J Struct Funct Genomics (2009) 10:107–125
DOI 10.1007/s10969-008-9056-5

Target selection and annotation for the structural genomics of the amidohydrolase and enolase superfamilies

Ursula Pieper · Ranyee Chiang · Jennifer J. Seffernick · Shoshana D. Brown · Margaret E. Glasner · Libusha Kelly · Narayanan Eswar · J. Michael Sauder · Jeffrey B. Bonanno · Subramanyam Swaminathan · Stephen K. Burley · Xiaojing Zheng · Mark R. Chance · Steven C. Almo · John A. Gerlt · Frank M. Raushel · Matthew P. Jacobson · Patricia C. Babbitt · Andrej Sali

Received: 8 August 2008 / Accepted: 12 December 2008 / Published online: 14 February 2009

Abstract To study the substrate specificity of enzymes, we use the amidohydrolase and enolase superfamilies as model systems; members of these superfamilies share a common TIM barrel fold and catalyze a wide range of chemical reactions. Here, we describe a collaboration between the Enzyme Specificity Consortium (ENSPEC) and the New York SGX Research Center for Structural Genomics (NYSGXRC) that aims to maximize the structural coverage of the amidohydrolase and enolase superfamilies. Using sequence- and structure-based protein comparisons, we first selected 535 target proteins from a variety of genomes for high-throughput structure determination by X-ray crystallography; 63 of these targets were not previously annotated as superfamily members. To date, 20 unique amidohydrolase and 41 unique enolase structures have been determined, increasing the fraction of sequences in the two superfamilies that can be modeled based on at least 30% sequence identity from 45% to 73%. We present case studies of proteins related to uronate isomerase (an amidohydrolase superfamily member) and mandelate racemase (an enolase superfamily member), to illustrate how this structure-focused approach can be used to generate hypotheses about sequence–structure–function relationships.

Keywords Amidohydrolase and enolase superfamilies · Structural genomics · Structure annotation · Target selection

Current affiliation of SGX Pharmaceuticals is Eli Lilly and Company, 10505 Roselle Street, San Diego, CA 92121, USA.

Electronic supplementary material The online version of this article (doi:10.1007/s10969-008-9056-5) contains supplementary material, which is available to authorized users.

U. Pieper (✉)
Departments of Biopharmaceutical Sciences and Pharmaceutical Chemistry, California Institute for Quantitative Biosciences, University of California at San Francisco, Byers Hall at Mission Bay, Office 501-32, 1700 4th Street, San Francisco, CA 94158, USA
e-mail: ursula@salilab.org
URL: http://salilab.org

R. Chiang · J. J. Seffernick · M. E. Glasner · L. Kelly · N. Eswar
Departments of Biopharmaceutical Sciences and Pharmaceutical Chemistry, California Institute for Quantitative Biosciences, University of California at San Francisco, Byers Hall at Mission Bay, Office 501, 1700 4th Street, San Francisco, CA 94158, USA

S. D. Brown
Department of Biopharmaceutical Sciences, California Institute for Quantitative Biosciences, University of California at San Francisco, Byers Hall at Mission Bay, Office 501C, 1700 4th Street, San Francisco, CA 94158, USA

J. M. Sauder · S. K. Burley
SGX Pharmaceuticals Inc, 10505 Roselle Street, San Diego, CA 92121, USA

J. B. Bonanno · S. C. Almo
Departments of Biochemistry and Physiology and Biophysics, Albert Einstein College of Medicine, Bronx, NY 10461, USA

S. Swaminathan
Biology Department, Brookhaven National Laboratory, Upton, NY 11973, USA

X. Zheng · M. R. Chance
Case Center for Proteomics & Bioinformatics, Case Western Reserve University, Cleveland, OH 44106, USA

J. A. Gerlt
Departments of Biochemistry and Chemistry, University of Illinois, Urbana, IL 61801, USA

Abbreviations

PDB	Protein Data Bank
NYSGXRC	New York SGX Research Center for Structural Genomics
ENSPEC	Enzyme specificity consortium
SFLD	Structure function linkage database
PSI	Protein structure initiative
NR	Non-redundant database of protein sequences
ESI	Electrospray ionization
HMM	Hidden Markov Model

Introduction

A long-standing challenge in biology is to predict the molecular function of proteins from their sequences and/or structures. This task is facilitated by a limited number of domain folds [1], restricting the set of structural types that must be studied in deducing a much larger set of functions. Special challenges, however, exist for functional prediction in different classes of proteins. For example, the function of an enzyme often cannot be correctly predicted because there are no clear links from the domain fold to the catalytic function and substrate specificity. Off-setting these problems, studies of genomes and sets of homologous proteins demonstrate that some aspects of catalysis are often conserved between evolutionarily-related proteins, even when these proteins catalyze different overall reactions [2–4]. This empirical observation restricts the functional space that must be considered, further facilitating prediction and leading to definitions of homologous sets of enzymes in terms of protein superfamilies and families based not only on structural conservation, but also on functional conservation [5]: Superfamily members share a common ancestor and potentially some aspects of function, while members of the same family are isofunctional, catalyzing the same overall reaction(s).

The large and diverse amidohydrolase and enolase superfamilies provide a particularly attractive opportunity to study the problem of predicting substrate specificity and enzymatic mechanisms from evolutionary and physical perspectives. These superfamilies are attractive targets because significant knowledge about the specificity of many of their members already exists, while there are still large areas of their sequence space where we do not have any structural or functional information.

Members of the amidohydrolase superfamily catalyze the hydrolysis of a wide range of substrates bearing amide or ester functional groups at carbon and phosphorus centers [6, 7]. A common feature for this superfamily is a mononuclear or binuclear metal center coordinated in a $(\beta/\alpha)_8$-barrel (TIM barrel) polypeptide chain fold. The active site is formed by loops at the C-terminal ends of the β-strands. Currently, 36 named families have been identified based on the experimentally verified catalytic reactions. The set of superfamily sequences has been clustered into 90 subgroups based on sequence and in some cases active site similarities (the Structure-Function Linkage Database [8]: http://sfld.rbvi.ucsf.edu). In some subgroups, additional information about chemical reactions catalyzed by subgroup members is available; for many of the subgroups, however, no information about functional specificity is available.

Enolase superfamily members catalyze the abstraction of a proton α to a carboxylic acid to form an enolate anion intermediate [9, 10]. Members of this superfamily share an N-terminal $\alpha+\beta$ capping domain, as well as a C-terminal $(\beta/\alpha)_7\beta$-barrel domain (modified TIM barrel). The active site is formed by loops at the C-terminal ends of the TIM barrel β-strands and two flexible loops from the capping domain; the active site also includes a Mg^{2+} ion [11]. Reactions catalyzed by enolases are less diverse than those of the amidohydrolases. The enolases are currently organized into 16 named families and 6 subgroups [8]. Approximately 50% of the sequences in the superfamily are of unknown function.

The amidohydrolase and enolase superfamilies are the focus of our Enzyme Specificity Consortium (ENSPEC), members of which include protein crystallographers, enzymologists, and computational biologists. We aim to predict the substrate specificity of an enzyme based on its experimentally determined and/or modeled structure

F. M. Raushel
Department of Chemistry, Texas A&M University,
P.O. Box 30012, College Station, TX 77842-3012, USA

M. P. Jacobson
Departments of Biopharmaceutical Sciences and Pharmaceutical Chemistry, California Institute for Quantitative Biosciences, University of California at San Francisco, 600 16th St., Box 2240, Genentech Hall at Mission Bay, Room N472C, San Francisco, CA 94158-2517, USA

P. C. Babbitt
Departments of Biopharmaceutical Sciences and Pharmaceutical Chemistry, California Institute for Quantitative Biosciences, University of California at San Francisco, Byers Hall at Mission Bay, Office 508E, 1700 4th Street, San Francisco, CA 94158, USA

A. Sali (✉)
Departments of Biopharmaceutical Sciences and Pharmaceutical Chemistry, California Institute for Quantitative Biosciences, University of California at San Francisco, Byers Hall at Mission Bay, Office 503B, 1700 4th Street, San Francisco, CA 94158, USA
e-mail: sali@salilab.org
URL: http://salilab.org

[2–4, 7, 10–42]. This goal has been enabled by determination of crystallographic structures representing many of the amidohydrolase and enolase families.

To maximize the number of experimentally determined structures, ENSPEC has collaborated with the New York SGX Research Center for Structural Genomics (NYSGXRC), which is one of the four large-scale production centers of the Protein Structure Initiative (PSI) (http://www.nigms.nih.gov/Initiatives/PSI; [43]). NIH guidelines mandate that 70% of the PSI targets come from diverse protein families selected by and shared among the four production centers [43]. About 15% of the targets are reserved for proteins of biomedical relevance defined by each center, and the remaining 15% are "community-nominated" targets. Several hundred of the NYSGXRC community targets are amidohydrolases and enolases nominated by ENSPEC. To date, the collaboration has determined 25 amidohydrolase and 50 enolase structures, contributing substantially to the total of 154 amidohydrolase and 89 enolase structures in the Protein Data Bank (PDB; 6/16/08) [44].

We begin by outlining the data sources and methods used for target selection and structure-based functional annotation (Materials and Methods). Second, we present the results of the target selection process, the status of the selected targets in the structural genomics pipeline, and the improvement in the modeling of the amidohydrolase and enolase superfamilies made possible by the new crystallographic structures (Results and Discussion). We conclude by discussing the biological impact of two sample target structures.

Materials and methods

Target selection

Target selection begins by identifying sequences of known members of the superfamilies (seed sequences), followed by filtering to obtain an initial target list. To identify additional members, we applied sequence- and structure-based expansion methods, followed by filtering for source organisms preferred by NYSGXRC. Superfamily membership for the additional targets was verified by expert curators by inspecting their sequences for probable catalytic residues. A web-based target selection tool was also constructed for further manual filtering to obtain the final target list.

Seed sequence sources

Verified amidohydrolase and enolase superfamily sequences (i.e., seed sequences) were obtained from the Structure Function Linkage Database (SFLD; http://sfld.rbvi.ucsf.edu/) [8]. The SFLD database is a manually constructed database that classifies enzymes hierarchically, based on specific sequence, structure, and functional criteria. The database is updated by a semi-automated method that detects new superfamily members by matching their sequences to Hidden Markov Models trained using the sequences of verified superfamily members, with subsequent manual inspection to verify the presence of catalytic residues. In June 2005, when our target list was constructed, the SFLD contained 3,701 amidohydrolases and 1,795 enolases.[1]

Filtering of seed sequences

PSI guidelines require that structural genomics targets share ~30% or less amino acid sequence identity to a known thee-dimensional structure. To satisfy this condition, the seed amidohydrolase and enolase sequences were processed using the automated comparative modeling server MODWEB (http://salilab.org/modweb) [45]. Sequences with more than 30% sequence identity to any structure in the PDB over at least 70% of their length were identified and excluded from further consideration.

