Discovery of new enzymes and metabolic pathways by using structure and genome context

Suwen Zhao1*, Ritesh Kumar2*, Ayano Sakai2*, Matthew W. Vetting3*, B. McKay Wood2*, Shoshana Brown4, Jeffery B. Bonanno3, Brandan S. Hillerich3, Ronald D. Seidel1, Patricia C. Babbitt4, Steven C. Almo3, Jonathan V. Sweedler2,5, John A. Gerlt2,5,6, John E. Cronan2,6,7 & Matthew P. Jacobson1

Assigning valid functions to proteins identified in genome projects is challenging: overprediction and database annotation errors are the principal concerns1. We and others2 are developing computation-guided strategies for functional discovery with ‘metabolite docking’ to experimentally derived3 or homology-based4 three-dimensional structures. Bacterial metabolic pathways often are encoded by ‘genome neighbourhoods’ (gene clusters and/or operons), which can provide important clues for functional assignment. We recently demonstrated the synergy of docking and pathway context by ‘predicting’ the intermediates in the glycolytic pathway in Escherichia coli5. Metabolite docking to multiple binding proteins and enzymes in the same pathway increases the reliability of in silico predictions of substrate specificities because the pathway intermediates are structurally similar. Here we report that structure-guided approaches for predicting the substrate specificities of several enzymes encoded by a bacterial gene cluster allowed the correct prediction of the in vitro activity of a structurally characterized enzyme of unknown function (PDB 2PMQ), 2-epimerization of trans-4-hydroxy-L-proline betaine (tHyp-B) and cis-4-hydroxy-D-proline betaine (cHyp-B), and also the correct identification of the catabolic pathway in which Hyp-B 2-epimerase participates. The substrate-ligated pose predicted by virtual library screening (docking) was confirmed experimentally. The enzymatic activities in the predicted pathway were confirmed by in vitro assays and genetic analyses; the intermediates were identified by metabolomics; and repression of the genes encoding the pathway by high and genetic analyses; the intermediates were identified by metabolomics. The automated TriEMBL annotations (Supplementary Table 1) fail to assign the in vitro activity of HpbB or to identify the metabolic pathway. P. bermudensis is not genetically tractable; we therefore studied P. denitrificans, which encodes one HpbD orthologue and two sets of orthologues of most of the genes neighbouring the P. bermudensis hpbDgene. Genome neighbourhoods are ‘conserved’ for other putative orthologues (about 20 can be identified in the sequence databases at http://slld.rbv.ucsf.edu/).

The in silico ligand docking library (87,098 members) included the KEGG metabolite library7 as well as other potential enolase superfamily substrates such as dipeptides, N-acetylated amino acids, acid sugars and the enolate anions obtained by abstraction of the α-proton (high-energy intermediates8) (Methods). The library was docked in the active site of HpbD by using Glide SP, and energy scoring functions rank-ordered the members of the library according to binding affinity. The best-scoring molecules were enriched with amino acid derivatives, especially proline analogues and N-capped amino acid derivatives (Fig. 3), permitting the prediction that HpbD is an amino acid racemase/epimerase, with the substrate probably having N-substitution.

The genome neighbourhood includes an ABC transporter with a periplasmic binding protein (HpbJ) annotated as binding ‘glycine betaine/α-proline’. The structure of a homologous binding protein with glycine betaine9 (PDB 1R9L) was used as the homology model template (HpbJ and 1R9L share 48% sequence identity) (Methods). The binding site contains three tryptophan residues (Fig. 1b) that form a π-cation ‘cage’ for a quaternary ammonium (betaine), which may also be electrostatically stabilized by Glu 42, located 5.4 Å from the quaternary nitrogen. Thus, the homology model permitted the prediction that its ligand is a betaine. A library of 31 betaines was docked to the model; tHyp-B had the highest rank (Supplementary Table 2), so we predicted that HpbJ participates in the transport of tHyp-B. In addition, the HpbD active site contains Trp 320 and Asp 292, which are similarly positioned relative to the predicted binding pose of betaines (Fig. 1c). The structural basis for the predicted specificity of HpbJ therefore refined the prediction that the substrate for HpbD is a proline betaine, for example tHyp-B.

A homology model was constructed for the Rieske-type protein (HpbB1) using a homologous (PDB 3N0Q) as the template (60% sequence identity; Fig. 1d) (Methods). The active site resembled the binding sites in the betaine-binding proteins (aromatic residues and Glu 200); indeed, some Rieske-type proteins are betaine demethylases10–12 (Supplementary Fig. 2 and Supplementary Table 3). We therefore predicted that the active site is sequestered from solvent by two closed loops and was therefore suitable for virtual metabolite docking for substrate prediction (Supplementary Fig. 1).

