Beilstein-Institut

Experimental Standard Conditions of Enzyme Characterizations,
September 13th – 16th, 2009, Rüdesheim/Rhein, Germany

# Functional Annotation of Orphan Enzymes within the Amidohydrolase Superfamily

**Frank M. Raushel**

Department of Chemistry, Texas A&M University,
College Station, TX 77843, U.S.A.

**E-Mail:** raushel@tamu.edu

*Received: 23rd December 2009 / Published: 14th September 2010*

## Abstract

The elucidation of the substrate profiles for enzymes of unknown function is a difficult and demanding problem. A general approach to this problem combines bioinformatics and operon context, computational docking to X-ray crystal structures, and the utilization of focused chemical libraries. These methods have been applied to the identification of novel substrates for enzymes of unknown function within the amidohydrolase superfamily. Operon context and X-ray crystallography was utilized in the identification of *N*-formimino-L-glutamate as the substrate for Pa5105 from *Pseudomonas aeruginosa* and D-galacturonate for Bh0493 from *Bacillus halodurans*. Focused substrate libraries were used to identify *N*-acetyl-D-glutamate as the substrate for Bb3285 from *Bordetella bronchiseptica* and L-Xaa-L-Arg/Lys as the substrate for Cc2672 from *Caulobacter crescentus*. Computational docking of potential high energy intermediates was used to determine that Tm0936 from *Thermotoga maritima* catalyzed the deamination of *S*-adenosyl homocysteine.

## Introduction

The recent advent of high throughput DNA sequencing efforts has significantly enhanced the number of completely sequenced bacterial genomes. The number of non-redundant genes that have been deposited in the public databases now exceeds 8 million entries. A close examination of these sequences indicates that a significant fraction of the proteins and enzymes coded by these gene sequences have an unknown, uncertain or incorrect functional annotation. This fact suggests that there are a substantial number of biochemical reactions that remain to be discovered. However, annotating enzymes of unknown function, based upon the protein sequences alone, is a difficult and demanding problem [1]. One strategy toward the solution to this problem utilizes a combination of bioinformatics, computational docking to X-ray structures or homology models, and library screening. Our efforts in this area have focused on the annotation of function for members of the amidohydrolase superfamily.

The amidohydrolase superfamily (AHS) was first identified in 1997 by Sander and Holm who recognized the structural similarities among urease, phosphotriesterase, and adenosine deaminase [2]. All three proteins fold as a distorted $(\beta/\alpha)_8$-barrel structure and possess either a binuclear or mononuclear metal centre within the active site [3]. Most of the experimentally characterized members of the AHS have been shown to catalyze the hydrolysis of amide and ester substrates contained within carbohydrates, peptides, and nucleic acids [3]. However, other members of the AHS have been shown to catalyze isomerization, hydration, and decarboxylation reactions. The metal centres in these proteins function to activate solvent water for nucleophilic attack and/or to enhance the reactivity of the substrate [4]. More than 12,000 unique protein sequences in the first 1,000 completely sequenced bacterial genomes that have been deposited in the NCBI have been identified as being part of the AHS. These sequences have been subclassified into 24 clusters of orthologous groups (COGs).