Sequence-based expansion of amidohydrolase and enolase superfamily members

For each seed amidohydrolase and enolase, homologous sequences in the UNIPROT database [46] were identified by the BUILD_PROFILE routine of MODELLER-9 [45]. BUILD_PROFILE is an iterative database-searching tool that relies on local dynamic programming to generate alignments and a robust estimate of their statistical significance. This method identified additional potential amidohydrolase and enolase sequences that were not present in the seed sequence pools.

Structure-based expansion of amidohydrolase superfamily members

In addition to the SFLD entries, we also used the known amidohydrolase superfamily structures to find additional potential amidohydrolase superfamily members (this expansion was not performed for the enolase superfamily). We began by splitting 100 PDB files containing known amidohydrolase structures (June 2005) into separate monomeric structures and clustering them at 80% sequence identity. The resulting 45 non-redundant structures were

[1] The numbers of sequences in the publicly accessible version of the SFLD differ from those cited here because large numbers of sequences are undergoing curation at any given time and are therefore not yet listed on the public site.

used for comparative modeling using the automated modeling server MODWEB [45].

First, each structure sequence was used as a query to find its homologs in UNIPROT using PSIBLAST [47]. Second, these homologs were modeled using the corresponding structure as a template. All models were deposited in our comprehensive MODBASE database of comparative protein structure models (http://salilab.org/modbase/; direct links to the datasets can be found in the supplemental materials) [48]. In addition, the amidohydrolase homologs found in UNIPROT were filtered by removing known amidohydrolase superfamily members, and then subjected to standard comparative modeling with MODWEB using all non-redundant chains in the PDB as potential templates. This step allowed us to eliminate sequences that are likely members of other superfamilies, judged by sequence identity and coverage.

Filtering by organism

While seed sequences could come from any genome, the additional amidohydrolase sequences identified by sequence- and structure-based expansions were filtered for ease of cloning to include only 79 organisms with genomic DNA available to NYSGXRC in 2005 and the marine metagenome from the Sargasso Sea sequencing project (formerly called environmental sequences) [49]. For simplicity, we call the 79 genomes plus the marine metagenome the "NYSGXRC genomes" (Table 1). The NYSGXRC reagent genomes have since been expanded to include over 115 organisms.

Verification of catalytic residues

The putative amidohydrolase sequences resulting from the sequence- and structure-based expansions were aligned to existing amidohydrolase Hidden Markov Models (HMMs) in the SFLD and manually inspected for probable catalytic residues. The final target list only includes sequences with at least 70% of the catalytic residues present.

Target selection tool

For final manual filtering of the target list, we constructed a web-based target selection tool. The tool comprises a combination of MySQL database tables with an interactive web-interface using LAMP [50]. It contains information about the sequences, including UNIPROT annotation, organism, sequence length, closest known structure, sequence identity to other cluster members, and domain boundaries for the TIM barrel domain obtained from SFLD. The interface allows searching for project datasets, organism groups, homologs based on sequence identity, and clusters of related sequences; the resulting sequences can be flagged for rejection or inclusion into the final target list.

Analysis of the target structures

The amidohydrolase and enolase superfamilies were annotated using computational tools. Cytoscape clustering gives an overview of how the targets are distributed across the superfamily [51]. Also, template-based modeling determines how many new sequences can be modeled with the new structural information [45].

Sequence clustering of amidohydrolase superfamily by cytoscape

The time required to perfom BLAST searches against the NCBI non-redundant database (NR) of protein sequences [52] was prohibitive due to the size and complexity of the superfamily. Thus, a custom database was created containing only the amidohydrolase sequences in the SFLD. To generate the all-by-all connections for cytoscape clustering, BLAST searches were then performed against this database at an E-value cutoff of 10^{-10}, using each sequence in the set as a query. Because this custom database contained only sequences known to be members of the amidohydrolase superfamily, the generation of E-values is biased. Consequently, the E-values from this analysis cannot be directly compared to those calculated by BLAST against the NCBI NR database. A cytoscape [51] network was created from these BLAST results. In the absence of established statistical techniques for selecting the E-value cutoff, we examined the superfamily networks at a number of different E-value cutoffs, and present here only one of the corresponding networks, at an E-value cutoff of 10^{-10}. Further discussion regarding choosing and interpreting E-value cutoffs for sequence similarity networks may be found in [53]. Each node in the network represents a single sequence and each edge represents the pairwise connection between two sequences with the most significant BLAST E-value (better than the cut-off) connecting the two sequences. Lengths of edges are not meaningful, except that sequences in tightly clustered groups are more similar to each other than sequences with few connections. The nodes were arranged using the yFiles organic layout provided in Cytoscape version 2.4. Tools for visualization of protein networks were created by the UCSF Resource for Biocomputing, Visualization, and Informatics (http://www.rbvi.ucsf.edu).

Sequence clustering of enolase superfamily by cytoscape

To generate the all-by-all connections for cytoscape clustering, BLAST analysis was performed against the NR

database, using the sequences in the mandelate racemase-like, glucarate dehydratase-like, mannonate dehydratase-like, and muconate cycloisomerase-like subgroups of the SFLD enolase superfamily. The enolase subgroup was not included in this analysis. Almost all of the enolase subgroup members are in the enolase family, the sequences of which are all isofunctional, i.e. they all perform the well-characterized enolase reaction, important in glycolysis. Only hits in the aforementioned subgroups were used for further analysis. The cytoscape network was created as described above, but using an E-value cutoff for this superfamily of 10^{-40}.

Table 1 List of 80 NYSGXRC genomes (as of June 2005)

Organism	Taxonomy ID	Organism	Taxonomy ID
Aeropyrum pernix	56636	*Listeria monocytogenes*	1639
Aquifex aeolicus	63363	Metagenome sequences (Gene synthesis)	256318
Arabidopsis thaliana	3702	*Methanococcus jannaschi*	2190
Archaeoglobus fulgidus	2234	*Mus musculus*	10090
Bacillus cereus	1396	*Mycobacterium turberculosis* H37Rv	83332
Bacillus halodurnas	86665	*Mycoplasma pneumonia*	2104
Bacillus subtilis	1423	*Neisseria gonorrhoeae*	485
Bacillus thuringiensis	1428	*Neisseria meningitidis*	487
Bartonella henselae	38323	*Nostoc*	1180
Bordetella pertussis	520	*Oryctolagus cuniculus*	9986
Borrelia burgdorferi	139	*Oryza sativa*	4530
Bos taurus	9913	*Ovis aries*	9940
Caenorhabditis elegans	6239	*Porphyromonas gingivalis*	837
Campylobacter jejuni	197	*Pseudomonas aeruginosa*	287
Candida albicans	5476	*Pseudomonas putida*	303
Canis familiaris	9615	*Pyrococcus furiosus*	2261
Capra hircus	9925	*Pyrococcus horikoshii*	53953
Caulobacter vibrioides	155892	*Rattus norvegicus*	10116
Clostridium acetobutylicum	1488	*Rhodobacter sphaeroides*	1063
Corynebacterium diphtheriae	1717	*Saccharomyces cerevisiae*	4932
Cryptococcus neoformans	5207	*Salmonella typhimurium*	602
Cryptosporidium parvum	5807	*Schizosaccharomyces pombe*	4896
Deinococcus radiodurans	1299	*Shigella Flexneri* type 2a	42897
Desulfovibrio vulgaris	881	Simian immunodeficiency virus	11723
Dictyostelium discoideum	44689	*Staphylococcus aureus*	1280
Drosophila melanogaster	7227	*Staphylococcus epidermidis*	1282
Enterobacter cloacae	550	*Streptococcus mutans*	1309
Enterococcus faecalis	1351	*Streptococcus pneumoniae*	1313
Equus caballus	9796	*Streptococcus pyogenes*	1314
Escherichia coli	562	*Sulfolobus solfataricus*	2287
Escherichia coli 0157:H7	83334	*Sus scrofa*	9823
Felis catus	9685	*Takifugu rubripes*	31033
Gallus gallus	9031	*Thermoplasma acidophilum*	2303
Haemophilus influenzae	727	*Thermoplasma volcanium*	50339
Halobacterium sp. NRC-1	64091	*Thermotoga maritima*	2336
Helicobacter pylori	210	*Ureaplasma urealyticum*	2130
Homo sapiens	9606	*Vibrio cholerae*	666
Human immunodeficiency virus type 1	11676	*Xenopus laevis*	8355
Klebsiella pneumoniae	573	*Xylella fastidiosa*	2371
Legionella pneumophila	446	*Zea mays*	4577

Table 2 Summary of new enolase and amidohydrolase X-ray crystal structures and automated template-based modeling results, including subgroup and family assignments

