Electronic Supplementary Material (ESI) for Integrative Biology.

Protein structure networks provide insight into active site flexibility in novel esterase/lipases from the carnivorous plant *Drosera capensis*

Vy T. Duong,[†,‡] Megha H. Unhelkar,[†] John E. Kelly,[†] Suhn H. Kim,[†] Carter T. Butts,[*,¶,§,‖] and Rachel W. Martin[*,†,‡]

Department of Chemistry, University of California, Irvine, Irvine, CA 92697, Department of Molecular Biology and Biochemistry, University of California, Irvine, Irvine CA 92697, Department of Sociology, UC Irvine, Irvine, CA, 92697 USA, Department of Electrical Engineering and Computer Science, UC Irvine, Irvine, CA, 92697 USA, and Department of Statistics, UC Irvine, Irvine, CA, 92697 USA

[*]To whom correspondence should be addressed
[†]Department of Chemistry, University of California, Irvine, Irvine, CA 92697
[‡]Department of Molecular Biology and Biochemistry, University of California, Irvine, Irvine CA 92697
[¶]Department of Sociology, UC Irvine, Irvine, CA, 92697 USA
[§]Department of Electrical Engineering and Computer Science, UC Irvine, Irvine, CA, 92697 USA
[‖]Department of Statistics, UC Irvine, Irvine, CA, 92697 USA

Sequence Alignments

Sequence alignments for the esterase/lipases from *D. capensis* are shown along with annotation reference sequences from other plants. Cluster 1 (Figure S1) contains enzymes with the traditional GDSL motif, including GDL1_CARPA from *Carica papaya*. Cluster 2 (Figure S2) contains only sequences from *D. capensis*, while Cluster 3 (Figure S3) contains two reference sequences from *Arabidopsis thaliana*. Cluster 4 is split into two figures for legibility (Figures S4 and S5). The alignment figures are annotated to highlight chemical properties of the amino acid residues as well as important sequence features. The amino acid attributes are color-coded as follows: cysteines are yellow, positively charged residues are blue, negatively charged residues are red, hydrophobic residues are green, and all others are black. Highly conserved residues are indicated with a dot above the sequence position. The catalytic triad residues are marked with colored arrows. SignalP 4.1[1] is used to predict the signal peptide cleavage site, which is specified by underlining the residues on either end of the cleavage point. The signal peptide itself is highlighted in light orange. Strikethrough text indicates sequence regions that are absent in the active enzyme, in this case the N-terminal signal peptide that is expressed but removed during maturation. Functional blocks I-IV are highlighted with colored boxes. Annotations were performed by homology to the annotations reference sequences from *C. papaya* and *A. thaliana* found in the UniProt database and identified by their UniProt IDs.

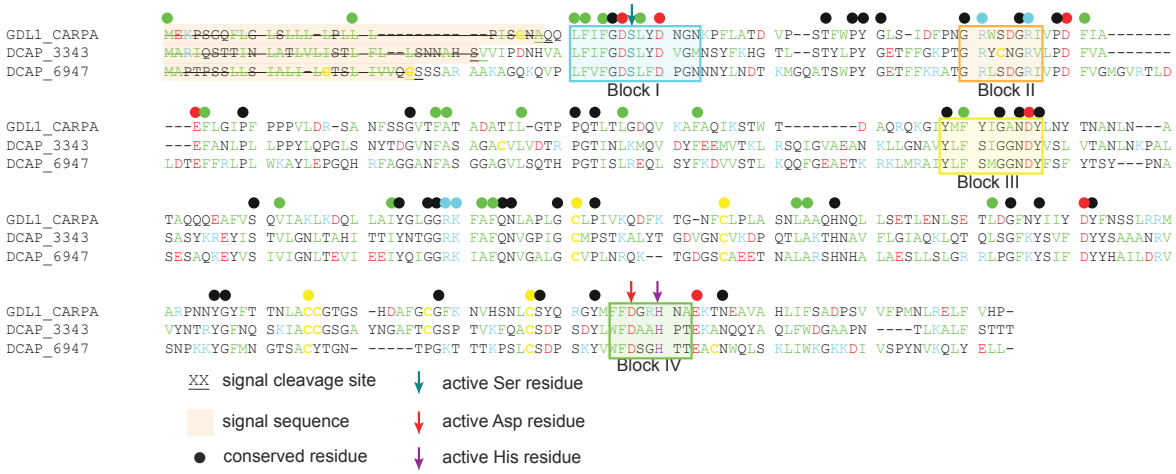

Figure S 1: Sequence alignment for Cluster 1 esterase/lipases, annotated by homology to the reference sequence GDL1_CARPA. The four functional blocks that are critical for enzyme function are highlighted using outlined colored boxes. The N-terminal signal peptide is highlighted in light orange. Colored arrows indicate the catalytic triad residues. Conserved residues are marked using colored dots: acidic (red), basic (blue), hydrophobic (green), and hydrophilic (black) residues.