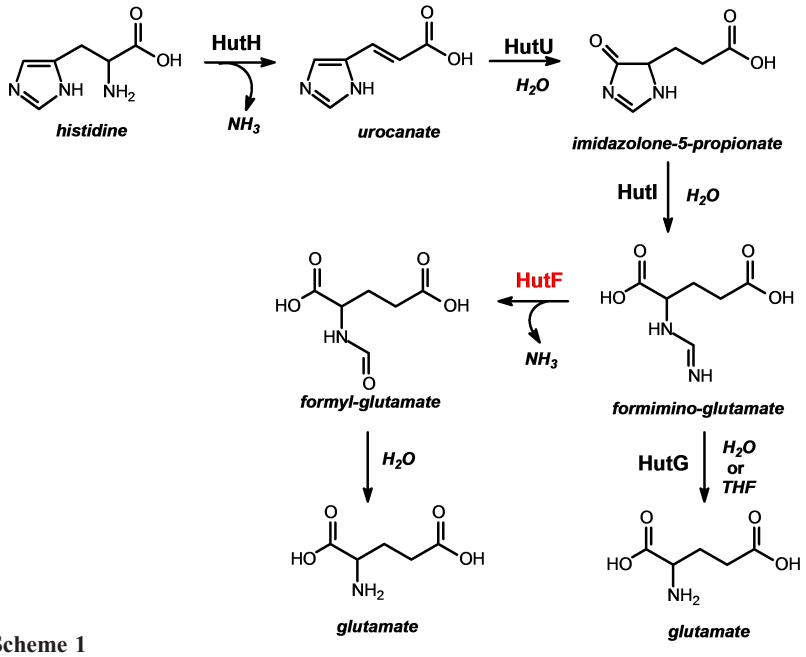
## Representative Examples

### *Pa5106: N-Formimino-L-Glutamate Deiminase*

One of the first examples for the elucidation of a previously unrecognized function for a member of the amidohydrolase superfamily occurred with Pa5106 [5]. This protein from *Pseudomonas aeruginosa* PA01 is a member of cog0402 and was misannotated as a "probable chlorohydrolase or cytosine deaminase". At the time of this investigation all of the functionally characterized members of cog0402 catalyzed the deamination of aromatic bases and these proteins included guanine deaminase and cytosine deaminase. The hallmark for this COG is an HxxE motif that is found at the end of beta-strand 5 in the $(\beta/\alpha)_8$-barrel structure. In this motif, the histidine residue coordinates to the single divalent cation in the

active site while the glutamate functions to shuttle a proton from the hydrolytic water molecule to the deaminated products (xanthine and uracil from guanine and cytosine, respectively).

Examination of the genomic context for Pa5106 within *P. aeruginosa* revealed that this gene was adjacent to a cluster of genes that are known to be involved in the degradation of histidine to glutamate. The histidine degradation pathway (hut operon) is shown in Scheme 1. In this pathway histidine is first deaminated to urocanate by histidine ammonia lyase (HutH) and then urocanase (HutU) converts this product to imidazolone-4-propionate. Imidazolone propionate amidohydrolase (HutI) catalyzes the hydrolysis of imidazolone-4-propionate to *N*-formimino-L-glutamate. In the last step the formimino group of *N*-formimino-L-glutamate is transferred to either tetrahydrofolate or to water by HutG to make the final product, L-glutamate. At first the localization of the gene for Pa5106 next to the hut operon was quite confusing (at least for us) since all four of the known proteins for the conversion of histidine to glutamate were accounted for in this pathway (HutH, HutU, HutI, and HutG) and there was no obvious need for an enzyme that we initially thought would catalyze the deamination of an aromatic base. However, it soon occurred to us that the formimino functional group of *N*-formimino-L-glutamate looked very much like that portion of guanine or cytosine that was deaminated by other members of cog0402 within the AHS. Therefore, we predicted that Pa5106 would catalyze the deimination of *N*-formimino-L-glutamate to *N*-formyl-L-glutamate (HutF) and that the protein designated as HutG would actually catalyze the hydrolysis of *N*-formyl-L-glutamate to formate and L-glutamate [5].