PDB code	Database accession number (Genpept GI IDs)	No of sequences in Psi-blast alignment	No of sequences with acceptable models and/or fold assignments	No of models >50% seq. ID (min 50% template coverage)	No of models 30–50% seq. ID (min 50% template coverage)	No of models <30% seq. ID (min 50% template coverage)	Subgroup assignment	Family assignment
Enolase super family								
2GL5	16420812	2,863	2,777	0	98	2,462	Mandelate racemase-like	Galactonate dehydratase
2GDQ	2633433	2,234	2,129	1	0	2,036	Mandelate racemase-like	None
2GSH	16420830	2,588	2,286	16	9	2,110	Mandelate racemase-like	L-Fuconate dehydratase
2HNE	21115341	2,746	2,712	83	20	2,527	Mandelate racemase-like	None
2HZG	77386310	2,667	2,341	1	1	2,248	Mandelate racemase-like	None
2I5Q	15832389	2,566	2,340	21	13	2,206	Mandelate racemase-like	None
2NQL	17743914	2,849	2,470	2	1	2,356	Mandelate racemase-like	None
2O56	16767118	3,016	2,968	15	127	2,735	Mandelate racemase-like	None
2OQH	21225834	2,690	2,668	2	32	2,630	Glucarate dehydratease-like	None
2OQY	23100298	2,700	2,631	1	0	3,004	Muconate cycloisomerase-like	None
2OVL	21221904	2,670	2,656	1	97	2,534	Mandelate racemase-like	None
2OG9	91786345	2,669	2,664	10	75	2,553	Mandelate racemase-like	L-Talarate/galactarate dehydratase
2OLA	88195610	2,719	2,697	5	3	2,652	Muconate cycloisomerase-like	*o*-Succinylbenzoate synthase
2OO6	91778214	3,271	3,221	3	2	3,111	Mandelate racemase-like	None
2OKT	57650581	2,723	2,705	5	3	2,664	Muconate cycloisomerase-like	*o*-Succinylbenzoate synthase
2OPJ	72161814	2,562	1,855	19	31	1,712	Mandelate racemase-like	*o*-Succinylbenzoate synthase
2OX4	56552160	2,733	2,639	11	136	2,449	Mandelate racemase-like	None
2OZ3	67154209	2,743	2,656	38	25	2,567	Mandelate racemase-like	None
2OZ8	13475907	2,821	2,674	0	0	2,641	Mandelate racemase-like	None
2POI	46136735	2,747	2,661	13	52	2561	Mandelate racemase-like	None
2OZT	22294898	2,816	2,726	0	16	2,722	Muconate cycloisomerase-like	*o*-Succinylbenzoate synthase
2PCE	83951697	2,693	2,683	1	16	2,635	Muconate cycloisomerase-like	None
2PGE	51244103	2,779	2,767	1	19	2,768	Muconate cycloisomerase-like	*o*-Succinylbenzoate synthase
2PGW	16263250	2,781	2,743	1	3	2,694	Mandelate racemase-like	None
2PMQ	114764387	2,881	2,760	3	14	2,723	Muconate cycloisomerase-like	None
2POD	53723090	2,745	2,732	12	97	2,585	Mandelate racemase-like	Galactonate dehydratase
2POZ	13488170	2,861	2,836	1	162	2,687	Mandelate racemase-like	None
2PPG	16262827	2,947	2,755	2	66	2,707	Mandelate racemase-like	None
2PS2	83774494	2,777	2,753	3	16	2,712	Muconate cycloisomerase-like	None
2QDE	56478643	2,930	2,670	1	62	2,595	Muconate cycloisomerase-like	None
2QGY	110347373	2,988	2,899	0	1	2,912	Mandelate racemase-like	None
2QQ6	108803396	3,238	3,216	0	201	3,081	Mandelate racemase-like	Galactonate dehydratase
2QYE	83951695	3,128	3,121	0	21	3,110	Muconate cycloisomerase-like	None

Table 2 continued

PDB code	Database accession number (Genpept GI IDs)	No of sequences in Psi-blast alignment	No of sequences with acceptable models and/or fold assignments	No of models >50% seq. ID (min 50% template coverage)	No of models 30–50% seq. ID (min 50% template coverage)	No of models <30% seq. ID (min 50% template coverage)	Subgroup assignment	Family assignment
3BJS	6791043	3,261	2,897	4	82	2,810	Mandelate racemase-like	None
2QDD	83951694	2,868	2,852	0	20	2,849	Muconate cycloisomerase-like	*o*-Succinylbenzoate synthase
3CAW	42522147	2,220	2,139	0	0	2,137	Muconate cycloisomerase-like	*o*-Succinylbenzoate synthase
3CT2	70731221	3,483	2,771	84	77	2,667	Muconate cycloisomerase-like	Muconate cycloisomerase
3CYJ	108805509	3,551	2,879	8	35	2,838	Mandelate racemase-like	None
3DDM	33575875	3,603	3,576	6	27	3,591	Mandelate racemase-like	None
3BSM	92115090	3,372	3,359	86	165	3,097	Mannonate dehydratase-like	Mannonate dehydratase
Total (unique sequences)		7,013	5,804	398	766	5,190		
Amidohydrolase superfamily								
2GOK	17742376	3,001	2,943	96	103	2,678	Imidazolonepropionase-like	Imidazolonepropionase
2OOD	27378991	3,609	3,572	0	160	3,440	Guanine deaminase-like	None
2OOF	83646866	3,588	3,578	142	154	3,270	Imidazolonepropionase-like	Imidazolonepropionase
2I5G	9951721	569	448	28	50	340	None	None
2I9U	15023121	3,386	3,363	5	96	3,198	Newfam59	None
2ICS	29342885	3,433	3,334	3	28	3,209	Unknown18	None
2IMR	9911007	3,790	3,502	1	3	3,498	None	None
2OGJ	17741648	3,527	3,510	5	18	3,395	Newfam71	None
2P9B	23466009	3,319	3,302	2	37	3,230	Unknown41	None
2PAJ	91783796	3,264	3,252	4	116	3,128	Unknown55	None
2QO1	13422863	460	263	35	14	172	Uronate isomerase-like	Uronate isomerase
2Q6B	15615056	306	189	3	0	167	Uronate isomerase-like	Uronate isomerase
2QS8	114773165	3,508	3,497	15	144	3,280	Unknown43	None
2QT3	32455889	3,723	3,693	1	49	3,606	Unknown95	None
2RAG	16126978	911	602	7	53	502	Newfam32	None
2R8C	4447959	3,649	3,632	19	195	3,359	Unknown47	None
2I9U	150231121	3,386	3,363	5	96	3,198	Newfam59	None
2OOF	83646866	3,588	3,578	142	154	3,270	Imidazolonepropinase-like	Imidazolonepropionase
3B40	9948434	1,149	656	16	34	504	Newfam190	None
3CJP	15896580	3,289	3,286	1	1,467	1,851	Newfam63	None
3BE7	4436882	3,697	3,198	4	112	3,042	Unknown42	None
Total (unique sequences)		12,101	11,628	302	2,429	8,912		

Only one entry is shown for structures determined in different crystal forms or ligand binding states. An acceptable model is defined to be based on a significant PSI-BLAST E-value (0.0001) or a favorable GA341 model score (>0.7) [60]

Template-based modeling by MODWEB

Automated comparative modeling of all known protein sequences using the new NYSGXRC crystallographic structures as templates was performed with MODWEB [45]. We relied on the MODWEB option that allows using a protein structure as input and results in models for all of the identifiable sequence homologs of the input structure from the NCBI NR database; these homologs were identified during ten PSI-BLAST iterations of the template sequence against NR (E-value cutoff is 0.0001). The results are available at http://salilab.org/modbase/models_nysgxrc_latest.html (Table 2).

Results and discussion

We first present the results of the target selection procedure. We also describe the current snapshot of the progress of the targets through our structural genomics pipeline (June 2008). We then indicate how the resulting crystallographic structures are distributed across the two superfamilies. Next, we determine the number of protein sequences in the comprehensive sequence databases that are detectably related to these protein structures (i.e., the modeling leverage). Finally, for each of the two superfamilies, we describe an example target with interesting biological features.

Target selection

Given the capacities of ENSPEC and NYSGXRC, the goal was to identify approximately 500 target sequences, approximately evenly distributed between the two superfamilies. These targets were obtained by selecting representatives from previously identified superfamily members as well as by identifying new superfamily members in a select set of genomes (Materials and Methods).

Targets for the amidohydrolase superfamily

From the SFLD, we obtained a list of 3,701 amidohydrolase superfamily members. The first filtering step resulted in 1,918 sequences with less than 30% sequence identity to a known structure and at least 250 amino acid residues in length, originating from 424 organisms. We chose the 30% sequence identity limit, in congruence with NIH PSI guidelines, to concentrate our efforts on protein sequences with limited structural knowledge; sequences related at less than 30% sequence identity to the closest known structure are frequently modeled inaccurately due to errors in the corresponding target-template alignments [54–56].

These 1,918 sequences were further filtered manually using the target selection tool to obtain the reduced set of 224 target sequences. The selected amidohydrolase superfamily members are evenly distributed among the various clades of the superfamily, thus representing the diversity within the superfamily. Preference was given to the NYSGXRC genomes, but other organisms were also considered.

The 224 targets can be divided into 76 clusters with less than 30% sequence identity between any pair of sequences from two different clusters, 126 clusters at 50% sequence identity, and 177 clusters at 80% sequence identity. The amidohydrolase superfamily members all contain the defining conserved TIM barrel domain with some variation in their lengths; all targets are between 224 and 628 amino acid residues long, with 90% of them shorter than 500 residues. The length variation stems mostly from loops that connect the main secondary structure elements of the TIM barrel fold and is consistent with the previously observed size range for TIM barrel domains (150 to 500 residues [57]).

In addition to the known superfamily members, the sequence- and structure-based expansions detected 63 putative amidohydrolase superfamily members that were not initially in the SFLD (Table 3). These new potential targets fall into two categories: (i) divergent sequences that were detected by the sequence-based approach (Fig. 1, blue box) and (ii) divergent sequences that were detected by the structure-based approach (Fig. 1, orange box). Of the 63 putative amidohydrolase superfamily sequences, 50 were subsequently verified using the SFLD update procedure. The presence of probable catalytic residues for the remaining 13 targets was verified manually. Nine of these 13 sequences were detected by both the sequence- and structure-based approaches, and four sequences were only detected by the structure-based approach. Thus, the sequence- and structure-based approaches yielded 13 additional targets that could not be identified as amidohydrolase superfamily members using previously available protocols (corresponding to 21% of the new putative members of the amidohydrolase superfamily).

In summary, the final amidohydrolase target list includes 224 previously identified amidohydrolase superfamily members, as well as the 63 newly identified sequences. The final list includes 287 sequences from 53 organisms that cover 22 (61%) of the named families in the superfamily (Fig. 2).

Targets for the enolase superfamily

We used a simpler selection scheme for the enolase superfamily members, because previous detailed studies have effectively found all of the superfamily members in publicly available sequence and structure databases (data not shown). Of the 1,795 sequences already established as enolase superfamily members, we selected as targets the