Figure 2 shows the genome neighbourhoods of the gene encoding 2PMQ (hpbD; Hpb-B 2-epimerase from P. bermudensis) (Methods). The automated TriEMBL annotations (Supplementary Table 1) fail to assign the in vitro activity of HpbB or to identify the metabolic pathway. P. bermudensis is not genetically tractable; we therefore studied P. denitrificans, which encodes one HpbD orthologue and two sets of orthologues of most of the genes neighbouring the P. bermudensis hpbDgene. Genome neighbourhoods are ‘conserved’ for other putative orthologues (about 20 can be identified in the sequence databases at http://slld.rbv.ucsf.edu/).

The marine bacterium Pelagibaca bermudensis encodes an uncharacterized member of the enolase superfamily (National Center for Biotechnology Information GI number 114543141) in which two lysine residues of the TIM-barrel domain are positioned to function as acid-base catalysts13–15. The New York SGX Research Consortium determined its structure (PDB 2PMQ) because it shared less than 30% sequence identity with structurally characterized enolase superfamily members. The only ligand was the Mg2+ that stabilizes the enolate anion intermediate obtained by abstraction of the α-proton of a carbonyl substrate. 

*These authors contributed equally to this work. 

1Department of Pharmaceutical Chemistry, University of California, San Francisco, California 94143, USA. 2Institute for Genomic Biology, University of Illinois at Urbana-Champaign, Urbana, Illinois 61801, USA. 3Department of Biochemistry, Albert Einstein College of Medicine, Bronx, New York 10461, USA. 4Department of Bioengineering and Therapeutic Sciences, University of California, San Francisco, California 94143, USA. 5Department of Chemistry, University of Illinois at Urbana-Champaign, Urbana, Illinois 61801, USA. 6Department of Biochemistry, University of Illinois at Urbana-Champaign, Urbana, Illinois 61801, USA. 7Department of Microbiology, University of Illinois at Urbana-Champaign, Urbana, Illinois 61801, USA.

©2013 Macmillan Publishers Limited. All rights reserved
substrate is a small betaine. (While this work was in progress, the X-ray structure of a Rieske-type Pro-B demethylase from *Sinorhizobium meliloti* (PDB 3VCP) was published\(^1\); its active site superimposed closely with our HpbB1 homology model.)

The results of library docking to the experimental apo structure of HpbD (2PMQ) and to homology models of Hpbj and HpbB1 enabled us to predict that HpbD uses Hyp-B or Pro-B as a substrate in a 1,1-proton transfer reaction. The betaines Gly-B and carnitine also were candidates (although these would be substrates for virtual reactions; the \(\alpha\)-carbons are prochiral). Although more than 25 functions have been assigned to members of the enolase superfamily, including \(N\)-succinylamino acid racemases and dipeptide epimerases, no amino acid or amino acid betaine was known to be a substrate\(^6\).

The kinetic constants for tHyp-B and \(l\)-Pro-B were determined for both HpbD orthologues (Fig. 4b) (Methods). Although the \(k_{cat}\) values are large, the \(K_m\) values are also large, so the \(k_{cat}/K_m\) values are modest. Betaines, including tHyp-B, are osmoprotectants accumulated by many bacteria, including pelagic (*P. bermudensis*) and plant-associated (*P. denitrificans*) species, to survive osmotic stress\(^{16-20}\); their intracellular concentrations can approach molar levels\(^{21,22}\). We determined that the intracellular concentration of Hyp-B is 170 mM in *P. denitrificans* grown on glucose in the presence of 0.5 M NaCl and 20 mM tHyp-B (Methods). Hyp-B 2-epimerase probably functions with a high intracellular concentration of tHyp-B, so the kinetic constants are both physiologically reasonable and expected\(^23\). That only four compounds were tested (tHyp-B, \(l\)-Pro-B, carnitine and Gly-B) and two have physiologically relevant kinetic constants confirms that pathway docking enables efficient functional prediction.

The 1.70-Å structure of HpbD was determined in the presence of D\(_2\)O (Methods). The \(\mathrm{H}^1\) NMR spectra with tHyp-B, D-Pro-B and Gly-B revealed exchange of the \(\alpha\)-proton with solvent deuterium (the latter being a virtual reaction); in addition, for tHyp-B, resonances associated with \(c\)Hyp-B, the 2-epimer, were observed (Supplementary Fig. 3). These results are expected for a 1,1-proton transfer reaction that equilibrates the configurations at carbon 2 of tHyp-B and Pro-B using two lysine acid–base catalysts.