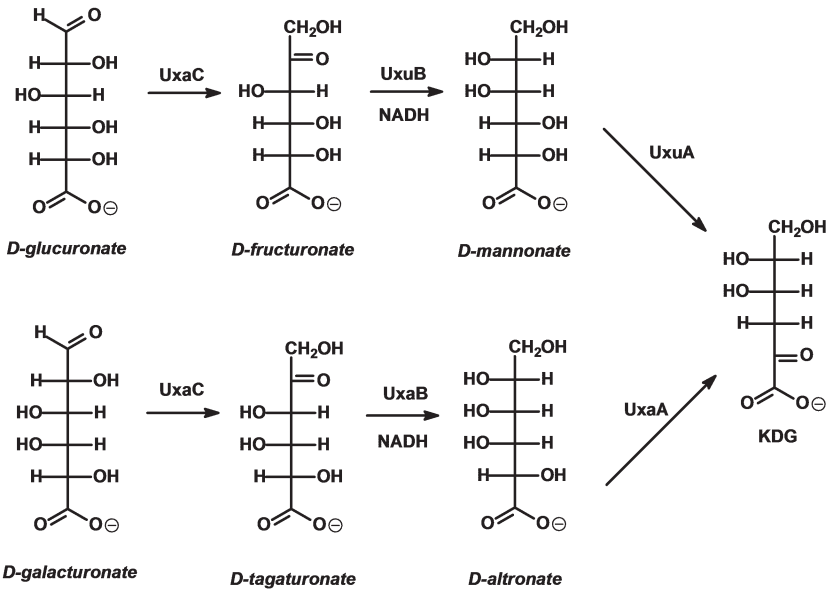

**Scheme 1**

These predictions proved to be correct. Pa5106 was found to catalyze the deimination of *N*-formimino-L-glutamate with values for $k_{cat}$, $K_m$, and $k_{cat}/K_m$ of 13 $s^{-1}$, 0.22 mM, and $6 \times 10^4$ $M^{-1}s^{-1}$, respectively. The protein originally annotated as HutG (Pa5091) was found to catalyze the hydrolysis of *N*-formyl-L-glutamate to formic acid and L-glutamate with values of $k_{cat}$, $K_m$, and $k_{cat}/K_m$ of 1 $s^{-1}$, 3.3 mM, and $3 \times 10^2$ $M^{-1}s^{-1}$, respectively. These reactions were first discovered over 50 years ago by Tabor and Mehler [6].

### *Bh0493: D-Galacturonate Isomerase*

One of the most diverged members of the AHS to be functionally characterized is Bh0493 from *Bacillus halodurans* [7]. When this protein was first interrogated the sequence identity to any other member of the amidohydrolase superfamily was less than 20% and there was significant doubt that this protein was even a member of the AHS. Nevertheless, the closest structurally characterized homologue that could be identified was uronate isomerase (Tm0064) from *Thermotoga maritima* [PDB code: 1j5 s]. Bh0493 contains a conserved WWF motif at the end of beta-strand 7 that is conserved in all of the known bacterial uronate isomerases but a conserved histidine at the end of beta-strand 5 is missing. Adding to the confusion about the functional identity of Bh0493 as a putative uronate isomerase is the presence of another protein in the genome of *B. halodurans* that is annotated as an uronate isomerase (Bh0705).

The transformations utilized by many bacteria for the metabolism of D-glucuronate and D-galacturonic are shown in Scheme 2. These two uronic acids are isomerised to D-fructuronate and D-tagaturonate, respectively, by a single enzyme, uronate isomerase. In *E. coli* D-fructuronate is subsequently converted to 2-keto-3-deoxy-D-gluconate (KDG) by the combined actions of UxuB and UxuA, whereas D-tagaturonate is transformed to KDG by a different pair of enzymes, UxaB and UxaA. Examination of the operon context for Bh0493 in *B. halodurans* proved informative since the gene for this enzyme was found to be adjacent to two genes homologous to UxaA (Bh0490) and UxaB (Bh0492) whereas the more prototypical uronate isomerase (Bh0705) was adjacent to UxuA (Bh0706) and UxuB (Bh0707). It was therefore postulated for *B. halodurans* that separate isomerases were utilized for the metabolism of D-glucuronate and D-galacturonate [7]. For Bh0705 the value of $k_{cat}/K_m$ for D-glucuronate was determined to be two orders of magnitude greater than for D-galacturonate. For Bh0493 the values of $k_{cat}/K_m$ for the two compounds are essentially the same. The operon contexts and the kinetic constants for the two enzymes capable of isomerising uronic acids in *B. halodurans* are consistent with the primary function for Bh0493 as a D-galacturonate isomerase.