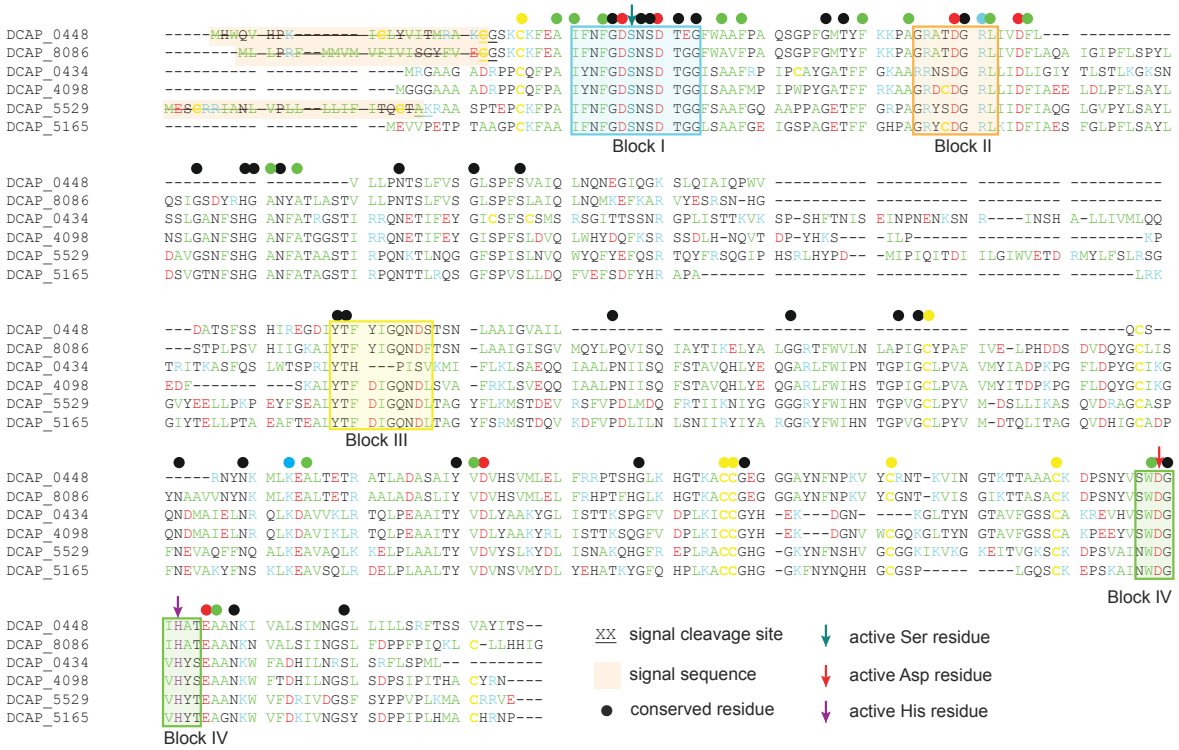

Figure S 2: Sequence alignment and annotation for Cluster 2. The four block regions are determined by sequence conservation and outlined with colored boxes. Three *D. capensis* esterase/lipases contain the N-terminal signal sequence (highlighted in light orange) and three lack it. The catalytic triad is indicated using colored arrows. Colored dots denote conserved residues.

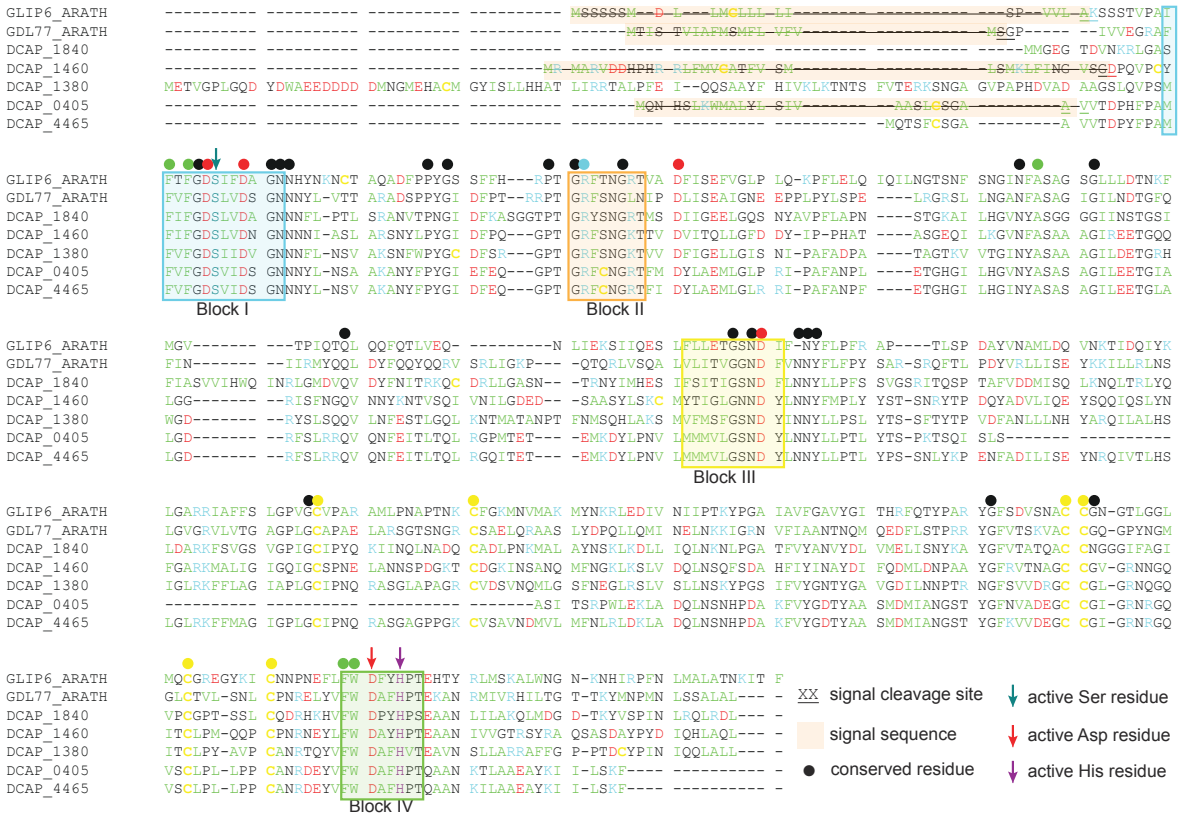

Figure S 3: Sequence alignment and annotation for Cluster 3. Reference sequences are GLIP6_ARATH and GDL77_ARATH. All but three Cluster 3 esterase/lipases contain a N-terminal signal peptide (highlighted in light orange). Functional block regions are outlined using colored boxes. Colored dots indicate conserved residues.