The kinetic constants for tHyp-B and \(l\)-Pro-B were determined for both HpbD orthologues (Fig. 4b) (Methods). Although the \(k_{cat}\) values are large, the \(K_m\) values are also large, so the \(k_{cat}/K_m\) values are modest. Betaines, including tHyp-B, are osmoprotectants accumulated by many bacteria, including pelagic (*P. bermudensis*) and plant-associated (*P. denitrificans*) species, to survive osmotic stress\(^{16-20}\); their intracellular concentrations can approach molar levels\(^{21,22}\). We determined that the intracellular concentration of Hyp-B is 170 mM in *P. denitrificans* grown on glucose in the presence of 0.5 M NaCl and 20 mM tHyp-B (Methods). Hyp-B 2-epimerase probably functions with a high intracellular concentration of tHyp-B, so the kinetic constants are both physiologically reasonable and expected\(^23\). That only four compounds were tested (tHyp-B, \(l\)-Pro-B, carnitine and Gly-B) and two have physiologically relevant kinetic constants confirms that pathway docking enables efficient functional prediction.

The 1.70-Å structure of HpbD was determined in the presence of tHyp-B (Methods, Supplementary Fig. 4 and Supplementary Table 4).

![Figure 1](https://example.com/figure1.png)  
**Figure 1** | Homology modelling and docking results for HpbD, Hpbj and HpbB1. *a* The reaction catalysed by HpbD, the Hyp-B 2-epimerase. *b* The binding site of the model of Hpbj, with the top-ranked ligand tHyp-B docked. The ligand surface is shown in magenta. *c*, Comparison of HpbD top-ranked docking pose of t-Pro-B (magenta) with the experimental pose of tHyp-B (cyan). The unliganded structure used in docking (PDB 2PMQ) and the subsequently determined liganded structure (PDB 4H2H) are shown in magenta and cyan, respectively. *d*, Superposition of the model of HpbB1 (magenta) and the closest characterized Rieske-type protein (cyan; PDB 1O7G, a naphthalene dioxygenase), showing that the active site of the model is too small to accept naphthalene as a substrate. Steric clashes identified by using a van der Waals overlap of 0.6 Å or more are shown in red lines.

![Figure 2](https://example.com/figure2.png)  
**Figure 2** | Genome contexts of HpbD in *P. bermudensis* and the orthologous genes in *P. denitrificans*. The genes encoding orthologues are highlighted with the same colour; the sequence identities relating orthologues in *P. bermudensis* and *P. denitrificans* are indicated. The ecological sources of tHyp-B would be seaweed (sargasso) for the Sargasso Sea bacterium *P. bermudensis*, and plants for the soil bacterium *P. denitrificans*. 

©2013 Macmillan Publishers Limited. All rights reserved
Figure 3 | Chemotype analysis of HpbD docking results. a, Enriched chemotypes in the top 120 hits. Most of them are amino acid derivatives, in which N-capped amino acid derivatives and proline analogues are the two most common subtypes. b, Proline analogues in the rank-ordered list of predicted ligands, illustrating the frequent occurrence of N-modified proline analogues. Pro-B, a substrate for HpbD, ranks at number 110 in the list (top 0.12% of the docking library).

The ligand electron density, with elevated $B$-factors, was interpreted as a mixture of tHyp-B (substrate) and cHyp-B (product) (Supplementary Fig. 5). The betaine forms a $\pi$-cation interaction with Trp 320 and is proximal to Asp 292, similar to the interactions in the Gly-B periplasmic binding protein (Fig. 1c). The predicted pose with D-Pro-B superimposes closely on the experimental pose, explaining the correct computation-based prediction of substrate specificity.

We also speculated that the binding proteins and enzymes encoded in the HpbD genome neighbourhoods constitute a catabolic pathway that degrades tHyp-B to $\alpha$-ketoglutarate (Fig. 4a), with HpbD catalysing the first step in which tHyp-B is 2-epimerized to cHyp-B; that is, the in vivo activity of HpbD is Hyp-B 2-epimerase. Subsequently, HpbB1/HpbC1, the Rieske-type protein, catalyses the demethylation of cHyp-B to N-methyl cHyp; HpbA, a flavin-dependent enzyme, converts N-methyl cHyp to cHyp; HpbG, a d-amino acid oxidase, catalyses the oxidation of cHyp to its imino acid; HpbB, a member of the dihydropicolinate synthase superfamily, catalyses the dehydrogenation of the 4-OH group and ‘hydrolysis’ of the 5-amino group to $\alpha$-ketoglutarate semialdehyde; and HpbE, an aldehyde dehydrogenase, catalyses the oxidation of $\alpha$-ketoglutarate semialdehyde to $\alpha$-ketoglutarate. This pathway would permit the utilization of tHyp-B as a carbon and nitrogen source. The activities predicted for HpbE and HpbG were described recently in pathways for tHyp catabolism in Pseudomonas aeruginosa, Pseudomonas putida and S. meliloti, however, the sequences of HpbE and HpbB are so divergent (less than 35% sequence identity) that the cHyp oxidase and cHyp imino acid dehydratase/deaminase functions could not be assigned to the P. bermudensis and P. denitrificans enzymes without additional information (Supplementary Fig. 6).