**Scheme 2**

### *Bb3285: N-Acetyl-D-Glutamate Deacetylase*

The function of Bb3285 from *Bordetella bronchiseptica* was determined primarily through the utilization of a focused chemical library screen since the operon context was of little use in the assignment of function [8]. This protein is found in cog3653 and some of the enzymes in this COG have been functionally annotated as deacetylases or peptidases. Therefore, small substrate libraries containing nearly all possible combinations of L-Xaa-L-Xaa, L-Xaa-D-Xaa, D-Xaa-L-Xaa, *N*-acyl-D-Xaa, and *N*-acyl-L-Xaa were tested as substrates for Bb3285 and the products of the hydrolysis reactions quantified by amino acid analysis. Of the substrate libraries tested, the only one that showed any significant formation of a free amino acid after the addition of enzyme was *N*-acetyl-D-Xaa. This library contained the twenty common amino acids derivatized with an *N*-acetyl group. The chromatogram from the HPLC analysis is presented in Figure 1. The only substrate for Bb3258 in this library is *N*-acetyl-D-glutamate. The kinetic constants for $k_{cat}$, $K_m$, and $k_{cat}/K_m$ were found to be 460 $s^{-1}$, 88 μM, and $5 \times 10^6$ $M^{-1}s^{-1}$, respectively [8].

The identification of *N*-acetyl-D-glutamate as the primary substrate for Bb3285 enabled us to design an analogue that proved to be a potent inhibitor of this enzyme. The *N*-methyl phosphonate derivative of D-glutamate resembles the tetrahedral intermediate that would be formed during substrate hydrolysis. The structure of this compound is presented in Scheme 3. This compound is a competitive inhibitor versus *N*-acetyl-D-glutamate with a $K_i$ value of 460 pM!

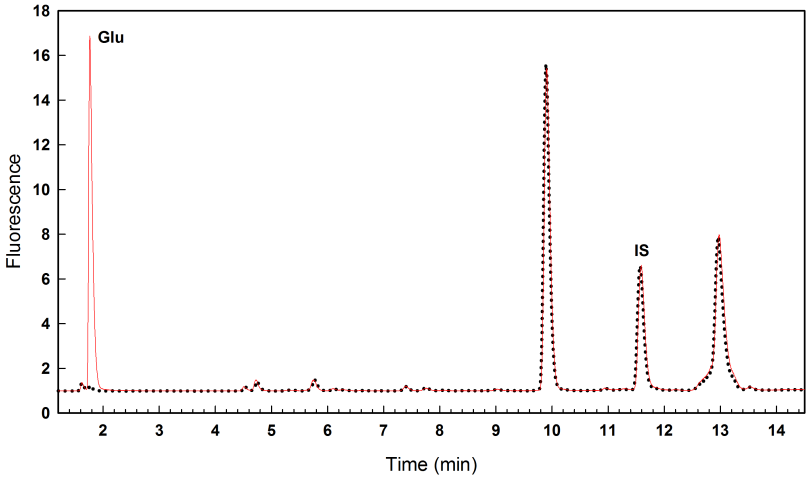
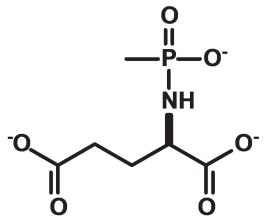



**Figure 1.** Chromatogram of the *N*-acetyl-D-Xaa library treated with no enzyme (black dots) and 20 nM Bb3285 (red line) for 1 hour at 30 °C. The OPA-derivatized D-glutamate was detected at a retention time of 1.7 minutes in the sample treated with Bb3285. The internal standard is labeled as IS.

In collaboration with the group of Steve Almo at the Einstein College of Medicine we were able to crystallize Bb3285 in the presence of this inhibitor and the molecular interactions are shown in Figure 2 (pbd code: 3giq). In this structure, the two phosphonate oxygens bridge the two metal ions in the binuclear metal centre. The α-carboxylate of the substrate is ion-paired with Lys-250, Arg-376, and Tyr-282. The recognition of the side-chain carboxylate is an ion-pair interaction with Arg-295. The structure of this complex has enabled us to identify another cluster of enzymes within cog3653 that specifically hydrolyze *N*-acyl-D-Hydrophobic amino acid derivatives. The specific examples included Gox1177 from *Gluconobacter oxydans* and Sco4986 from *Streptomyces coelicolor* [8].