Figure S 4: Sequence alignment and annotation of Cluster 4a (first set), annotated by homology to EXL3_ARATH. Cluster 4 is separated into two parts (4a and 4b) for clarity. Block regions I-IV are shown in colored boxes with active site residues marked by colored arrows. Colored dots indicate conserved residues. When present, the N-terminal signal peptide is highlighted in light orange.

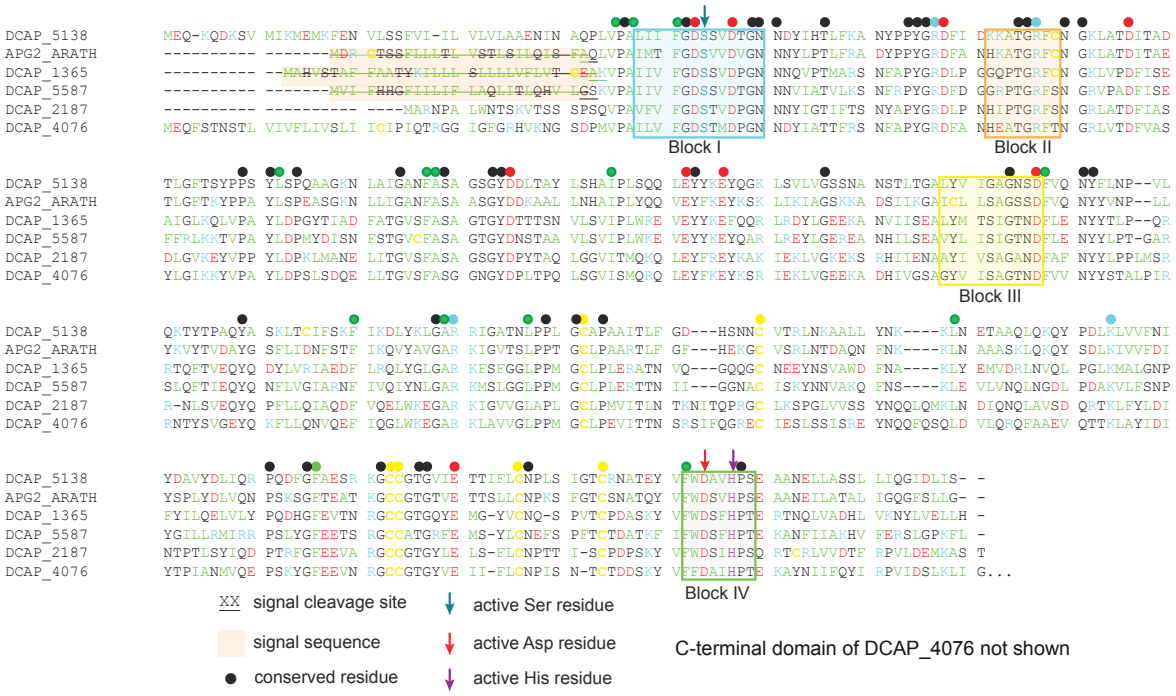

Figure S 5: Sequence alignment and annotation of Cluster 4b (second set), annotated by homology to APG2_ARATH. Cluster 4 is separated into two parts (4a and 4b) for clarity. Block regions I-IV are shown in colored boxes with active site residues marked by colored arrows. Colored dots indicate conserved residues. When present, the N-terminal signal peptide is highlighted in light orange. DCAP_4076 has an additional C-terminal domain (shown in Figure S8).

Preliminary Structural Models and *In silico* Maturation

Preliminary models for the esterase/lipases were produced using the online Robetta implementation[2] of Rosetta[3]. The Rosetta structures contain the full sequences, including the N-terminal signal peptides that are cleaved during maturation. We performed *in silico* maturation, which we have previously described for cysteine proteases,[4] for each protein. The initial Rosetta structure for each enzyme includes the signal peptide and lacks post-translational modifications. During *in silico* maturation, the signal sequence is removed and the structure is equilibrated for 500 ps in explicit TIP3P solvent using NAMD.[5] Figures of predicted structures were generated using Chimera.[6] Figure S6A shows the workflow of the overall enzyme discovery process. Panels (B) and (C) show an example of a Cluster

2 esterase/lipase, DCAP_8086, before (B) and after (C) the *in silico* maturation process. Further comparison of a Cluster 3 esterase/lipase (DCAP_1460) to Cluster 4 enzymes and a cutin synthase from *Solanum lycopersicum* (tomato), G1DEX3_SOLLC, is shown in Figure S7. Functional sequence blocks DCAP_1460 and G1DEX3_SOLLC are highlighted by color (Figure S7). DCAP_4076, has an additional C-terminal domain. A PSI-BLAST search for the sequence of this domain indicated that it is related to the negative regulator of systemic acquired resistance proteins previously discovered in other plants,[7] with approximately 36% sequence identity to the SNI1 proteins from *Arabidopsis thaliana* (Uniprot ID: SNI1_ARATH) and *Glycine max* (Uniprot ID: Q0ZFU8_SOYBN). The *Arabidopsis* protein negatively regulates DNA recombination and gene expression during short-term stress responses. It has been suggested that SNI1_ARATH provides a scaffold for other proteins involved in regulation of transcription to bind;[8] it is possible that this domain is playing a similar role here. DCAP_4076 lacks the N-terminal secretion signal common to many of the esterase/lipases, suggesting an intracellular function (Figure S8).

The template structures used by Rosetta to calculate the predicted structures for a representative esterase / lipase, DCAP_0434, are tabulated in Supplementary Table S1.