Table 3 Putative amidohydrolase superfamily members

Database ID (GenPept GI IDs)	Method	Organism	Length	Annotation available at target selection	Verification
7462218	Structure-based	*Thermotoga maritima*	434	Conserved hypothetical protein	HMM
7497374	Structure-based	*Caenorhabditis elegans*	818	Hypothetical protein C44B7.10	HMM
7500805	Structure-based	*Caenorhabditis elegans*	313	T21966 hypothetical protein F38E11.3—*Caenorhabditis elegans*	HMM
9948434	Structure-based	*Pseudomonas aeruginosa* PAO1	448	Probable dipeptidase precursor (*Pseudomonas aeruginosa*)	HMM
10173106	Structure-based	*Bacillus halodurans*	427	BH0493	HMM
10175729	Structure-based	*Bacillus halodurans*	571	DNA-dependent DNA polymerase beta chain	HMM
13700943	Structure-based	*Staphylococcus aureus* subsp. *aureus* N315	570	DNA-dependent DNA polymerase beta chain	HMM
14600641	Structure-based	*Aeropyrum pernix*	313	313aa long hypothetical microsomal dipeptidase	HMM
14601853	Template	*Aeropyrum pernix*	394	Hypothetical protein (*Aeropyrum pernix*)	HMM
14602106	Structure-based	*Aeropyrum pernix*	327	Hypothetical protein (*Aeropyrum pernix*)	HMM
15600589	Structure-based	*Pseudomonas aeruginosa* PAO1	325	D82971 hypothetical protein PA5396 (imported)—*Pseudomonas aeruginosa* (strain PAO1)	HMM
15612748	Structure-based	*Bacillus halodurans*	448	BH0185	HMM
15614834	Structure-based	*Bacillus halodurans*	310	Dipeptidase	HMM
15791917	Structure-based	*Campylobacter jejuni* subsp. *jejuni* NCTC	265	Hypothetical protein Cj0556	HMM
15805850	Structure-based	*Deinococcus radiodurans* R1	418	Hydrolase, putative	HMM
15896580	Structure-based	*Clostridium acetobutylicum*	262	Predicted amidohydrolase (dihydroorotase family)	HMM
15898656	Structure-based	*Sulfolobus solfataricus*	314	Microsomal dipeptidase	HMM
15925570	Structure-based	*Staphylococcus aureus* subsp. aureus N315	336	Conserved hypothetical protein	HMM
16125737	Structure-based	*Caulobacter vibrioides*	487	Uronate isomerase (EC 5.3.1.12) (Glucuronate isomerase) (UronicDE isomerase)	HMM
16126978	Structure-based	*Caulobacter vibrioides*	417	Dipeptidase	HMM
16127409	Structure-based	*Caulobacter vibrioides*	353	Hypothetical protein	HMM
16130781	Structure-based	*Escherichia coli* K12	464	Soluble protein involved in cell viability at the beginning of stationary phase; soluble protein involved in cell viability at the beginning of stationary phase, contains urease domain	HMM
16410647	Structure-based	*Listeria monocytogenes* EGD-e	570	lmo1231	HMM
17556402	Structure-based	*Caenorhabditis elegans*	352	Hypothetical protein Y71D11A.3a	HMM
19705473	Structure-based	*Rattus norvegicus*	336	2-amino-3-carboxymuconate-6-semialdehyde decarboxylase	HMM
19911227	Structure-based	*Homo sapiens*	336	2-amino-3-carboxylmuconate-6-semialdehyde decarboxylase	HMM
19911231	Structure-based	*Caenorhabditis elegans*	401	2-amino-3-carboxylmuconate-6-semialdehyde decarboxylase	HMM
24379660	Structure-based	*Streptococcus mutans* UA159	267	conserved hypothetical protein	HMM
33592291	Structure-based	*Bordetella pertussis* Tohama I	284	Putative 2-pyrone-4,6-dicarboxylic acid hydrolase	HMM
33593502	Structure-based	*Bordetella pertussis* Tohama I	341	Putative dipeptidase	HMM
39976001	Sequence- and structure-based	*Magnaporthe grisea* 70–15	417	Hypothetical protein	HMM

Table 3 continued

Database ID (GenPept GI IDs)	Method	Organism	Length	Annotation available at target selection	Verification
42527610	Structure-based	*Treponema denticola* ATCC 35405	371	Dihydroorotase, putative	HMM
42631159	Structure-based	*Haemophilus influenzae*	330	Hypothetical protein	HMM
51012913	Structure-based	*Saccharomyces cerevisiae*	313	YMR262W	HMM
51968376	Structure-based	*Arabidopsis thaliana*	346	Unnamed protein product	HMM
51968996	Structure-based	*Arabidopsis thaliana*	346	Unnamed protein product	HMM
55980841	Structure-based	*Thermus thermophilus* HB8	369	Amidohydrolase family protein	HMM
60279993	STRUCTURE-based	*Pseudomonas aeruginosa*	403	PvdM HMM	
66807941	Structure-based	*Dictyostelium discoideum*	359	Hypothetical protein	HMM
66808659	Structure-based	*Dictyostelium discoideum*	322	Hypothetical protein	HMM
1065989	Sequence-based	*Bacillus subtilis* subsp. *subtilis* str. 1	577	Adenine deaminase	HMM
15023784	Sequence-based	*Clostridium acetobutylicum*	570	Adenine deaminase	HMM
24636152	Structure-based	*Caenorhabditis elegans*	403	Hypothetical protein C44B7.12	HMM
29377069	Structure-based	*Enterococcus faecalis* V583	444	Chlorohydrolase family protein	HMM
40788915	Structure-based	*Homo sapiens*	777	Q93075_chr3:10265710-10295706_H233R_V272I_L374P PUTATIVE DEOXYRIBONUCLEASE KIAA0218 (EC 3.1.21.-)	HMM
45446932	Sequence- and structure-based	*Drosophila melanogaster*	774	CG32626-PA, isoform A	HMM
56203368	Sequence- and structure-based	*Homo sapiens*	776	Adenosine monophosphate deaminase 1 (isoform M	HMM
56203369	Sequence-based	*Homo sapiens*	780	OTTHUMP00000059283	HMM
57230710	Structure-based	*Filobasidiella neoformans*	469	Hydrolase, putative	HMM
63055053	Structure-based	Homo sapiens	761	TatD DNase domain containing 2	HMM
68250266	Structure-based	*Haemophilus influenzae*	251	Conserved putative deoxyribonuclease	HMM
429129	Sequence-based	*Saccharomyces cerevisiae*	797	YB9Z_YEAST HYPOTHETICAL 92.9 KD PROTEIN IN SSH1-APE3 INTERGENIC REGION	Manual
7293948	Sequence-based	*Drosophila melanogaster*	520	CG5998-PA	Manual
11463854	Sequence-based	*Drosophila melanogaster*	561	Male-specific IDGF	manual
14602062	Structure-based	*Aeropyrum pernix*	375	Hypothetical protein [*Aeropyrum pernix*]	Manual
15898896	Structure-based	*Sulfolobus solfataricus*	269	Conserved hypothetical protein	Manual
16264026	Template	*Sinorhizobium meliloti*	466	HYPOTHETICAL PROTEIN	Manual
17646150	Sequence- and structure-based	*Drosophila melanogaster*	506	Adenosine deaminase-related growth factor C	Manual
23093239	Sequence-based	*Drosophila melanogaster*	561	CG32178-PA	Manual
25009707	Sequence-based	*Drosophila melanogaster*	561	AT05468p	Manual
33593596	Structure-based	*Bordetella pertussis* Tohama I	523	Conserved hypothetical protein	Manual
40744823	Structure-based	*Aspergillus nidulans* FGSC A4	562	HYPOTHETICAL protein	Manual
47678365	Sequence-based	*Homo sapiens*	511	Cat eye syndrome critical region protein 1 [Homo sapiens]	Manual
49116836	Sequence- and structure-based	*Xenopus laevis*	510	Hypothetical protein	Manual

Tables listing all amidohydrolase and enolase superfamily targets can be found at http://salilab.org/projects/enspec/ (*HMM* Hidden Markov Model verification)

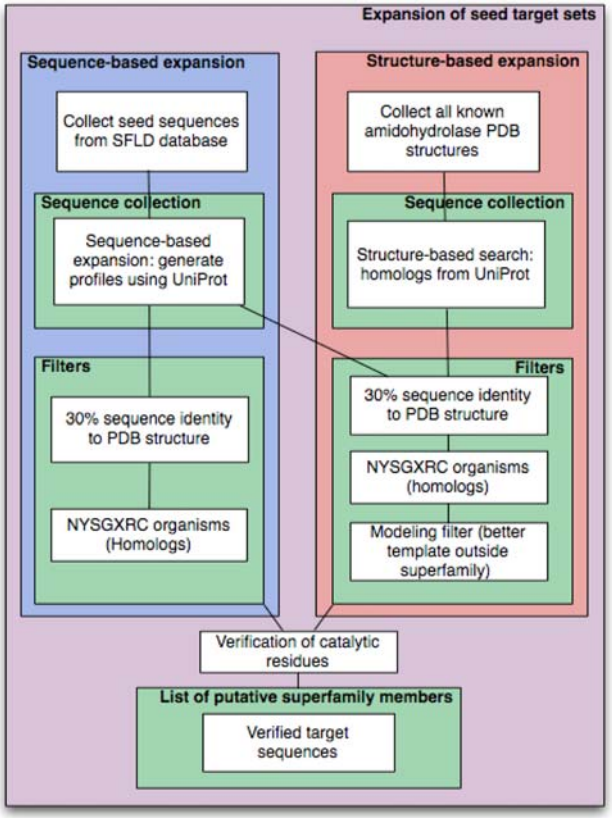


Fig. 1 Flowchart of the target expansion strategy of sequence-based target expansion (left) and structure-based target expansion (right)

255 sequences with less than 30% sequence identity to a known structure over at least 250 residues in length, originating from 98 organisms. These targets form 74 clusters at the 30% sequence identity cutoff, 126 clusters at 50% sequence identity, and 196 clusters at 80% sequence identity. The length distribution is 200 to 656 amino acid residues, with 90% of the sequences between 200 and 405 residues in length.

A complete list of the selected amidohydrolase and enolase superfamily targets can be found at http://salilab.org/projects/enspec/.

Structural genomics pipeline attrition

To date, 254 amidohydrolase (88%) and 206 enolase (80%) superfamily members have been attempted using the NYSGXRC/ENSPEC X-ray crystallographic structure determination pipeline. Progress to date and attrition rate at each stage of the pipeline are documented in Table 4 (June 2008). The project has not yet been completed, and a number of targets are still progressing through the pipeline. Also, a few targets in the target list have not yet been entered in the experimental pipeline. Therefore, the final overall success rate should be higher than that presented in Table 4. Experimental results for all NYSGXRC Community-nominated targets are updated weekly in PepcDB (http://pepcdb.pdb.org/).

Clear trends are observed in the success rates of crystallization and subsequent crystallographic structure determination for the amidohydrolase and enolase superfamily members. While only 38% of the purified targets were members of the enolase superfamily, they comprise 67% of the unique experimental structures. If crystals are obtained for an enolase superfamily member, there is a good chance that its structure will be successfully determined. On the other hand, for at least a quarter of the amidohydrolase proteins, we observed unusually broad peaks in the electrospray ionization (ESI) mass spectra of the intact proteins, indicative of heterogeneity in the preparation. Proteolytic digestion followed by tandem mass spectrometry analysis was carried out on the heterogeneous proteins; multiple sites of oxidation and methylation were identified with 90% of the protein sequence typically identified. These modifications were the source of the sample heterogeneity, and thus one reason for the limited success in obtaining usable crystallographic datasets from crystals of these amidohydrolases.

Of structural and functional interest was the fact that the oxidation sites were primarily located at histidine residues adjacent to Fe^{2+} ions in the presumed active sites of the amidohydrolases. Excess oxidation can be avoided using an alternate expression system (e.g. baculovirus) or adding excess Mn^{2+} and an iron chelator such as 2,2′-dipyridyl prior to induction during *E. coli* expression. In contrast, oxidation was not been observed in members of the enolase superfamily, since these proteins bind only a divalent metal ion such as Mg^{2+} or Mn^{2+} and not iron.