**Scheme 3**

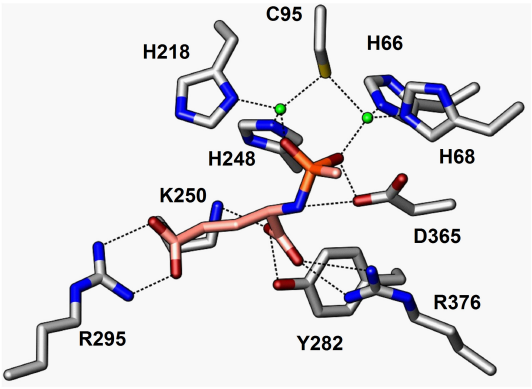



**Figure 2.** Binuclear Zn (green spheres) active site of Bb3285 with bound inhibitor (pink carbons, orange phosphorus). Enzyme-substrate contacts within 2.0 – 3.5 Å are indicated by dashed lines.

### *Cc2672: L-Xaa-L-Arg/Lys Dipeptidase*

The substrate profile for Cc2672 from *Caulobacter crescentus* CB15 was determined using a combination of library screening and X-ray crystallography [9]. At the start of this investigation Cc2672 was annotated in the NCBI as a L-Xaa-L-Pro dipeptidase. Similar to the situation with Bb3285 discussed above, the operon context for Cc2672 was of no help in the search for the reaction catalyzed by this enzyme. Therefore, the initial test of catalytic activity employed a broad set of dipeptide libraries that covered most of the nearly 1600 combinations of the twenty common amino acids in the D- and L-configurations. Of the various dipeptide libraries tested, the only ones that displayed significant catalytic turnover with Cc2672 were of the type L-Zaa-L-Xaa. In these substrate libraries a fixed amino acid was placed at the N-terminus (L-Zaa) with a combination of the 20 common amino acids at the C-terminus (L-Xaa). The liberation of free amino acids was monitored as a function of time with ninhydrin as a preliminary measure of the number of dipeptides in the library that serve as substrates. Quantitative amino acid analysis was utilized with a single dipeptide library (for example, L-Ala-L-Xaa) to determine the specific dipeptides that are hydrolyzed and the relative rates of hydrolysis for all of the dipeptides contained within a given dipeptide library. When this was conducted with Cc2672 the only free amino acids detected were L-lysine and L-arginine (and the fixed amino acid at the N-terminus). The substrate specificity for Cc2672 is therefore L-Zaa-L-Arg/Lys. There was very little discrimination among the twenty common amino acids at the N-terminus but an absolute requirement for either arginine or lysine at the C-terminus.

Thus far, attempts to determine the X-ray crystal structure of Cc2672 have failed. However, we have been able, in collaboration with the group of Steve Almo at the Einstein College of Medicine, to determine the structure of a close homologue of this enzyme. The homologue,

designated as Sgx9359b (gi|44368820), is a protein whose DNA was originally isolated from the Sargasso Sea. The crystal structure was solved to a resolution of 2.3 Å and two divalent cations were found in the active site (pdb code: 3be7 and 3dug). In one of the subunits arginine is found bound as a product in the active site and thus this structure illustrated how the terminal carboxylate of dipeptide substrates is recognized and reveals the structural determinants for the C-terminal substrate specificity. The α-carboxylate is ion paired with a histidine (His-225) that is found at the end of β-strand 5 and the guanidino group is ion paired with a glutamate (Glu-289) that is found in the loop after β-strand 7. These interactions are illustrated in Figure 3. The structural determinants for substrate specificity have helped to identify the substrate profiles for other members of this superfamily that are specific for the hydrolysis of L-Xaa-L-Hydrophobic dipeptides [10] and dipeptides that terminate in proline [11].