When tHyp-B is used as an osmoprotectant (sea water is about 0.6 M NaCl), its catabolism should be depressed to maintain high intracellular concentrations. However, in the absence of osmotic stress, bacteria should be able to catabolize tHyp-B as a carbon and nitrogen source. P. denitrificans utilizes both tHyp-B and cHyp-B as carbon and nitrogen sources at low salt concentrations, as expected if the genome encodes the proposed catabolic pathway (Methods). Moreover, tHyp-B alleviates growth inhibition at high salt concentration (0.5 M NaCl) in glucose medium, arguing that P. denitrificans uses tHyp-B as an osmoprotectant (Supplementary Figs 7–9). Growth stimulation by tHyp-B in high-salt glucose medium could result from both osmoprotection and catabolism of tHyp-B. To address this possibility we used strain RPd4, which lacks the demethylases that convert the isomers of Hyp-B to the isomers of N-methyl Hyp and therefore cannot utilize tHyp-B or cHyp-B as a carbon source (Supplementary Table 8). Strain RPd4 grew almost as well on high-salt glucose medium in the presence of tHyp-B or cHyp-B as did the culture without salt supplementation, but growth on high-salt glucose medium in the absence of tHyp-B or cHyp-B was strongly inhibited (Supplementary Fig. 8). The results establish that tHyp-B and cHyp-B function as osmoprotectants.

We identified the metabolites obtained from tHyp-B at low salt concentrations (Methods). In addition to Hyp-B (21 mM; Supplementary Table 10), N-methyl Hyp and Hyp (the carbon-2 epimers cannot be distinguished), the predicted downstream $\Delta^1$-pyrroline-4-hydroxy-2-carboxylate, $\alpha$-KGSA, $\alpha$-ketoglutarate semialdehyde and $\alpha$-ketoglutarate were observed (Supplementary Figs 10 and 11). The metabolites were not detected with succinate as a carbon source. In high-salt glucose medium
containing tHyp-B, the intracellular concentration of Hyp-B was 170 mM (as expected for an osmolyte; Supplementary Table 10); however, its downstream metabolites were not detected. Thus, the flux through the pathway is regulated so that tHyp-B is not catabolized when it is needed as osmoprotectant. No Hyp-B was detected in cells grown on high-salt glucose medium, establishing that P. denitrificans lacks an anabolic pathway for tHyp-B.

We used quantitative PCR with reverse transcription (qRT–PCR) to investigate expression of the genes encoding the catabolic pathway (Methods and Supplementary Table 6). P. denitrificans encodes one orthologue of Hyp-B 2-epimerase (HpbD) and the FAD-dependent N-methyl Hyp demethylase but two orthologues of the remaining proteins and enzymes involved in the transport of tHyp-B and its catabolism (Fig. 2). The genes encoding the pathway are upregulated by tHyp-B and cHyp-B, as expected if their encoded products are involved in the catabolic pathway.

The effects of high salt concentration were determined using equimolar concentrations of glucose and either tHyp-B or cHyp-B. Salt (0.5 M NaCl) enhanced the expression of the transporters (HpbN/HpbO/HpbH/Hpbj and HpbX/HpbY/HpbZ). In contrast, salt decreased the expression of the genes encoding Hyp-B 2-epimerase (HpbD), both Hyp-B demethylases (HpbB1/HpbC1 and HpbB2/HpbC2) and the single N-methyl Hyp demethylase (HpbA) (Supplementary Table 6). Transport of HpbN/HpbO/HpbH/Hpbj is required for uptake as osmolytes as well as carbon and nitrogen sources; expression of their transporters is enhanced, whereas epimerization and demethylation are suppressed, thereby allowing tHyp-B/cHyp-B to be retained as osmolytes.

The genes encoding the P. denitrificans pathway were individually disrupted by the insertion of antibiotic-resistance cassettes (Methods and Supplementary Table 7). The growth phenotypes are consistent with the predicted functions (Supplementary Discussion).

Here we have used homology modelling and metabolite docking to several proteins encoded by a gene cluster to guide the in vitro assignment of the previously undocumented Hyp-B 2-epimerase activity to 2PMQ, a structure determined by the Protein Structure Initiative. With knowledge of the catalytic capabilities of enzyme superfamilies, we also predicted the pathway that catabolizes Hyp-B 2-0-ketogluutarate. These predictions were verified by metabolomics and genetics. Finally, we used transcriptomics to demonstrate that Hyp-B 2-epimerase is a ‘switch’ that determines whether the tHyp-B is accumulated as an osmolyte or catabolized as carbon and nitrogen source.