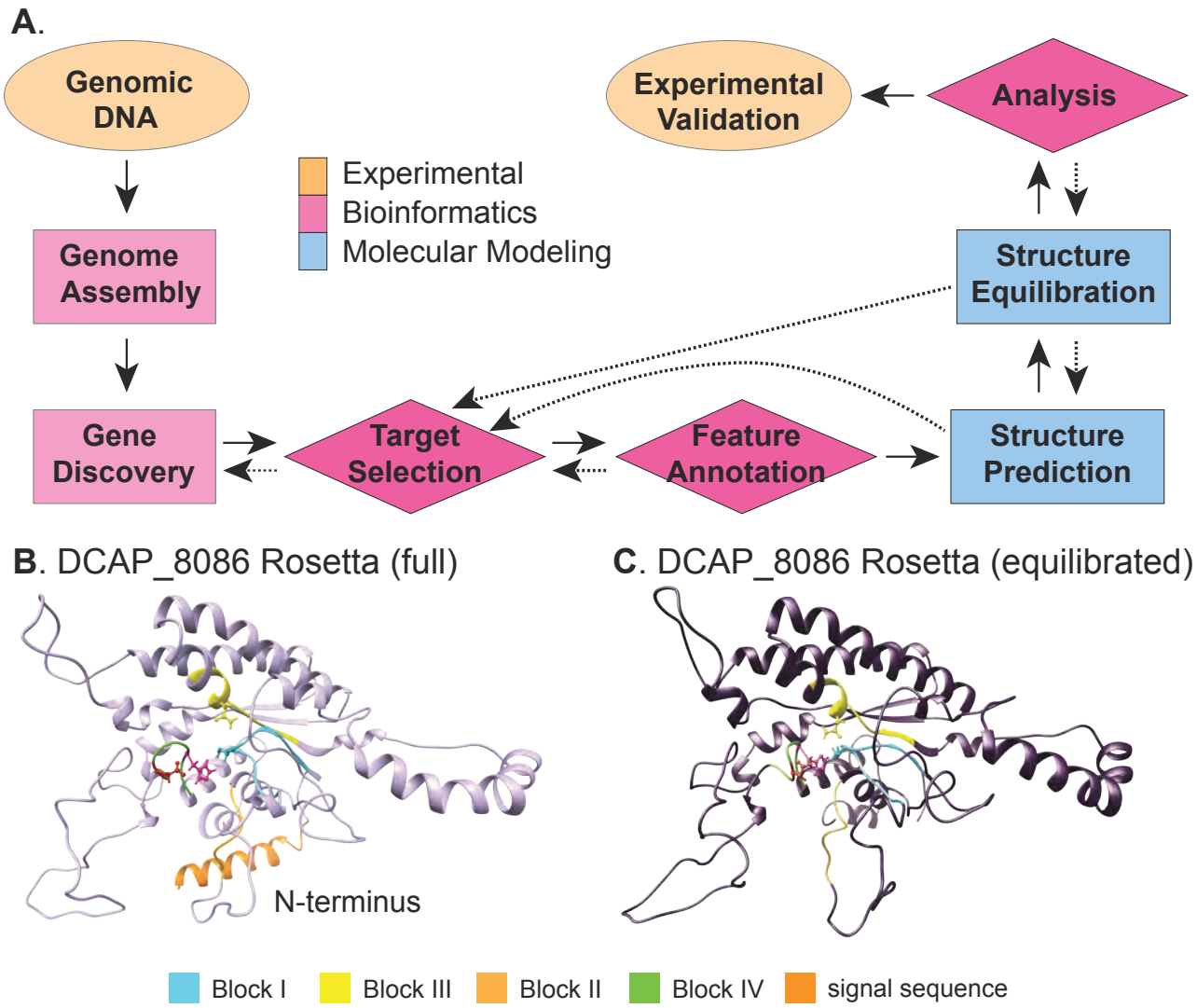

Figure S 6: (A) Flow chart illustrating the overall strategy for identifying enzymatic targets from genomic DNA. The workflow is indicated with solid arrows, while dotted arrows represent steps where information from a later stage of the pipeline enables refinement of earlier stages in an iterative manner. After genome sequencing, assembly, and gene discovery, target proteins are identified based on putative enzymatic activity. Functional sequence features are identified by analogy to annotation reference sequences found in the UniProt database. Structures are predicted using the Rosetta software, and equilibrated in explicit solvent after removal of sequence regions not present in the mature enzyme. Structures are compared using network analytic methods, enabling strategic selection of enzymes for experimental characterization in a future study. (B) DCAP_8086 before and (C) after *in silico* maturation. The light orange helix in part A is the N-terminal signal sequence, which is cleaved upon maturation. Important residues are color-coded as follows: dark cyan (catalytically active serine), red (active site aspartic acid), purple (active site histidine).