Analysis of the resulting crystallographic structures

Leverage of new crystallographic structures by modeling

To determine the impact of a structure on the structural mapping of the protein sequence space, we determine how many known protein sequences can be modeled based on the structure (i.e., the modeling leverage) (Table 2). Each enolase structure is a useful template for calculating comparative models for 2,500 to 3,200 other protein sequences in the NR database; a template is considered useful when the resulting model is based on a significant PSI-BLAST E-value (0.0001) or a favorable GA341 model score (>0.7). In contrast, the amidohydrolase superfamily structures fall into two categories: most are detectably related to 3,000–3,800 other proteins, but five structures (PDB Codes: 2I5G, 2Q01, 2Q6E, 2RAG, and 3B40) are related to a significantly smaller number of sequences (approximately 300–1,000).

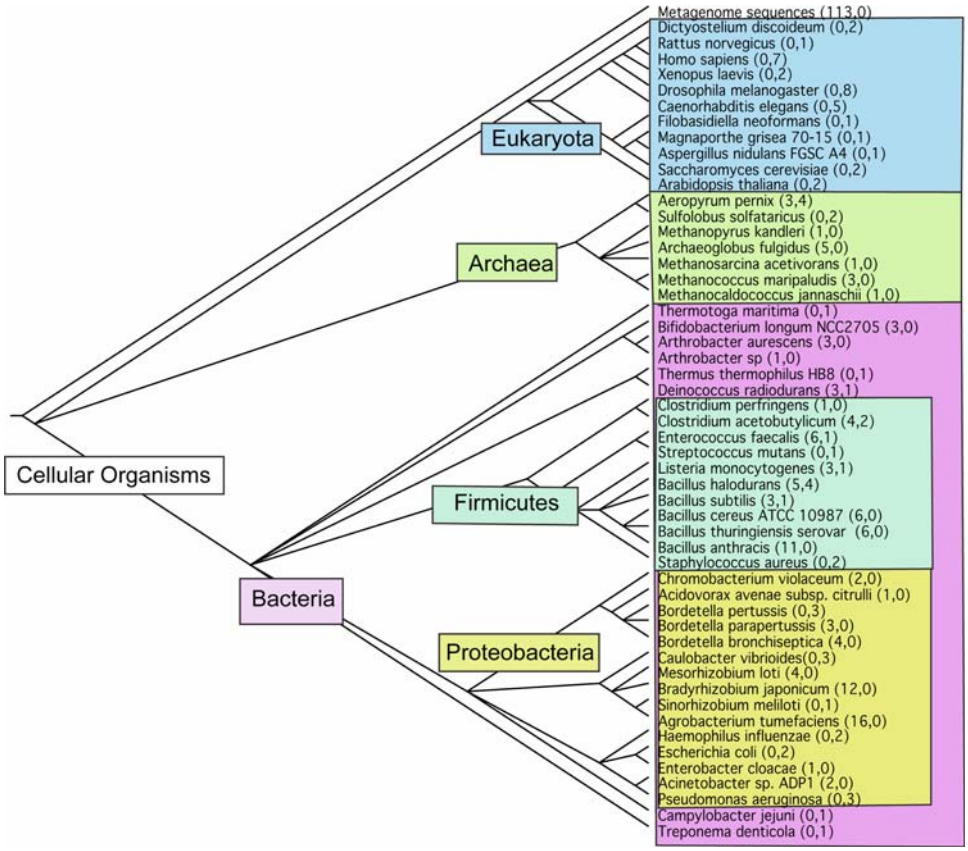


Fig. 2 Phylogenetic tree of the organisms for the selected amidohydrolase targets. The numbers in parentheses represent the number of targets for confirmed (first number) and putative (second number) amidohydrolase superfamily members. The tree was generated using the NCBI Taxonomy Browser [61]

A comparison of these numbers to the template-based modeling results for all NYSGXRC structures as of May 2007 (Table 5) shows that the average number of models per structure is significantly higher for the amidohydrolase and enolase superfamilies than for all structures determined by NYSGXRC (2,681 vs. 1,964). This difference reflects the relatively large sizes of the amidohydrolase and enolase superfamilies; according to the Superfamily database (http://supfam.org, [58]), across all of the superfamilies in the database, there are on average 1,770 protein sequences per superfamily.

Table 4 Success rates for the steps in the structural genomics pipeline as of June 2008

Step	Amidohydrolase superfamily		Enolase superfamily		Both superfamilies	
	Total	Fraction (%)	Total	Fraction (%)	Total	Fraction (%)
In pipeline	279		222		501	
Cloned	254	91	206	93	460	92
Expressed	225	88	177	86	402	87
Soluble	167	74	112	63	279	69
Purified	110	66	67	60	177	63
Crystallized	63	57	44	66	107	60
Unique structures	20	32	41	93	61	57
All structures	25		50		75	

Breaking down the modeling leverage by sequence identity reveals that the modeling leverage for the amidohydrolase and enolase superfamily structures is higher and lower than that for all NYSGXRC structures below and above the sequence identity cutoff of 30%, respectively. These differences are likely due in part to the relatively high diversity in the amidohydrolase and enolase superfamilies.

Upon initiation of the ENSPEC/NYSGXRC effort in June 2005, 45% of all known members of the amidohydrolase and enolase superfamilies were related to a known structure with a sequence identity higher than 30%. Due to the increased number of templates from the amidohydrolase and enolase superfamilies contributed by our consortia, this number increased to from 45% to 73%.

The total number of unique sequences modeled using the new amidohydrolase and enolase superfamily structures is 11,097, approximately 30% more than the number of known sequences from the amidohydrolase and enolase superfamilies. Among these additional sequences, we expect both members of other superfamilies with the TIM

Table 5 Comparison of template-based modeling statistics for the 61 ENSPEC/NYSGXRC structures and all 327 NYSGXRC structures (May 2007)

	Amidohydrolase and enolase superfamily members	All
Average number of sequences with acceptable models	2,681	1,964
Minimum/maximum number of sequences with acceptable models	189/3693	30/6320
Average number of sequences with >50% sequence identity, at least 50% coverage	15	20
Average number of sequences with 30–50% sequence identity, at least 50% coverage	59	113
Average number of sequences with <30% sequence identity, at least 50% coverage	2,572	1,400

An acceptable model is defined to be based on a significant PSI-BLAST E-value (0.0001) or a favorable GA341 model score (>0.7)

barrel fold, as well as currently unidentified members of the amidohydrolase and enolase superfamilies, because the sequence databases have been growing by approximately 50% since 2005, and also because we concentrated on selecting only targets from the NYSGXRC genomes in the target selection process for this project.

Distribution of targets over the amidohydrolase and enolase superfamilies

For large groups of related sequences, such as the amidohydrolase superfamily network-based visualization of their relationships is helpful in generating hypotheses about how various enzymes in the superfamily evolved, and on how closely the subgroups are related to each other. We have plotted cytoscape networks for the amidohydrolase and enolase superfamilies, based on clustering by sequence similarity, and marked previously known structures, and the final targets and the structures from this project (Fig. 3). For clarity, we circled a few distinct subgroups. Another network representation with all sub-group assignments can be found in the supplemental materials.

Many subgroups in the large amidohydrolase superfamily, such as the urease-like subgroup and the uronate isomerase-like subgroup, are distinctly separated from the other superfamily members. This separation can most simply be interpreted as the result of the extreme divergence of these subgroups; thus, they are "outliers" in the overall context of the superfamily (see below for further discussion of this subgroup).

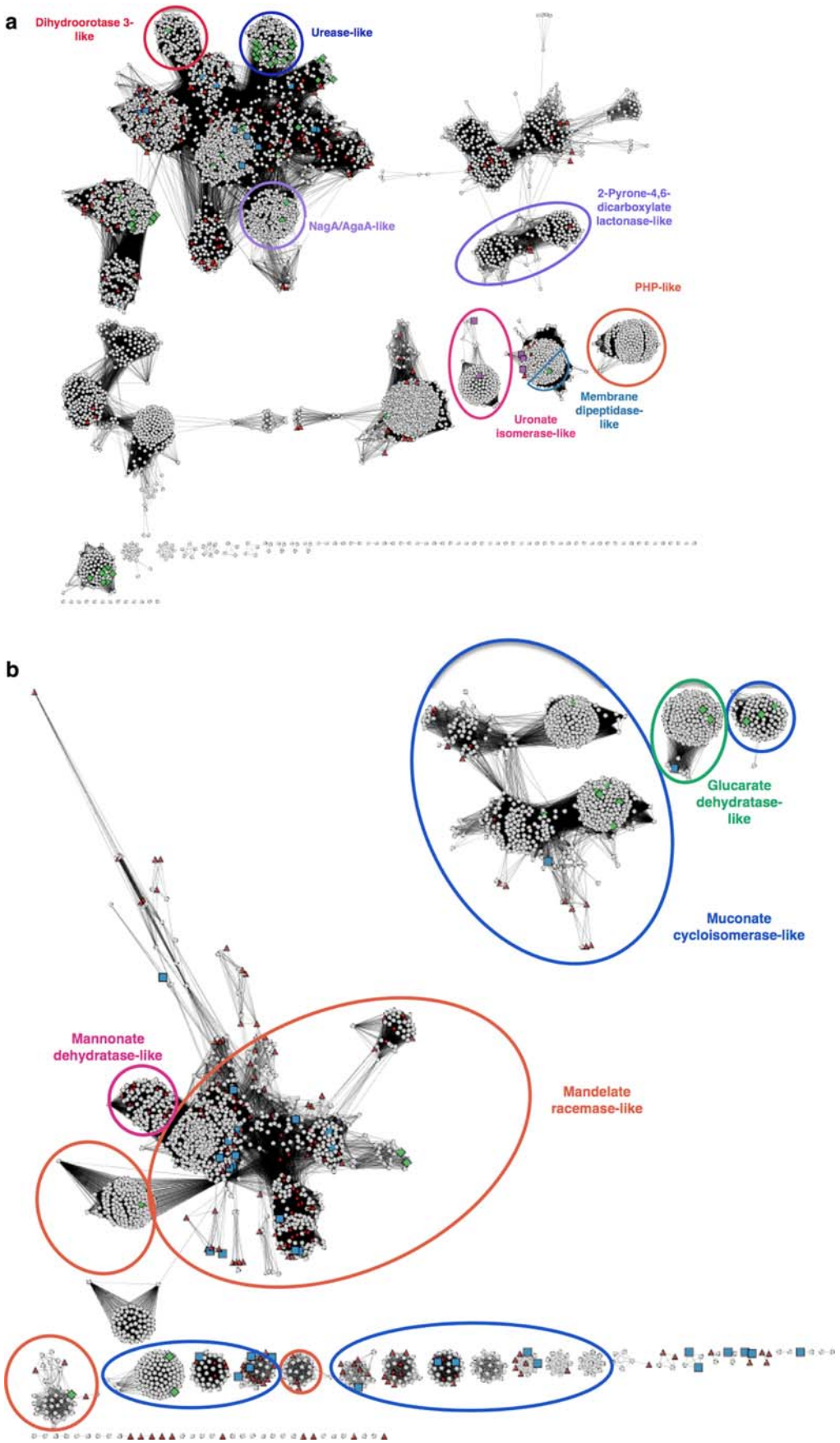
Four of the five divergent amidohydrolase structures with a considerably smaller number of homologs are separated from the main amidohydrolase network, even at the relatively non-stringent E-value cut-off of 10^{-10} required to visualize connections between nodes. Two of them (2Q01, 2Q6E) belong to the uronate isomerase-like subgroup. Another two of these structures (2RAG, 3B40) are clustered together with a number of unclassified sequences as well as several membrane dipeptidase-like amidohydrolase superfamily members, possibly indicating that these targets are additional members of the membrane dipeptidase subgroup. This subgroup membership is also supported by their annotation as putative dipeptidases in UniProt.

