile Hydratase from *Pseudomonas putida*, *Biochemistry.* 50 (2011) 4923-4935.
- [55] Somarowthu, S., H.R. Brodtkin, J.A. D'Aquino, D. Ringe, M.J. Ondrechen, and P.J. Beuning, A tale of two isomerases: compact versus extended active sites in ketosteroid isomerase and phosphoglucose isomerase, *Biochemistry.* 50 (2011) 9283-95.
- [56] Walsh, J.M. and P.J. Beuning, Synthetic nucleotides as probes of DNA polymerase specificity, *J Nucleic Acids.* 2012 (2012) 530963.
- [57] Shehadi, I.A., A. Abyzov, A. Uzun, Y. Wei, L.F. Murga, V. Ilyin, and M.J. Ondrechen, Active site prediction for comparative model structures with thematics, *J Bioinform Comput Biol.* 3 (2005) 127-43.
- [58] Kuhad, R.C., R. Gupta, and A. Singh, Microbial cellulases and their industrial applications, *Enzyme Res.* 2011 (2011) 280696.