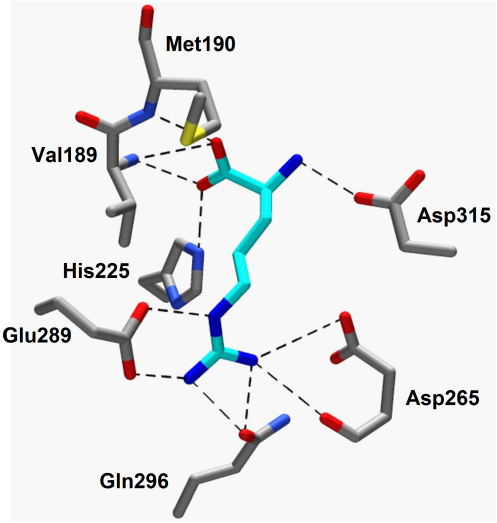
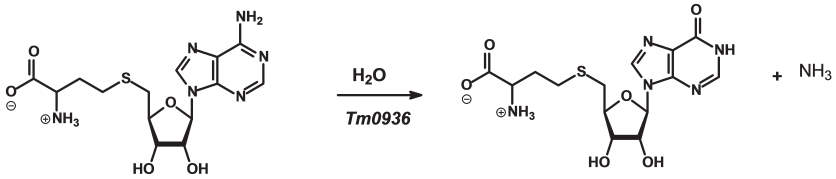



**Figure 3.** Structure of the active site of Sgx9359b showing the interactions of the product arginine with various residues. Taken from pdb code: 3dug.

### *Tm0936: S-Adenosyl Homocysteine Deaminase*

The identification of the catalytic function for Tm0936 from *Thermotoga maritima* was accomplished largely through the utilization of computational docking to an existing X-ray crystal structure [12]. These calculations were done in collaboration with the group of Brian Shoichet at UC-San Francisco. The X-ray structure of Tm0936 showed that this enzyme contained a single zinc in the active site and the conserved HxxE motif at the end of β-strand 5 placed this enzyme within cog0402 (pdb code: 1plm and 1j6 p). This enzyme is in the same COG as Pa5106, cytosine deaminase and guanine deaminase, and thus it was highly likely that the overall reaction would involve a deamination reaction. For the docking calculations, the Shoichet laboratory created small molecule mimics that resembled the

putative transition state intermediates for the hydrolysis reactions. The strategy here was based on the assumption that molecules that resembled the transition states would be more selective for the active site than simple ground state molecules [13]. The obvious complication for these types of docking calculations is the high potential for conformational changes in the protein structure upon binding of the substrate to the active site. Nevertheless, the entire KEGG library of compounds that possess a hydrolytic site was computationally docked into the active site of Tm0936. Of the top 100 hits, nearly 40% of the compounds were modifications of adenosine. This result provided a high degree of confidence that an adenosine derivative would be deaminated by Tm0936. Of the compounds tested, the best substrate was *S*-adenosyl homocysteine (SAH) followed by thiomethyl adenosine (TMA) and adenosine itself. The values of $k_{cat}/K_m$ for the three compounds were greater than $10^5$ $M^{-1}s^{-1}$. As a test of the correctness of the proposed docking pose relative to the binding of actual ligands to the active site we enzymatically prepared the product of the reaction by incubating SAH with Tm0936 and then isolated the product, *S*-inosyl homocysteine (SIH). The SIH was then used in a co-crystallization of Tm0936. The structural overlay between the docking pose of the proposed high energy intermediate and the conformation of SIH bound in the active site of Tm0936, determined by X-ray diffraction methods, was excellent (pdb code: 2plm). The conversion of SAH to SIH had not previously been recognized as a metabolic transformation and it is shown in Scheme 4.




**Scheme 4**

## Conclusions

A significant fraction of the genes contained within recently sequenced bacterial genomes have an unknown function. We have attempted to develop a broad-based strategy for determining the function of enzymes belonging to the amidohydrolase superfamily. Employment of bioinformatics and operon context, computational docking of intermediates and substrates to active sites, and screening with focused chemical libraries have enabled the identification of novel substrates for a variety of enzymes of unknown function.

## Acknowledgment

The work described in this paper was supported in part by the National Institutes of Health (GM 71790) and the Robert A. Welch Foundation (A-840).