For the enolase superfamily, we chose to generate a cytoscape network that represents only four subgroups, containing the majority of the targets. The targets were mostly chosen from the mandelate racemase-like subgroup, because it is the largest subgroup with little previous structural coverage, and from the more divergent muconate cycloisomerase subgroup. The cytoscape networks illustrate that the targets and the resulting structures are indeed concentrated in regions of superfamily sequence space that lacked structural characterization prior to the start of the project, as desired for our target selection.

Examples of biological impact resulting from new structures obtained in this study

Amidohydrolase superfamily example: atypical uronate isomerase Bh0493

The enzymes in the uronate isomerase family are members of the amidohydrolase superfamily, although they are extremely diverged from other clusters of the amidohydrolase superfamily network (Fig. 3a). Target 9247a (gi 10173106, Bh0493) from *Bacillus halodurans* was identified by our structure-based expansion as a putative member of the amidohydrolase superfamily and has recently been experimentally confirmed as a uronate isomerase [29]. In most organisms, both glucuronic acid and galacturonic acid are first isomerized by a single uronate isomerase, followed by further modification by several sugar specific dehydrogenases and dehydratases. In *B. halodurans*, as in several other organisms, two uronate isomerase genes are found, in


a
Dihydroorotase 3-like
Urease-like
NagA/AgaA-like
2-Pyrone-4,6-dicarboxylate lactonase-like
PHP-like
Uronate isomerase-like
Membrane dipeptidase-like
b
Glucarate dehydratase-like
Muconate cycloisomerase-like
Mannonate dehydratase-like
Mandelate racemase-like

◀ **Fig. 3** **a** Cytoscape clustering for the amidohydrolase superfamily. The most homogeneous subgroups have been named. An additional figure with full subgroup coloring is available in Supplemental Materials. *Green diamonds* Structures determined prior to the start of the ENSPEC/NYSGXRC project in June 2005. *Red triangles* Superfamily members in the target list. *Purple squares* Five divergent structures determined by ENSPEC/NYSGXRC. *Blue squares* All other structures determined by ENSPEC/NYSGXRC. Ovals indicate subgroups: *red* dihydroorotase 3-like; *dark blue* urease-like; *purple* NagA/AgaA-like; *light-blue*: 2-Pyrone-4,6-dicarboxylate lactonase-like; *pink* uronate-isomerase-like; *orange* PHP-like; *delft-blue* membrane dipeptidase-like. **b** Cytoscape clustering for the enolase superfamily. Subgroup clusters are marked for four subgroups. The full subgroup assignments can be found in Supplemental Materials. *Green diamonds* Structures determined prior to the start of the ENSPEC/NYSGXRC project in June 2005. *Red triangles* Superfamily members in the target list. *Blue squares* All structures determined by ENSPEC/NYSGXRC. Ovals indicate subgroups: *pink* mannonate dehydratase-like; *orange* mandelate racemase-like; *blue* muconate cycloisomerase-like; *green* glucarate dehydratase-like

operons containing dehydrogenase as well as dehydratase enzymes, consistent with this assignment of activity. We characterized both uronate isomerase genes, a "typical" uronate isomerase, Bh0705, and Bh0493, an "outlier" relative to other characterized members of this family (Fig. 4a). Although the results showed that each enzyme can isomerize both substrates, galacturonate and glucuronate, the Bh0705 uronate isomerase preferentially isomerizes glucuronic acid (approximately 100 times faster than galacturonic acid). In contrast, Bh0493 isomerizes glucuronic acid and galacturonic acid almost equally efficiently. These observations indicate that in *B. halodurans,* the "typical" uronate isomerase (Bh0705) has specialized its catalytic activity to preferentially isomerize glucuronic acid, perhaps because the isomerization of galacturonic acid is sufficiently achieved by Bh0493.

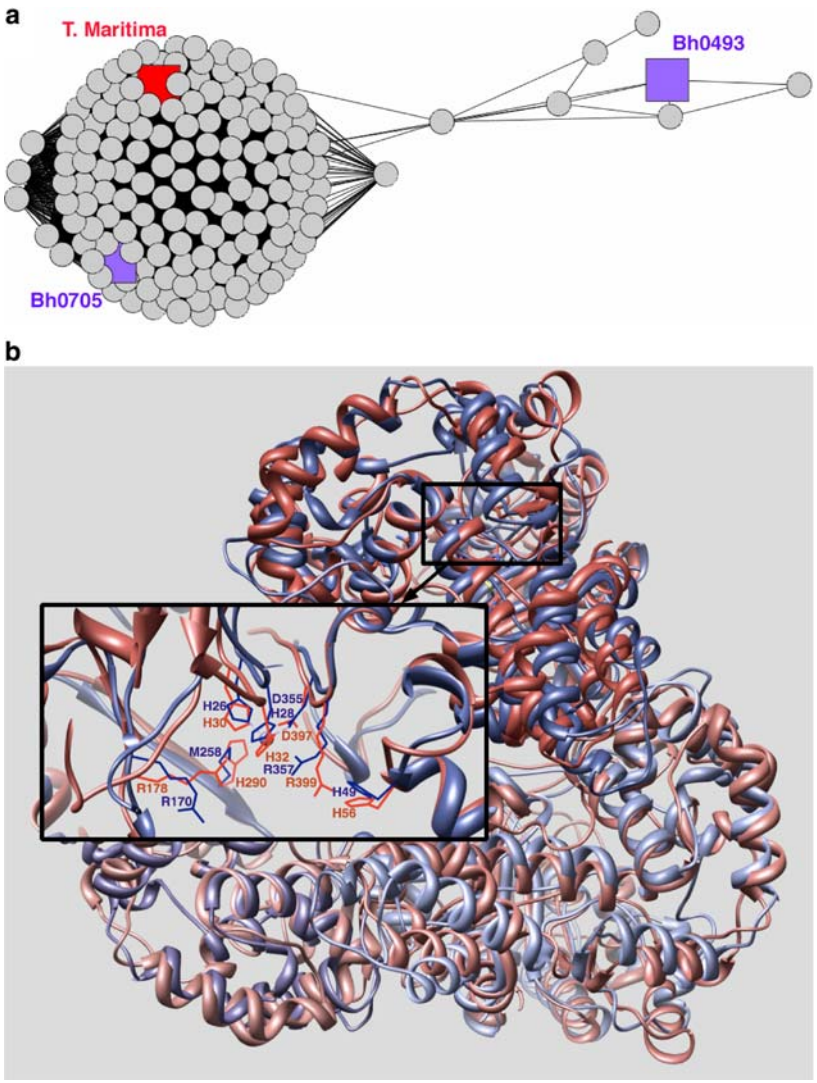


Fig. 4 **a** Cytoscape network showing the uronate isomerase family. The E-value threshold for displaying edges is 10^{-10}. The large cluster represents the "typical" uronate isomerases; sequences in this cluster are more similar to other members of the amidohydrolase superfamily than is Bh0493. Bh0705 is shown in purple and the structurally characterized enzyme from *Thermotoga maritima* is shown in red. On the right, the outlier uronate isomerase, Bh0493, is shown in purple along with a small number of sequences of unknown function. **b** Ribbon diagram [62] of a superposition of the trimeric structures of Bh0493 (2Q6E, blue) and a uronate isomerase from *Thermotoga maritima* (1J5S, red). The detailed box shows the active site residues of chain A including a Zn^{2+} ion for 2Q6E

To gain further insight into the structural differences between Bh0493 and the "typical" uronate isomerases (and between uronate isomerases and other members of the amidohydrolase superfamily), and in the absence of a structure of Bh0705, we compared the structure of Bh0493 (PDB codes 2Q08 and 2Q6E) to another "typical" uronate isomerase from *Thermotoga maritima* (PDB code 1J5S). As shown in Fig. 4b, the functionally important residues Arg170, Arg357 and His49, are conserved and cluster together within the enzyme active site both in the *T. maritima* enzyme and Bh0493. However, an additional metal-coordinating histidine that is usually found at the end of β-strand five in "typical" uronate isomerases (H290 in the 1J5S) is missing in Bh0493, which has a Met (M258) in that position. The Zn^{2+} ion is coordinated by two histidine residues (His28 and His26) plus Asp355. Guided by these structures, further biochemical and computational studies to examine the differences between these two types of uronate isomerases, and how they may be related to their different specificities, are currently in progress.

Enolase superfamily example: mandelate racemase subgroup

The SFLD currently describes 17 different families in the enolase superfamily, each performing a different overall reaction associated with different substrates and products. For the approximately 50% of the superfamily sequences whose functions are yet unknown, we estimate that roughly 15–20 novel functions (i.e. new families) will be identified. Across the superfamily, the sequences whose functions are not yet identified can be clustered into three primary subgroups and several smaller ones based on sequence and structural differences, including differences in the constellations of active site residues involved in binding specificity and catalysis [10]. In the mandelate racemase subgroup, most of the enzymes with characterized reactions are dehydratases acting on acid sugars, with the "outlier" enzyme being mandelate racemase itself. All structurally characterized members of the subgroup can be distinguished by a His-Asp dyad at the ends of β-strands six and seven that is associated with proton abstraction of substrates in the R-configuration [59]. Mandelate racemase and several acid sugar dehydratases that were previously structurally and functionally characterized also have a conserved Lys-X-Lys motif on β-strand two, with the second Lys in this motif involved in proton abstraction of substrates in the S-configuration [42]. Within this subgroup, we also observe divergence in this motif among several members of both known [32] and unknown function.