Orthologues of HpbD can be identified in 20 microbial species (http://sld.rvb.uncfs.edu/), so both the in vitro activity and the in vivo functional assignments identified in this study can be extended to these proteins and organisms. Moreover, we expect that the Hyp-B 2-epimerase activity assigned to HpbD will be used to facilitate the discovery of the proteins and organisms. Moreover, we expect that the Hyp-B 2-epimerase activity assigned to HpbD will be used to facilitate the discovery of the proteins and organisms.

We propose pathway docking as an efficient strategy for predicting fundamental determinant in the evolution of new catalytic activities. Sequence-based assignment and the FAD-dependent N-methyl Hyp demethylase (HpbA) (Supplementary Table 6). Transport of HpbN/HpbO/HpbH/Hpbj is required for uptake as osmolytes as well as carbon and nitrogen sources; expression of their transporters is enhanced, whereas epimerization and demethylation are suppressed, thereby allowing tHyp-B/cHyp-B to be retained as osmolytes.

The genes encoding the P. denitrificans pathway were individually disrupted by the insertion of antibiotic-resistance cassettes (Methods and Supplementary Table 7). The growth phenotypes are consistent with the predicted functions (Supplementary Discussion).

Here we have used homology modelling and metabolite docking to several proteins encoded by a gene cluster to guide the in vitro assignment of the previously undocumented Hyp-B 2-epimerase activity to 2PMQ, a structure determined by the Protein Structure Initiative. With knowledge of the catalytic capabilities of enzyme superfamilies, we also predicted the pathway that catabolizes Hyp-B 2-0-ketogluutarate. These predictions were verified by metabolomics and genetics. Finally, we used transcriptomics to demonstrate that Hyp-B 2-epimerase is a ‘switch’ that determines whether the tHyp-B is accumulated as an osmolyte or catabolized as carbon and nitrogen source.

Orthologues of HpbD can be identified in 20 microbial species (http://sld.rvb.uncfs.edu/), so both the in vitro activity and the in vivo functional assignments identified in this study can be extended to these proteins and organisms. Moreover, we expect that the Hyp-B 2-epimerase activity assigned to HpbD will be used to facilitate the discovery of the proteins and organisms.

We propose pathway docking as an efficient strategy for predicting fundamental determinant in the evolution of new catalytic activities. Sequence-based assignment and the FAD-dependent N-methyl Hyp demethylase (HpbA) (Supplementary Table 6). Transport of HpbN/HpbO/HpbH/Hpbj is required for uptake as osmolytes as well as carbon and nitrogen sources; expression of their transporters is enhanced, whereas epimerization and demethylation are suppressed, thereby allowing tHyp-B/cHyp-B to be retained as osmolytes.

The genes encoding the P. denitrificans pathway were individually disrupted by the insertion of antibiotic-resistance cassettes (Methods and Supplementary Table 7). The growth phenotypes are consistent with the predicted functions (Supplementary Discussion).

Here we have used homology modelling and metabolite docking to several proteins encoded by a gene cluster to guide the in vitro assignment of the previously undocumented Hyp-B 2-epimerase activity to 2PMQ, a structure determined by the Protein Structure Initiative. With knowledge of the catalytic capabilities of enzyme superfamilies, we also predicted the pathway that catabolizes Hyp-B 2-0-ketogluutarate. These predictions were verified by metabolomics and genetics. Finally, we used transcriptomics to demonstrate that Hyp-B 2-epimerase is a ‘switch’ that determines whether the tHyp-B is accumulated as an osmolyte or catabolized as carbon and nitrogen source.

Orthologues of HpbD can be identified in 20 microbial species (http://sld.rvb.uncfs.edu/), so both the in vitro activity and the in vivo functional assignments identified in this study can be extended to these proteins and organisms. Moreover, we expect that the Hyp-B 2-epimerase activity assigned to HpbD will be used to facilitate the discovery of the proteins and organisms.

We propose pathway docking as an efficient strategy for predicting fundamental determinant in the evolution of new catalytic activities. Sequence-based assignment and the FAD-dependent N-methyl Hyp demethylase (HpbA) (Supplementary Table 6). Transport of HpbN/HpbO/HpbH/Hpbj is required for uptake as osmolytes as well as carbon and nitrogen sources; expression of their transporters is enhanced, whereas epimerization and demethylation are suppressed, thereby allowing tHyp-B/cHyp-B to be retained as osmolytes.

The genes encoding the P. denitrificans pathway were individually disrupted by the insertion of antibiotic-resistance cassettes (Methods and Supplementary Table 7). The growth phenotypes are consistent with the predicted functions (Supplementary Discussion).