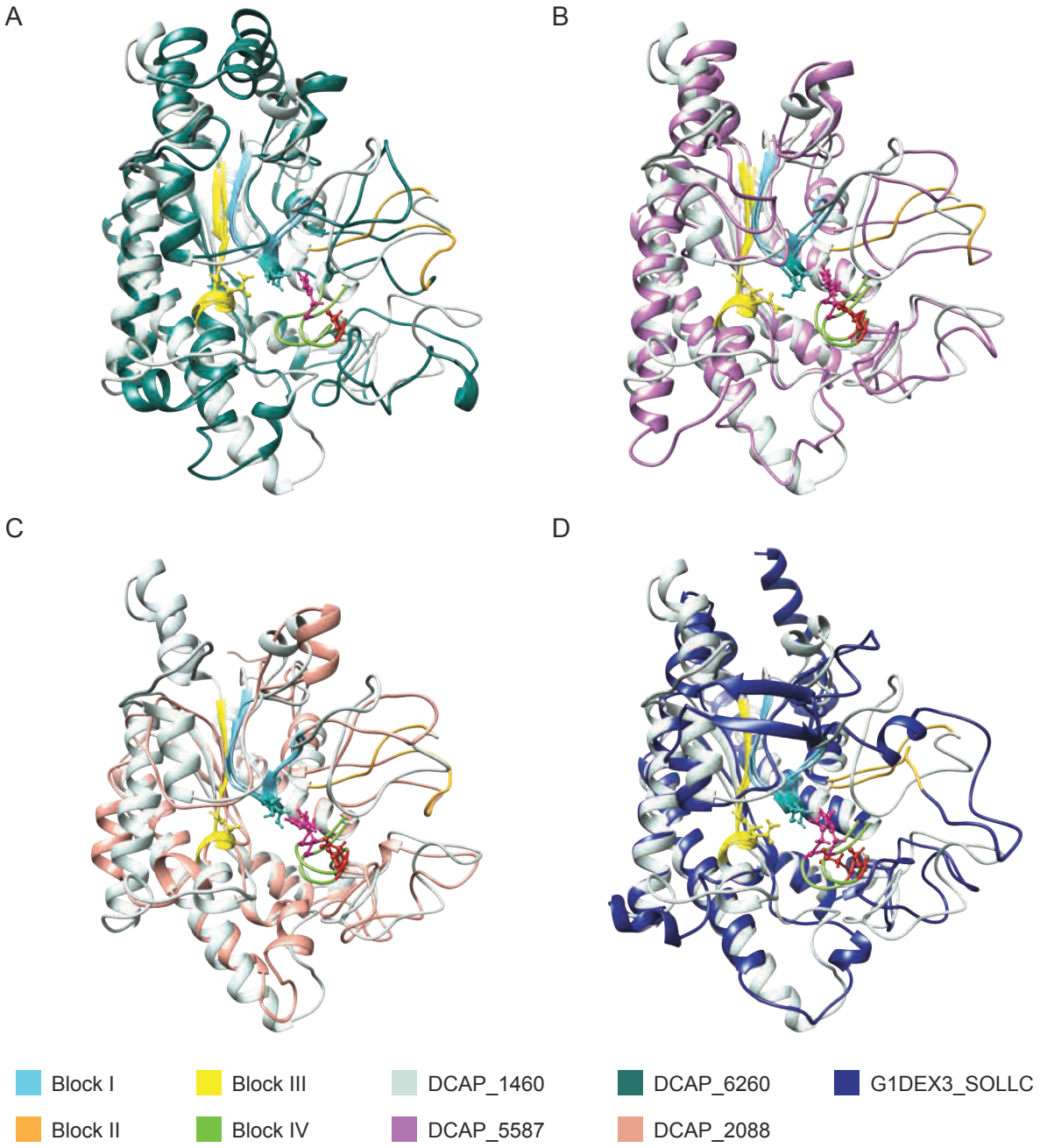

Figure S 7: Comparison of DCAP_1460 (Cluster 3) to *D. capensis* esterase/lipases from each of the other clusters. These pairwise alignments of structural models provide an indication of the type and magnitude of structural differences between clusters: in general, the overall fold and secondary structural elements is conserved, although considerable variation can be observed in their relative positions and the conformations of loops and termini. Alignment was performed using the matchmaker feature of Chimera with default settings.[6] Functional block regions I-IV are colored accordingly while the catalytic triad (Ser-His-Asp) residues are colored dark cyan, red, and purple. Active site residues are located in block I and IV, binding residues in block II-III. A. Comparison of DCAP_1460 to esterase/lipase DCAP_6260 (Cluster 4a). B. Comparison of DCAP_1460 to DCAP_5587 (Cluster 4b). C. Comparison of DCAP_1460 to DCAP_2088 (Cluster 4a). D. Comparison of DCAP_1460 to model esterase/lipase, G1DEX3_SOLLC, from *Solanum lycopersicum* (tomato).