Three members of the mandelate racemase subgroup whose structures were determined by NYSGXRC, 2GL5 and 2O56 from *Salmonella typhimurium* and 2OX4 from *Zymomonas mobilis*, were found to have a Lys-Val-Asp sequence motif at this position, possibly indicating a different catalytic mechanism or yet other novel function(s). The three structures align within 50% sequence identity to each other. The next closest structures (30% sequence identity) are also members of the mandelate racemase subgroup: 2POZ from *Mesorhizobium loti* and 2POD from *Burkholderia pseudomallei* have Lys-Phe-Tyr and Lys-Ile-Trp motifs at this position, respectively, providing further evidence for divergent catalytic function(s). Their structures reveal details of differences relative to that of well-characterized subgroup members containing a "canonical" Lys-X-Lys motif, providing information expected to be useful in identifying their functions. Figure 5 shows superpositions of mandelate racemase with 2GL5 and 2POD, illustrating the differences in this motif. Guided by these new structures, these enzymes are now being further analyzed computationally and experimentally.

Conclusion

The Enzyme Specificity Consortium and the New York SGX Research Center for Structural Genomics made significant progress towards characterizing the structures and functions in the amidohydrolase and enolase superfamilies. New members of the amidohydrolase superfamily have been identified through a combination of sequence- and structure-based expansions of the pool of known superfamily members. The structure-based expansion was particularly successful in identifying previously unrecognized superfamily members. The 63 crystallographic structures from the structural genomics pipeline increased the fraction of the sequences in these two superfamilies that can be modeled based on at least 30% sequence identity from 45% to 73%.

As an annotation tool for the targets in the two superfamilies, template-based modeling of all sequences that had detectable homology to a known structure in the amidohydrolase or enolase superfamily allowed us to suggest previously un-annotated amidohydrolase sequences, several of which were subsequently verified by experiment, as shown for Bh0493 in this paper. This demonstrates the power of combining sequence- and structure-based approaches for the structural genomics of two large and diverse enzyme superfamilies.

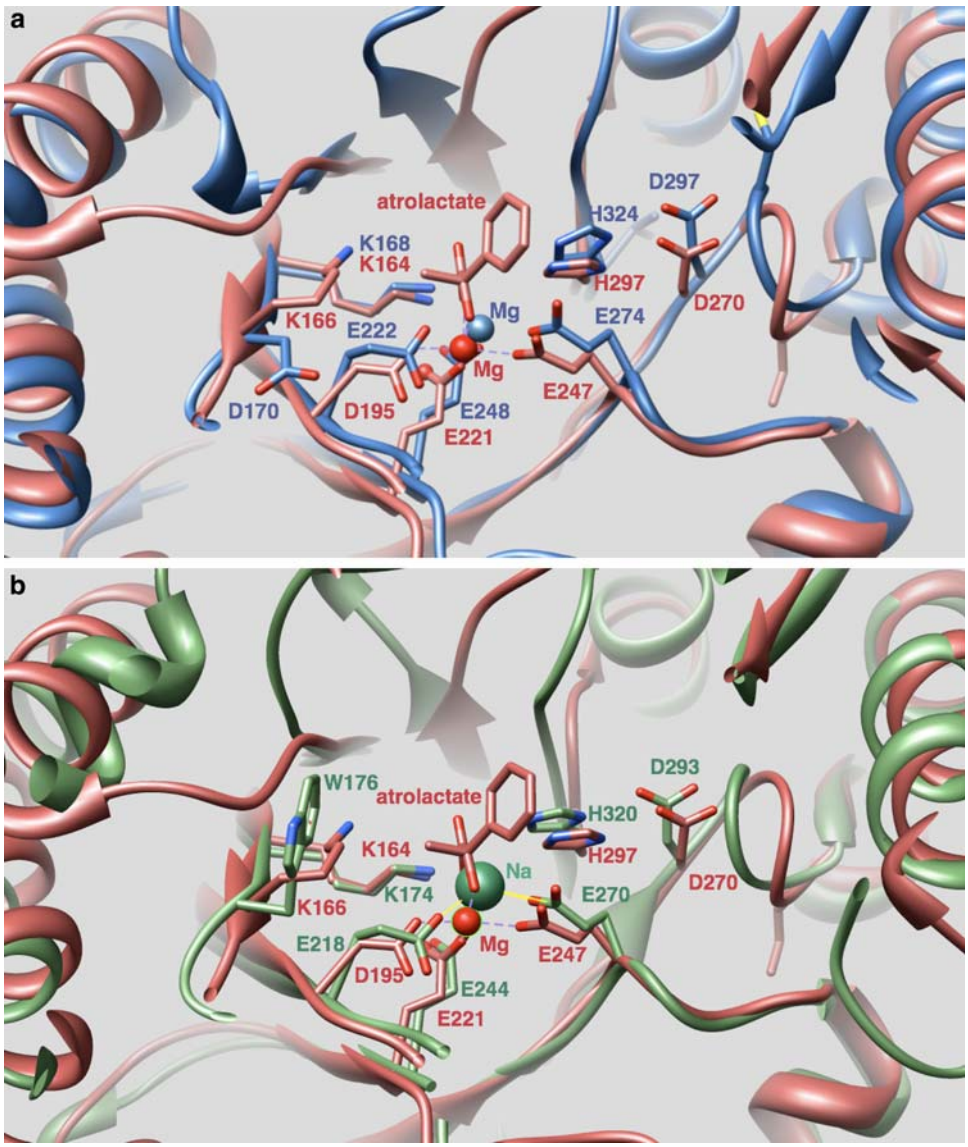


Fig. 5 Mandelate racemase bound to a substrate analog, atrolactate, (1MDR: red), is shown superimposed with two structures of unknown function. In both superpositions, active site metal ligands D195, E221, E247, the active site His-Asp dyad (H297, D270), and a Lys-X-Lys motif (K164, K166) conserved in 1MDR and other members of the mandelate racemase subgroup are labeled (1MDR numbering). **a** Superposition of 2GL5 (*blue*) with 1MDR shows conservation of all of these active site residues, except for the second Lys in the Lys-X-Lys motif of 1MDR, which is replaced in 2GL5 by Asp170. This residue faces away from the active site in 2GL5. **b** Superposition of 2POD (*green*) with 1MDR also shows conservation of all of listed residues, except for the second Lys in the Lys-X-Lys motif that is replaced in 2POD by W176

Acknowledgements This work was supported by NIH (U54 GM074945 (Principal Investigator: Stephen K. Burley) and P01 GM71790), the Sandler Family Supporting Foundation, Fight for Mike Foundation, Ron Conway, Hewlett-Packard, NetApp, IBM, and Intel. Use of the Advanced Photon Source was supported by the U.S. Department of Energy, Office of Basic Energy Sciences. Use of the SGX-CAT beam line facilities at Sector 31 of the APS was provided by SGX Pharmaceuticals, which constructed and operates the facility. Use of the NSLS beamline X29 was supported by the DOE and P41-EB-01979. Molecular graphics images were produced using the UCSF Chimera package from the Resource for Biocomputing, Visualization, and Informatics at the University of California, San Francisco (supported by NIH P41 RR-01081).

References

1. Andreeva A, Murzin AG (2006) Evolution of protein fold in the presence of functional constraints. Curr Opin Struct Biol 16:399–408
2. Gerlt JA, Babbitt PC (2001) Divergent evolution of enzymatic function: mechanistically diverse superfamilies and functionally distinct suprafamilies. Annu Rev Biochem 70:209–246
3. Glasner ME, Gerlt JA, Babbitt PC (2006) Evolution of enzyme superfamilies. Curr Opin Chem Biol 10:492–497
4. Todd AE, Orengo CA, Thornton JM (2001) Evolution of function in protein superfamilies, from a structural perspective. J Mol Biol 307:1113–1143
5. Pegg SC, Brown S, Ojha S, Huang CC, Ferrin TE, Babbitt PC et al (2005) Representing structure-function relationships in mechanistically diverse enzyme superfamilies. Pac Symp Biocomput 358–369
6. Holm L, Sander C (1997) An evolutionary treasure: unification of a broad set of amidohydrolases related to urease. Proteins 28:72–82