Here we have used homology modelling and metabolite docking to several proteins encoded by a gene cluster to guide the in vitro assignment of the previously undocumented Hyp-B 2-epimerase activity to 2PMQ, a structure determined by the Protein Structure Initiative. With knowledge of the catalytic capabilities of enzyme superfamilies, we also predicted the pathway that catabolizes Hyp-B 2-0-ketogluutarate. These predictions were verified by metabolomics and genetics. Finally, we used transcriptomics to demonstrate that Hyp-B 2-epimerase is a ‘switch’ that determines whether the tHyp-B is accumulated as an osmolyte or catabolized as carbon and nitrogen source.

Orthologues of HpbD can be identified in 20 microbial species (http://sld.rvb.uncfs.edu/), so both the in vitro activity and the in vivo functional assignments identified in this study can be extended to these proteins and organisms. Moreover, we expect that the Hyp-B 2-epimerase activity assigned to HpbD will be used to facilitate the discovery of the proteins and organisms.

We propose pathway docking as an efficient strategy for predicting fundamental determinant in the evolution of new catalytic activities. Sequence-based assignment and the FAD-dependent N-methyl Hyp demethylase (HpbA) (Supplementary Table 6). Transport of HpbN/HpbO/HpbH/Hpbj is required for uptake as osmolytes as well as carbon and nitrogen sources; expression of their transporters is enhanced, whereas epimerization and demethylation are suppressed, thereby allowing tHyp-B/cHyp-B to be retained as osmolytes.

The genes encoding the P. denitrificans pathway were individually disrupted by the insertion of antibiotic-resistance cassettes (Methods and Supplementary Table 7). The growth phenotypes are consistent with the predicted functions (Supplementary Discussion).

Here we have used homology modelling and metabolite docking to several proteins encoded by a gene cluster to guide the in vitro assignment of the previously undocumented Hyp-B 2-epimerase activity to 2PMQ, a structure determined by the Protein Structure Initiative. With knowledge of the catalytic capabilities of enzyme superfamilies, we also predicted the pathway that catabolizes Hyp-B 2-0-ketogluutarate. These predictions were verified by metabolomics and genetics. Finally, we used transcriptomics to demonstrate that Hyp-B 2-epimerase is a ‘switch’ that determines whether the tHyp-B is accumulated as an osmolyte or catabolized as carbon and nitrogen source.

Orthologues of HpbD can be identified in 20 microbial species (http://sld.rvb.uncfs.edu/), so both the in vitro activity and the in vivo functional assignments identified in this study can be extended to these proteins and organisms. Moreover, we expect that the Hyp-B 2-epimerase activity assigned to HpbD will be used to facilitate the discovery of the proteins and organisms.
29. Van Spanning, R. J. et al. A method for introduction of unmarked mutations in the genome of Paracoccus denitrificans: construction of strains with multiple mutations in the genes encoding periplasmic cytochromes c550, c551, and c553. J. Bacteriol. 173, 6962–6970 (1991).

30. Matsson, M., Ackrell, B. A., Cochran, B. & Hederstedt, L. Carboxin resistance in Paracoccus denitrificans conferred by a mutation in the membrane-anchor domain of succinate:quinone reductase. Arch. Microbiol. 170, 27–37 (1998).

Supplementary Information is available in the online version of the paper.

Acknowledgements This research was supported by cooperative agreements from the US National Institutes of Health (U54GM093342, U54GM074945 and U54GM094662). Molecular graphics and analyses were performed with the University of California, San Francisco (UCSF) Chimera package. Chimera is developed by the Resource for Biocomputing, Visualization, and Informatics at UCSF (supported by National Institutes of Health P41-GM103311). Use of the Advanced Photon Source, an Office of Science User Facility operated for the US Department of Energy (DOE) Office of Science by Argonne National Laboratory, was supported by the US DOE under contract no. DE-AC02-06CH11357. Use of the Lilly Research Laboratories Collaborative Access Team (LRL-CAT) beamline at Sector 31 of the Advanced Photon Source was provided by Eli Lilly Company, which operates the facility.

Author Contributions S.Z., R.K., A.S., M.W.V., B.M.W., S.B., J.B.B., B.S.H., R.D.S., P.C.B., S.C.A., J.V.S., J.A.G., J.E.C. and M.P.J. designed the research. S.Z., R.K., A.S., M.W.V., B.M.W., J.B.B., B.S.H. and R.D.S. performed the research. S.Z., R.K., A.S., M.W.V., B.M.W., J.B.B., B.S.H., R.D.S., P.C.B., S.C.A., J.V.S., J.A.G., J.E.C. and M.P.J. analysed data. S.Z., R.K., A.S., M.W.V., B.M.W., S.B., J.B.B., B.S.H., R.D.S., P.C.B., S.C.A., J.V.S., J.A.G., J.E.C. and M.P.J. wrote the paper.