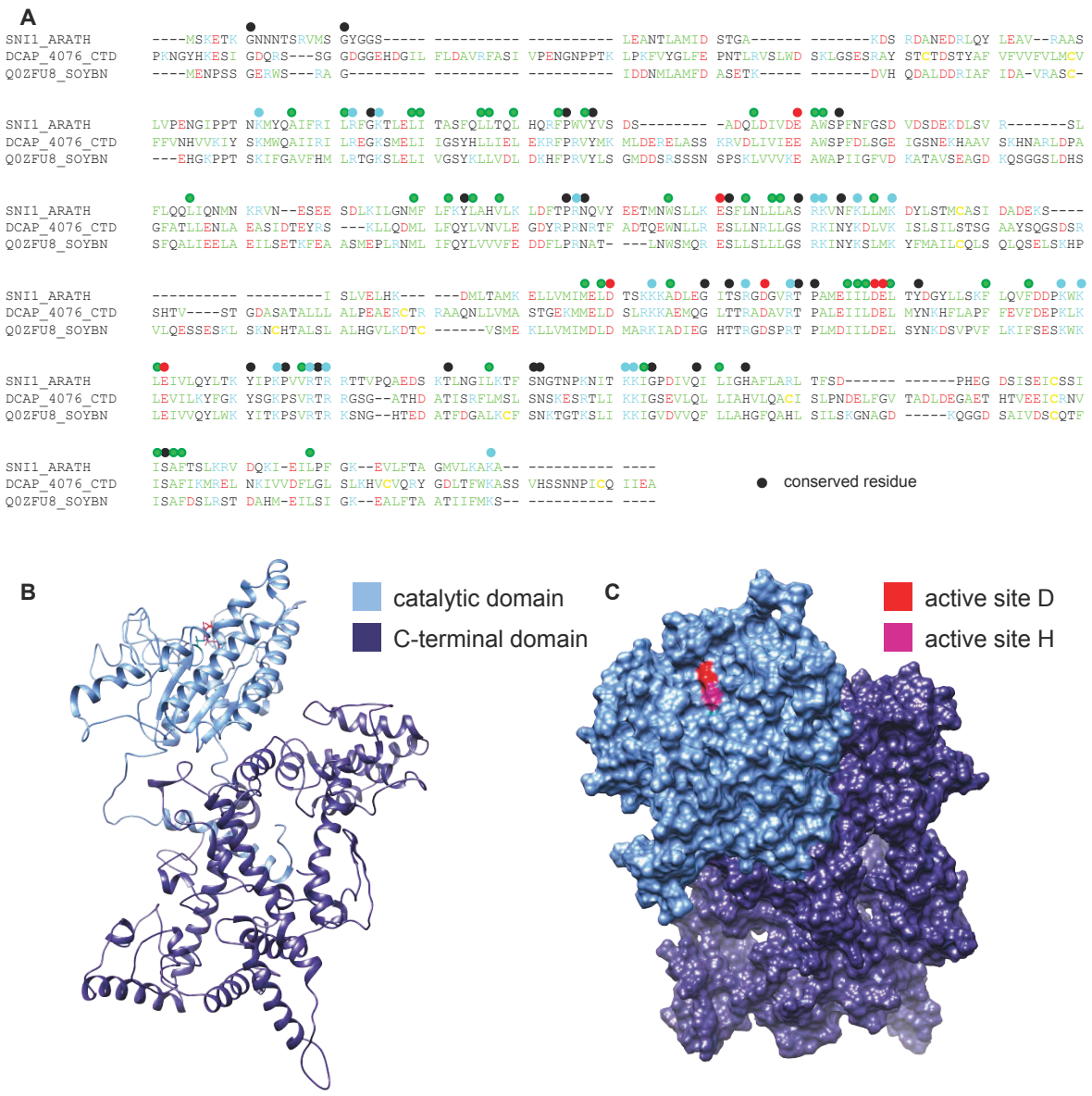

Figure S 8: A. Sequence alignment of the C-terminal domain of DCAP_4076 with the SNI1 proteins from *Arabidopsis thaliana* (Uniprot ID: SNI1_ARATH) and *Glycine max* (Uniprot ID: Q0ZFU8_SOYBN). B. Ribbon structure of DCAP_4076, with the catalytic domain in light blue and the C-terminal domain in dark blue. C. Structural model of DCAP_4076 showing the surface representation. The active site D (red) and H (magenta) residues are visible at the top of the model.

All initial and equilibrated structures available for download as PDB files are tabulated in Supplementary Tables 1 and 2, respectively.

Supplementary Table 1: Rosetta structures for esterase / lipases (PDB files available for download)

Protein	Organism	Sequence Elements included	File Name
GDL1_CARPA	*Carica papaya*	signal, active region	GDL1_CARPA_m1.pdb
DCAP_3343	*D. capensis*	signal, active region	DCAP_3343_m1.pdb
DCAP_6947	*D. capensis*	signal, active region	DCAP_6947_m1.pdb
DCAP_0448	*D. capensis*	signal, active region	DCAP_0448_m1.pdb
DCAP_8086	*D. capensis*	signal, active region	DCAP_8086_m1.pdb
DCAP_0434	*D. capensis*	active region	DCAP_0434_m1.pdb
DCAP_4098	*D. capensis*	active region	DCAP_4098_m1.pdb
DCAP_5529	*D. capensis*	signal, active region	DCAP_5529_m1.pdb
DCAP_5165	*D. capensis*	active region	DCAP_5165_m1.pdb
GLIP6_ARATH	*A. thaliana*	signal, active region	GLIP6_ARATH_m1.pdb
GDL77_ARATH	*A. thaliana*	signal, active region	GDL77_ARATH_m1.pdb
DCAP_1840	*D. capensis*	active region	DCAP_1840_m1.pdb
DCAP_1460	*D. capensis*	signal, active region	DCAP_1460_m1.pdb
DCAP_1380	*D. capensis*	active region	DCAP_1380_m1.pdb
DCAP_0405	*D. capensis*	signal, active region	DCAP_0405_m1.pdb
DCAP_4465	*D. capensis*	active region	DCAP_4465_m1.pdb
DCAP_6218	*D. capensis*	active region	DCAP_6218_m1.pdb
DCAP_6260	*D. capensis*	active region	DCAP_6260_m1.pdb
EXL3_ARATH	*A. thaliana*	signal, active region	EXL3_ARATH _m1.pdb
DCAP_1761	*D. capensis*	active region	DCAP_1761_m1.pdb
DCAP_6217	*D. capensis*	signal, active region	DCAP_6217_m1.pdb
DCAP_5461	*D. capensis*	signal, active region	DCAP_5461_m1.pdb
DCAP_0158	*D. capensis*	signal, active region	DCAP_0158_m1.pdb
DCAP_2088	*D. capensis*	active region	DCAP_2088_m1.pdb
DCAP_2089	*D. capensis*	active region	DCAP_2089_m1.pdb
DCAP_5138	*D. capensis*	active region	DCAP_5138_m1.pdb
APG2_ARATH	*A. thaliana*	signal, active region	APG2_ARATH _m1.pdb
DCAP_1365	*D. capensis*	signal, active region	DCAP_1365_m1.pdb
DCAP_5587	*D. capensis*	signal, active region	DCAP_5587_m1.pdb
DCAP_2187	*D. capensis*	active region	DCAP_2187_m1.pdb
DCAP_4076	*D. capensis*	active region	DCAP_4076_m1.pdb

Supplementary Table 2: Mature structures for esterase / lipases (PDB files available for download)