7. Seibert CM, Raushel FM (2005) Structural and catalytic diversity within the amidohydrolase superfamily. Biochemistry 44:6383–6391
8. Pegg SC, Brown SD, Ojha S, Seffernick J, Meng EC, Morris JH, Chang PJ, Huang CC, Ferrin TE, Babbitt PC (2006) Leveraging enzyme structure–function relationships for functional inference and experimental design: the structure–function linkage database. Biochemistry 45:2545–2555
9. Babbitt PC, Hasson MS, Wedekind JE, Palmer DR, Barrett WC, Reed GH, Rayment I, Ringe D, Kenyon GL, Gerlt JA (1996) The enolase superfamily: a general strategy for enzyme-catalyzed abstraction of the alpha-protons of carboxylic acids. Biochemistry 35:16489–16501
10. Gerlt JA, Babbitt PC, Rayment I (2005) Divergent evolution in the enolase superfamily: the interplay of mechanism and specificity. Arch Biochem Biophys 433:59–70
11. Vick JE, Gerlt JA (2007) Evolutionary potential of (beta/alpha)8-barrels: stepwise evolution of a "new" reaction in the enolase superfamily. Biochemistry 46:14589–14597
12. Akana J, Fedorov AA, Fedorov E, Novak WR, Babbitt PC, Almo SC, Gerlt JA (2006) D-Ribulose 5-phosphate 3-epimerase: functional and structural relationships to members of the ribulose-phosphate binding (beta/alpha)8-barrel superfamily. Biochemistry 45:2493–2503
13. Almo SC, Bonanno JB, Sauder JM, Emtage S, Dilorenzo TP, Malashkevich V, Wasserman SR, Swaminathan S, Eswaramoorthy S, Agarwal R, Kumaran D, Madegowda M, Ragumani S, Patskovsky Y, Alvarado J, Ramagopal UA, Faber-Barata J, Chance MR, Sali A, Fiser A, Zhang ZY, Lawrence DS, Burley SK (2007) Structural genomics of protein phosphatases. J Struct Funct Genomics 8:121–140
14. Bonanno JB, Almo SC, Bresnick A, Chance MR, Fiser A, Swaminathan S, Jiang J, Studier FW, Shapiro L, Lima CD, Gaasterland TM, Sali A, Bain K, Feil I, Gao X, Lorimer D, Ramos A, Sauder JM, Wasserman SR, Emtage S, D'Amico KL, Burley SK (2005) New York-Structural GenomiX Research Consortium (NYSGXRC): a large scale center for the protein structure initiative. J Struct Funct Genomics 6:225–232
15. Brown SD, Gerlt JA, Seffernick JL, Babbitt PC (2006) A gold standard set of mechanistically diverse enzyme superfamilies. Genome Biol 7:R8
16. Gerlt JA (2007) A protein structure (or function?) initiative. Structure 15:1353–1356
17. Glasner ME, Fayazmanesh N, Chiang RA, Sakai A, Jacobson MP, Gerlt JA, Babbitt PC (2006) Evolution of structure and function in the o-succinylbenzoate synthase/N-acylamino acid racemase family of the enolase superfamily. J Mol Biol 360:228–250
18. Glasner ME, Gerlt JA, Babbitt PC (2007) Mechanisms of protein evolution and their application to protein engineering. Adv Enzymol Relat Areas Mol Biol 75:193–239 xii–xiii
19. Hall RS, Brown S, Fedorov AA, Fedorov EV, Xu C, Babbitt PC, Almo SC, Raushel FM (2007) Structural diversity within the mononuclear and binuclear active sites of N-acetyl-D-glucosamine-6-phosphate deacetylase. Biochemistry 46:7953–7962
20. Hermann JC, Ghanem E, Li Y, Raushel FM, Irwin JJ, Shoichet BK (2006) Predicting substrates by docking high-energy intermediates to enzyme structures. J Am Chem Soc 128:15882–15891
21. Hermann JC, Marti-Arbona R, Fedorov AA, Fedorov E, Almo SC, Shoichet BK, Raushel FM (2007) Structure-based activity prediction for an enzyme of unknown function. Nature 448:775–779
22. Imker HJ, Fedorov AA, Fedorov EV, Almo SC, Gerlt JA (2007) Mechanistic diversity in the RuBisCO superfamily: the "enolase" in the methionine salvage pathway in *Geobacillus kaustophilus*. Biochemistry 46:4077–4089
23. Irwin JJ, Raushel FM, Shoichet BK (2005) Virtual screening against metalloenzymes for inhibitors and substrates. Biochemistry 44:12316–12328
24. Li Y, Raushel FM (2005) Inhibitors designed for the active site of dihydroorotase. Bioorg Chem 33:470–483
25. Liao RZ, Yu JG, Raushel FM, Himo F (2008) Theoretical investigation of the reaction mechanism of the dinuclear zinc enzyme dihydroorotase. Chemistry 14:4287–4292
26. Marti-Arbona R, Raushel FM (2006) Mechanistic characterization of N-formimino-L-glutamate iminohydrolase from *Pseudomonas aeruginosa*. Biochemistry 45:14256–14262
27. Marti-Arbona R, Thoden JB, Holden HM, Raushel FM (2005) Functional significance of Glu-77 and Tyr-137 within the active site of isoaspartyl dipeptidase. Bioorg Chem 33:448–458
28. Marti-Arbona R, Xu C, Steele S, Weeks A, Kuty GF, Seibert CM, Raushel FM (2006) Annotating enzymes of unknown function: N-formimino-L-glutamate deiminase is a member of the amidohydrolase superfamily. Biochemistry 45:1997–2005
29. Nguyen TT, Brown S, Fedorov AA, Fedorov EV, Babbitt PC, Almo SC, Raushel FM (2008) At the periphery of the amidohydrolase superfamily: Bh0493 from *Bacillus halodurans* catalyzes the isomerization of D-galacturonate to D-tagaturonate. Biochemistry 47:1194–1206
30. Nowlan C, Li Y, Hermann JC, Evans T, Carpenter J, Ghanem E, Shoichet BK, Raushel FM (2006) Resolution of chiral phosphate, phosphonate, and phosphinate esters by an enantioselective enzyme library. J Am Chem Soc 128:15892–15902
31. Porter TN, Li Y, Raushel FM (2004) Mechanism of the dihydroorotase reaction. Biochemistry 43:16285–16292
32. Rakus JF, Fedorov AA, Fedorov EV, Glasner ME, Vick JE, Babbitt PC, Almo SC, Gerlt JA (2007) Evolution of enzymatic activities in the enolase superfamily: D-Mannonate dehydratase from *Novosphingobium aromaticivorans*. Biochemistry 46:12896–12908
33. Sakai A, Xiang DF, Xu C, Song L, Yew WS, Raushel FM, Gerlt JA (2006) Evolution of enzymatic activities in the enolase superfamily: N-succinylamino acid racemase and a new pathway for the irreversible conversion of D- to L-amino acids. Biochemistry 45:4455–4462
34. Song L, Kalyanaraman C, Fedorov AA, Fedorov EV, Glasner ME, Brown S, Imker HJ, Babbitt PC, Almo SC, Jacobson MP, Gerlt JA (2007) Prediction and assignment of function for a divergent N-succinyl amino acid racemase. Nat Chem Biol 3:486–491
35. Thoden JB, Taylor Ringia EA, Garrett JB, Gerlt JA, Holden HM, Rayment I (2004) Evolution of enzymatic activity in the enolase superfamily: structural studies of the promiscuous *o*-succinylbenzoate synthase from Amycolatopsis. Biochemistry 43:5716–5727
36. Tyagi R, Eswaramoorthy S, Burley SK, Raushel FM, Swaminathan S (2008) A common catalytic mechanism for proteins of the *Hut*I family. Biochemistry 47:5608–5615
37. Vick JE, Schmidt DM, Gerlt JA (2005) Evolutionary potential of (beta/alpha)8-barrels: in vitro enhancement of a "new" reaction in the enolase superfamily. Biochemistry 44:11722–11729
38. Weeks A, Lund L, Raushel FM (2006) Tunneling of intermediates in enzyme-catalyzed reactions. Curr Opin Chem Biol 10:465–472
39. Williams L, Nguyen T, Li Y, Porter TN, Raushel FM (2006) Uronate isomerase: a nonhydrolytic member of the amidohydrolase superfamily with an ambivalent requirement for a divalent metal ion. Biochemistry 45:7453–7462
40. Yew WS, Fedorov AA, Fedorov EV, Rakus JF, Pierce RW, Almo SC, Gerlt JA (2006) Evolution of enzymatic activities in the enolase superfamily: L-fuconate dehydratase from *Xanthomonas campestris*. Biochemistry 45:14582–14597

41. Yew WS, Fedorov AA, Fedorov EV, Wood BM, Almo SC, Gerlt JA (2006) Evolution of enzymatic activities in the enolase superfamily: D-tartrate dehydratase from *Bradyrhizobium japonicum*. Biochemistry 45:14598–14608
42. Yew WS, Fedorov AA, Fedorov EV, Almo SC, Gerlt JA (2007) Evolution of enzymatic activities in the enolase superfamily: L-talarate/galactarate dehydratase from *Salmonella typhimurium* LT2. Biochemistry 46:9564–9577
43. Norvell JC, Berg JM (2007) Update on the protein structure initiative. Structure 15:1519–1522
44. Berman H, Henrick K, Nakamura H, Markley JL (2007) The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. Nucleic Acids Res 35:D301–D303
45. Eswar N, John B, Mirkovic N, Fiser A, Ilyin VA, Pieper U, Stuart AC, Marti-Renom MA, Madhusudhan MS, Yerkovich B, Sali A (2003) Tools for comparative protein structure modeling and analysis. Nucleic Acids Res 31:3375–3380
46. Wu CH, Apweiler R, Bairoch A, Natale DA, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, Martin MJ, Mazumder R, O'Donovan C, Redaschi N, Suzek B (2006) The Universal Protein Resource (UniProt): an expanding universe of protein information. Nucleic Acids Res 34:D187–D191
47. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389–3402
48. Pieper U, Eswar N, Davis FP, Braberg H, Madhusudhan MS, Rossi A, Marti-Renom M, Karchin R, Webb BM, Eramian D, Shen MY, Kelly L, Melo F, Sali A (2006) MODBASE: a database of annotated comparative protein structure models and associated resources. Nucleic Acids Res 34:D291–D295
49. Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, Eisen JA, Wu D, Paulsen I, Nelson KE, Nelson W, Fouts DE, Levy S, Knap AH, Lomas MW, Nealson K, White O, Peterson J, Hoffman J, Parsons R, Baden-Tillson H, Pfannkoch C, Rogers YH, Smith HO (2004) Environmental genome shotgun sequencing of the Sargasso Sea. Science 304:66–74
50. Lee J, Ware B (2003) Open source web development with LAMP: using Linux, Apache, MySQL, Per, and PHP. Addison-Wesley, Boston
51. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 13:2498–2504
52. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL (2008) GenBank. Nucleic Acids Res 36:D25–D30
53. Atkinson HJ, Morris JH, Ferrin TE, Babbitt PC et al (2009) Using sequence similarity networks for visualization of relationships across diverse protein superfamilies. PLoS One (submitted)
54. Martin AC, MacArthur M W, Thornton JM et al (1997) Assessment of comparative modeling in CASP2. Proteins, Suppl 1:14–28
55. Sanchez R, Sali A (1997) Advances in comparative protein-structure modelling. Curr Opin Struct Biol 7:206–214
56. Vitkup D, Melamud E, Moult J, Sander C (2001) Completeness in structural genomics. Nat Struct Biol 8:559–566
57. Nagano N, Orengo CA, Thornton JM (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
58. Gough J, Karplus K, Hughey R, Chothia C (2001) Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure. J Mol Biol 313:903–919
59. Schafer SL, Barrett WC, Kallarakal AT, Mitra B, Kozarich JW, Gerlt JA, Clifton JG, Petsko GA, Kenyon GL (1996) Mechanism of the reaction catalyzed by mandelate racemase: structure and mechanistic properties of the D270 N mutant. Biochemistry 35:5662–5669
60. Melo F, Sanchez R, Sali A (2002) Statistical potentials for fold assessment. Protein Sci 11:430–448
61. Wheeler DL, Barrett T, Benson DA, Bryant SH, Canese K, Chetvernin V, Church DM, Dicuccio M, Edgar R, Federhen S, Feolo M, Geer LY, Helmberg W, Kapustin Y, Khovayko O, Landsman D, Lipman DJ, Madden TL, Maglott DR, Miller V, Ostell J, Pruitt KD, Schuler GD, Shumway M, Sequeira E, Sherry ST, Sirotkin K, Souvorov A, Starchenko G, Tatusov RL, Tatusova TA, Wagner L, Yaschenko E (2008) Database resources of the National Center for Biotechnology Information. Nucleic Acids Res 36:D13–D21
62. Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, Ferrin TE (2004) UCSF Chimera—a visualization system for exploratory research and analysis. J Comput Chem 25:1605–1612