Author Information The atomic coordinates and structure factors for APO Hyp-B 2-epimerase (HpbD) and tHyp-B-liganded HpbD are deposited in the Protein Data Bank under accession numbers 2PMQ and 4H2H, respectively. Reprints and permissions information is available at www.nature.com/reprints. The authors declare competing financial interests: details accompany the paper on www.nature.com/nature. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to P.C.B. (babbitt@cgl.ucsf.edu), S.C.A. (steve.almo@einstein.yu.edu), J.V.S. (jswedle@illinois.edu), J.A.G. (j-gerlt@illinois.edu), J.E.C. (j-cronan@life.uiuc.edu) or M.P.J. (matt.jacobson@ucsf.edu).
Homology modelling and docking. Sequence similarity network analysis. All sequences from the MLE subgroup in the Structure–Function Linkage Database (SFLD)\textsuperscript{2} were used in the MLE subgroup network analysis. BLAST analyses were performed with these sequences as queries in an all-by-all fashion. The details have been described previously\textsuperscript{4}.

Sequences in the cHyp oxidase and PyrH2C deaminase networks were collected by BLAST, using red and blue dots in Supplementary Fig. 6 as queries, and 10\textsuperscript{−100} as BLAST E-value cutoff. The Pythoscose v. 1.0 program\textsuperscript{11} was used to make the two networks.

Homology modelling and docking. The models of HpbJ and HpbB1 were built with our in-house software Protein Local Optimization (PLOP, marketed as Prime by Schrödinger LLC). The template PDBs used for HpbJ and HpbB1 were 1R9L and 3N0Q, respectively. The sequence alignment of each pair of target and template was made by the I-INS-i method in MAFFT v. 6.925b (ref. 32). While constructing the models we included both the metal ions and the co-crystallized ligands (if any) from the templates. For docking, 2PMQ, the 1.72-A˚ X-ray apo structure of HpbD, was used. The structures were processed by Protein Preparation Wizard in Schrödinger Suite 2009 (ref. 33) before docking.

Two different libraries were used for docking in the active site of HpbD. The large metabolite library is the KEGG metabolite library plus potential substrates for members of the enolase superfamily not found in KEGG. The small library for focused docking to HpbD contained 31 betaines and betaine-like metabolites.

The KEGG metabolite library was generated by the following steps. First, we obtained 14,039 compounds from the KEGG COMPOUND database; then, we used LigPrep\textsuperscript{18} in Schrödinger Suite 2009 to convert each compound from two dimensions to three dimensions and to enumerate up to 32 chiral forms. During this process, compounds with unspecified chemical groups (listed as ‘R’), polymers and monomeric ions were automatically removed. Next, we removed compounds with molecular masses greater than 400 Da because we did not expect these to fit into the active site of HpbD, as well as duplicates generated by LigPrep preparation. We obtained 82,952 unique KEGG ligands.

Potential substrates for the enolase superfamily proteins include all dipeptides (formed by 20 standard amino acids), several types of N-capped (N-succinyl, N-acetyl), N-formiminino, N-formyl and N-carbamoyl amino acids, acid sugars (monosaccharides, diacid sugars, uronate sugars, 6-deoxy acid sugars and phospho sugars) and their corresponding enolates (that is, high-energy intermediates); these also were processed by LigPrep. After combining the KEGG metabolite library with these additional potential substrates for members of the enolase superfamily and removing duplicates, the library used for docking into the active site of HpbD contained 87,098 unique ligands.

The betaine library used for docking to the active site of HpbD contains 31 betaines and betaine-like metabolites, including dimethylsulfoniopropionate (DMSP), ectoine, 5-hydroxyectoine and trigonelline; the compounds are listed in Supplementary Table 2. The members of this library also were processed by LigPrep.

Two docking methods were used. Glide XP docking followed by MM-GBSA was used with HpbD; the details have been described previously\textsuperscript{30}. The Glide XP docking method\textsuperscript{30} was used with HpbJ.

In vitro activity measurements. Cloning, expression, and purification of the 2PMQ (HpbD). The protein sample was provided by the NYSGXRC structural genomics centre (PSI-2; U54GM07495).

Cloning, expression, and purification of the 2PMQ orthologue (HpbD) from P. denitrificans. The protein sample was provided by the NYSGXRC structural genomics centre.

Cloning, expression and purification of the HypF from P. denitrificans. The hypf gene was amplified by PCR using primers P17 and P18 and genomic DNA of P. denitrificans as templates. The PCR product was digested with NdeI and BglII and ligated to pET15b expression vector, yielding plasmid pRK9. The cloned HypF gene was amplified by PCR using primers P17 and P18 and genomic DNA of P. denitrificans 2PMQ (HpbD). The protein sample was provided by the NYSGXRC structural genomics centre. The vector was transformed into Rosetta2 (DE3) with a high expression efficiency.