Protein	Organism	Sequence Elements included	File Name
GDL1_CARPA	*Carica papaya*	active region	GDL1_CARPA_mature_m1.pdb
DCAP_3343	*D. capensis*	active region	DCAP_3343_mature_m1.pdb
DCAP_6947	*D. capensis*	active region	DCAP_6947_mature_m1.pdb
DCAP_0448	*D. capensis*	active region	DCAP_0448_mature_m1.pdb
DCAP_8086	*D. capensis*	active region	DCAP_8086_mature_m1.pdb
DCAP_0434	*D. capensis*	active region	DCAP_0434_mature_m1.pdb
DCAP_4098	*D. capensis*	active region	DCAP_4098_mature_m1.pdb
DCAP_5529	*D. capensis*	active region	DCAP_5529_mature_m1.pdb
DCAP_5165	*D. capensis*	active region	DCAP_5165_mature_m1.pdb
GLIP6_ARATH	*A. thaliana*	active region	GLIP6_ARATH_mature_m1.pdb
GDL77_ARATH	*A. thaliana*	active region	GDL77_ARATH_mature_m1.pdb
DCAP_1840	*D. capensis*	active region	DCAP_1840_mature_m1.pdb
DCAP_1460	*D. capensis*	active region	DCAP_1460_mature_m1.pdb
DCAP_1380	*D. capensis*	active region	DCAP_1380_mature_m1.pdb
DCAP_0405	*D. capensis*	active region	DCAP_0405_mature_m1.pdb
DCAP_4465	*D. capensis*	active region	DCAP_4465_mature_m1.pdb
DCAP_6218	*D. capensis*	active region	DCAP_6218_mature_m1.pdb
DCAP_6260	*D. capensis*	active region	DCAP_6260_mature_m1.pdb
EXL3_ARATH	*A. thaliana*	active region	EXL3_ARATH _mature_m1.pdb
DCAP_1761	*D. capensis*	active region	DCAP_1761_mature_m1.pdb
DCAP_6217	*D. capensis*	active region	DCAP_6217_mature_m1.pdb
DCAP_5461	*D. capensis*	active region	DCAP_5461_mature_m1.pdb
DCAP_0158	*D. capensis*	active region	DCAP_0158_mature_m1.pdb
DCAP_2088	*D. capensis*	active region	DCAP_2088_mature_m1.pdb
DCAP_2089	*D. capensis*	active region	DCAP_2089_mature_m1.pdb
DCAP_5138	*D. capensis*	active region	DCAP_5138_mature_m1.pdb
APG2_ARATH	*A. thaliana*	active region	APG2_ARATH _mature_m1.pdb
DCAP_1365	*D. capensis*	active region	DCAP_1365_mature_m1.pdb
DCAP_5587	*D. capensis*	active region	DCAP_5587_mature_m1.pdb
DCAP_2187	*D. capensis*	active region	DCAP_2187_mature_m1.pdb
DCAP_4076	*D. capensis*	active region	DCAP_4076_mature_m1.pdb

Supplementary Table 3: Templates used for structure prediction of DCAP_0434, designated by PDBID.

PDBID	protein	organism	citation
3KVN (A)	EstA	*Pseudomonas aeruginosa*	[9]
3KVN (X)	EstA	*Pseudomonas aeruginosa*	[9]
1ESC (A)	Streptomyces scabies esterase	*Streptomyces scabiei*	[10]
3RJT (A)	lipolytic protein	*Alicyclobacillus acidocaldarius*	[11]
3MIL (A)	isoamyl acetate- hydrolyzing esterase	*Saccharomyces cerevisiae*	[12]
4JJ6 (A)	Axe2 variant H194A	*Geobacillus stearothermophilus*	[13]
4OAP (A)	Axe2 variant W190I	*Geobacillus stearothermophilus*	[14]
3W7V (A)	Axe2	*Geobacillus stearothermophilus*	[15]
4JKO (A)	Axe2 variant S15A	*Geobacillus stearothermophilus*	[16]
4HYQ (A)	phospholipase A1	*Streptomyces albidoflavus* NA297	[17]
4ZR8 (A)	uroporphyrinogen decarboxylase	*Acinetobacter baumannii*	[18]
4WSH (B)	probable uroporphyrinogen decarboxylase	*Pseudomonas aeruginosa*	[19]
4R7G (A)	Phosphoribosylformylglycinamidine synthase	*Salmonella enterica*	[20]

Active Site Network Constraint Measures

To assess the extent to which each active site was structurally constrained, four base constraint measures and one derived measure (the first principal component of these measures following standardization) were computed as described in the main text. Figure S9 shows the values of each studied protein on five constraint measures; proteins are ordered vertically by rank on the omnibus site constraint measure.

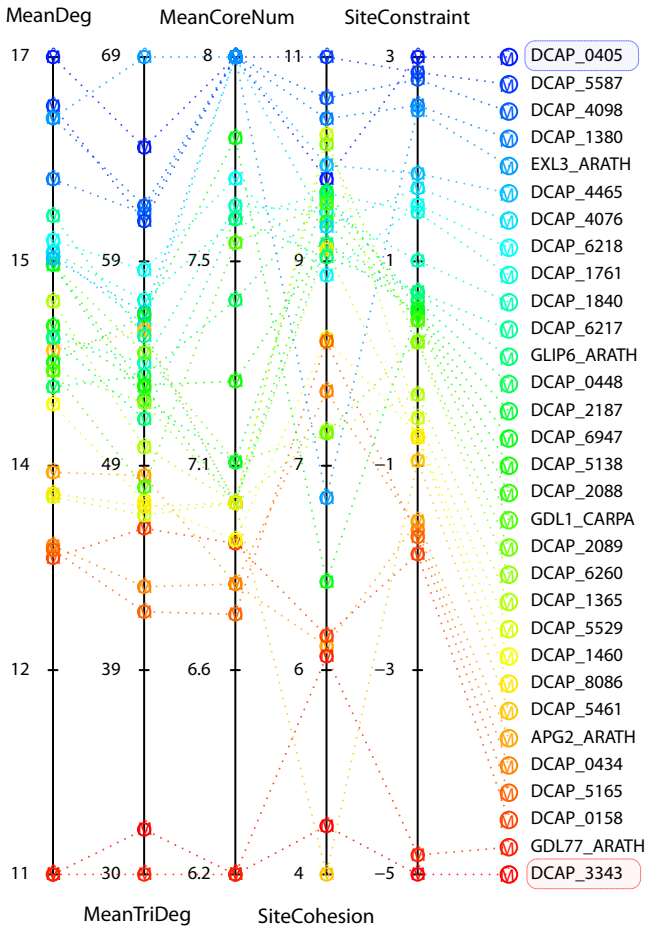

Figure S 9: Constraint measures for the active site networks. Vertical axes indicate values on each of the four base constraint measures and the omnibus derived measure, as described in the main text.