## References

[1] Gerlt, J.A. and Babbitt, P.C. (2000) Can sequence determine function? *Genome Biology* **5**:1 – 10.
doi: http://dx.doi.org/10.1186/gb-2000-1-5-reviews0005.

[2] Holm, L. and Sander, C. (1997) An evolutionary treasure: unification of a broad set of amidohydrolases related to urease. *Proteins* **28**:72 – 82.
doi: http://dx.doi.org/10.1002/(SICI)1097-0134(199705)28:1<72::AID-PROT7>3.0.CO;2-L.

[3] Seibert, C.M. and Raushel, F.M. (2005) Structural and Catalytic Diversity within the Amidohydrolase Superfamily. *Biochemistry* **44**:6383 – 6391.
doi: http://dx.doi.org/10.1021/bi047326v.

[4] Aubert, S.D., Li, Y., and Raushel, F.M. (2004) Mechanism for the Hydrolysis of Organophosphates by the Bacterial Phosphotriesterase. *Biochemistry* **43**:5707 – 5715.
doi: http://dx.doi.org/10.1021/bi0497805.

[5] Marti-Arbona, R., Xu, C., Steele, S., Weeks, A., Kuty, G.F., Seibert, C.M. and Raushel, F.M. (2006) Annotating Enzymes of Unknown Function: *N*-Formimino-L-glutamate Deiminase Is a Member of the Amidohydrolase Superfamily. *Biochemistry* **45**:1997 – 2005.
doi: http://dx.doi.org/10.1021/bi0525425.

[6] Tabor, H. and Mehler, A.H. (1954) Isolation of *N*-Formyl-L-Glutamic Acids as an Intermediate in the Enzymatic Degradation of L-Histidine. *J. Biol. Chem*. **210**(2):559 – 568.

[7] Nguyen, T.T., Brown, S., Fedorov, A.A., Fedorov, E.V., Babbitt, P.C., Almo, S.C. and Raushel, F.M. (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.
doi: http://dx.doi.org/10.1021/bi7017738.

[8] Cummings, J., Fedorov, A.A., Xu, C., Brown, S., Fedorov, E.V., Babbitt, P.C., Almo, S.C. and Raushel, F.M. (2009) Annotating Enzymes of Uncertain Function: The Deacylation of D-Amino Acids by Members of the Amidohydrolase Superfamily. *Biochemistry* **48**:6469 – 6481.
doi: http://dx.doi.org/10.1021/bi900661b.

[9] Xiang, D.F., Patskovsky, Y., Xu, C., Meyer, A., Sauder, J.M., Burley, S.K., Almo, S.C., and Raushel, F.M. (2009) Functional Identification of Incorrectly Annotated Prolidases from the Amidohydrolase Superfamily of Enzymes. *Biochemistry* **48**:3730 – 3742.
doi: http://dx.doi.org/10.1021/bi900111q.

[10] Xiang, D.F., Xu, C., Kumaran, D., Brown, A.C., Sauder, J.M., Burley, S.K., Swaminathan, S., Raushel, F.M. (2009) Functional Annotation of Two New Carboxypeptidases from the Amidohydrolase Superfamily of Enzymes. *Biochemistry* **48**:4567 – 4576.
doi: http://dx.doi.org/10.1021/bi900453u.

[11] Xiang, D.F. and Raushel, F.M. (2009) unpublished observations.

[12] Hermann, J., Marti-Arbona, R., Fedorov, A.A., Fedorov, E., Almo, S.C., Shoichet, B.K. and Raushel, F.M. (2007) Structure-based activity prediction for an enzyme of unknown function. *Nature* **448**: 775 – 779.
doi: http://dx.doi.org/10.1038/nature05981.

[13] Hermann, J.C., Ghanem, E., Li, Y., Raushel, F.M., Irwin, J.J. and Shoichet, B.K. (2006) Predicting Substrates by Docking High-Energy Intermediates to Enzyme Structures. *J. Amer. Chem. Soc.* **128**:15882 – 15891.
doi: http://dx.doi.org/10.1021/ja065860f.