References

(1) Petersen, T.; Brunak, S.; von Heijne, G.; Henrik Nielsen, H. SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nature Methods* **2011**, *8*, 785–786.

(2) Raman, S.; Vernon, R.; Thompson, J.; Tyka, M.; Sadreyev, R.; Pei, J.; Kim, D.; Kellogg, E.; DiMaio, F.; Lange, O. et al. Structure prediction for CASP8 with all-atom refinement using Rosetta. *Proteins* **2009**, *77*, 89–99.

(3) Kim, D.; Chivian, D.; Baker, D. Protein Structure Prediction and Analysis Using the Robetta Server. *Nucleic Acids Research* **2004**, *32*, W526–31.

(4) Butts, C. T.; Zhang, X.; Kelly, J. E.; Roskamp, K. W.; Unhelkar, M. H.; Freites, J. A.; Tahir, S.; Martin, R. W. Sequence comparison, molecular modeling, and network analysis predict structural diversity in cysteine proteases from the Cape sundew, *Drosera capensis*. *Computational and Structural Biotechnology Journal* **2016**, *14*, 271–282.

(5) Phillips, J. C.; Braun, R.; Wang, W.; Gumbart, J.; Tajkhorshid, E.; Villa, E.; Chipot, C.; Skeel, R. D.; Kalé, L.; Schulten, K. Scalable molecular dynamics with NAMD. *Journal of Computational Chemistry* **2005**, *26*, 1781–1802.

(6) Pettersen, E. F.; Goddard, T. D.; Huang, C. C.; Couch, G. S.; Greenblatt, D. M.; Meng, E. C.; Ferrin, T. E. UCSF Chimera—a visualization system for exploratory research and analysis. *Journal of Computational Chemistry* **2004**, *25*, 1605–1612.

(7) Yan, S.; Wang, W.; Marqués, J.; Mohan, R.; Saleh, A.; Durrant, W. E.; Song, J.; Dong, X. Salicylic acid activates DNA damage responses to potentiate plant immunity. *Molecular Cell* **2013**, *52*, 602–610.

(8) Mosher, R. A.; Durrant, W. E.; Wang, D.; Song, J.; Dong, X. A comprehensive structure-function analysis of *Arabidopsis* SNI1 defines essential regions and transcriptional repressor activity. *The Plant Cell* **2006**, *18*, 1750–1765.

(9) van den Berg, B. Crystal structure of a full-length autotransporter. *Journal of Molecular Biology* **2010**, *396*, 627–633.

(10) Wei, Y.; Schottel, J.; Derewenda, U.; Swenson, L.; Patkar, S.; Derewenda, Z. A novel variant of the catalytic triad in the Streptomyces scabies esterase. *Nature Structural Biology* **1995**, *2*, 218–223.

(11) Chang, C.; Chhor, G.; Bearden, J.; Joachimiak, A. Crystal structure of lipolytic protein G-D-S-L family from *Alicyclobacillus acidocaldarius* subsp. acidocaldarius DSM 446. *To be published* **2011**,

(12) Ma, J.; Lu, Q.; Yuan, Y.; Ge, H.; Li, K.; Zhao, W.; Gao, Y.; Niu, L.; Teng, M. Crystal structure of isoamyl acetate-hydrolyzing esterase from *Saccharomyces cerevisiae* reveals a novel active site architecture and the basis of substrate specificity. *Proteins* **2011**, *79*, 662–668.

(13) Lansky, S.; Alalouf, O.; Solomon, V.; Alhassid, A.; Belrahli, H.; Govada, L.; Chayan, N.; Shoham, Y.; Shoham, G. Crystal structure of a catalytic mutant of Axe2 (Axe2-H194A), an acetylxylan esterase from *Geobacillus stearothermophilus*. *To be published* **2014**,

(14) Lansky, S.; Alalouf, O.; Solomon, V.; Alhassid, A.; Belrahli, H.; Govada, L.; Chayan, N.; Shoham, Y.; Shoham, G. An Axe2 mutant (W190I), an acetyl-xylooligosaccharide esterase from *Geobacillus stearothermophilus*. *To be published* **2014**,

(15) Lansky, S.; Alalouf, O.; Solomon, V.; Alhassid, A.; Belrahli, H.; Govada, L.; Chayan, N.; Shoham, Y.; Shoham, G. Crystal Structure of Axe2, an Acetylxylan Esterase from *Geobacillus stearothermophilus*. *To be published* **2014**,

(16) Lansky, S.; Alalouf, O.; Solomon, V.; Alhassid, A.; Belrahli, H.; Govada, L.; Chayan, N.; Shoham, Y.; Shoham, G. Crystal Structure of Axe2, an Acetylxylan Esterase from *Geobacillus stearothermophilus*. *Acta Crystallographica, Section D: Biological Crystallography* **2014**, *70*, 261–278.

(17) Murayama, K.; Kano, K.; Matsumoto, Y.; Sugimori, D. Crystal structure of phospholipase A1 from *Streptomyces albidoflavus* NA297. *Journal of Structural Biology* **2013**, *182*, 192–196.

(18) Abendroth, J.; Fairman, J.; Lorimer, D.; Edwards, T. Structure of uroporphyrinogen decarboxylase from Acinetobacter baumannii. *To be published* **2015**,

(19) Abendroth, J.; Dranow, D.; Lorimer, D.; Edwards, T. Crystal structure of probable uroporphyrinogen decarboxylase (UPD) (URO-D) from *Pseudomonas aeruginosa*. *To be published* **2014**,

(20) Tanwar, A.; Sindhikara, D.; Hirata, F.; Anand, R. Determination of the formylglycinamide ribonucleotide amidotransferase ammonia pathway by combining 3D-RISM theory with experiment. *ACS Chemical Biology* **2015**, *10*, 698–704.