In vitro activity measurements. Expression of HpbD. Plasmid 9417a2BNN12p1, obtained from NYSGXRC stock clones\textsuperscript{37}, consists of a codon optimized HpbD gene in pSGX2, a derivative of pET28B (Novagen), with the amino-terminal methionine of Hpdb changed to the sequence MAHHHHHHSL. The vector was transformed into Rosetta2 (DE3) with a high expression efficiency.

Crystallization and structure solution of HpbD. Crystals were obtained by vapour diffusion at 18 °C using the sitting-drop vapour-diffusion method in 96-well IntelliPlates (Art Robbins). Equal volumes of protein (24.6 mg ml\textsuperscript{−1} in 10 mM HEPES pH 7.5, 150 mM NaCl, 20 mM imidazole, 10% (w/v) glycerol) supplemented with 0.1% (v/v) Tween 20 and disrupted by sonication. Cellular debris was removed by centrifugation, and the supernatant was applied to a 1.0 ml metal-affinity column (Ni\textsuperscript{2+} Sepharose High Performance; GE Healthcare) pre-equilibrated with buffer A. The column was washed with five column volumes of buffer A and subsequently eluted with two column volumes of the same buffer containing 300 mM imidazole. Eluted protein was pooled and applied to a 120-ml Superdex 2000 column (GE Healthcare) equilibrated with buffer B (10 mM HEPES pH 7.5, 150 mM NaCl, 5% (v/v) glycerol). Fractions with more than 95% purity by SDS–PAGE analysis were pooled, concentrated by centrifugal ultrafiltration, snap-frozen in liquid nitrogen and stored at −80 °C.

Crystalization and structure solution of HpbD. Crystals were obtained by vapour diffusion at 18 °C using the sitting-drop vapour-diffusion method in 96-well IntelliPlates (Art Robbins). Equal volumes of protein (24.6 mg ml\textsuperscript{−1} in 10 mM HEPES pH 7.5, 150 mM NaCl, 20 mM imidazole, 10% (w/v) glycerol) supplemented with 0.1% (v/v) Tween 20 and disrupted by sonication. Cellular debris was removed by centrifugation, and the supernatant was applied to a 1.0 ml metal-affinity column (Ni\textsuperscript{2+} Sepharose High Performance; GE Healthcare) pre-equilibrated with buffer A. The column was washed with five column volumes of buffer A and subsequently eluted with two column volumes of the same buffer containing 300 mM imidazole. Eluted protein was pooled and applied to a 120-ml Superdex 2000 column (GE Healthcare) equilibrated with buffer B (10 mM HEPES pH 7.5, 150 mM NaCl, 5% (v/v) glycerol). Fractions with more than 95% purity by SDS–PAGE analysis were pooled, concentrated by centrifugal ultrafiltration, snap-frozen in liquid nitrogen and stored at −80 °C.

Crystalization and structure solution of HpbD. Crystals were obtained by vapour diffusion at 18 °C using the sitting-drop vapour-diffusion method in 96-well IntelliPlates (Art Robbins). Equal volumes of protein (24.6 mg ml\textsuperscript{−1} in 10 mM HEPES pH 7.5, 150 mM NaCl, 20 mM imidazole, 10% (w/v) glycerol) supplemented with 0.1% (v/v) Tween 20 and disrupted by sonication. Cellular debris was removed by centrifugation, and the supernatant was applied to a 1.0 ml metal-affinity column (Ni\textsuperscript{2+} Sepharose High Performance; GE Healthcare) pre-equilibrated with buffer A. The column was washed with five column volumes of buffer A and subsequently eluted with two column volumes of the same buffer containing 300 mM imidazole. Eluted protein was pooled and applied to a 120-ml Superdex 2000 column (GE Healthcare) equilibrated with buffer B (10 mM HEPES pH 7.5, 150 mM NaCl, 5% (v/v) glycerol). Fractions with more than 95% purity by SDS–PAGE analysis were pooled, concentrated by centrifugal ultrafiltration, snap-frozen in liquid nitrogen and stored at −80 °C.
CaCl₂ 0.04, MgSO₄·7H₂O 0.001, FeSO₄·7H₂O 1.1, with or without 1.6 g of NH₄Cl as nitrogen source. *P. denitrificans* PD1222 was grown aerobically at 30 °C in mineral medium supplemented with either glucose/succinate, uHyp-B or methanol at the same concentration (20 mM). *E. coli* strain TOP10 (Invitrogen) was used for plasmid maintenance, propagation and cloning purposes. *E. coli* strain S17-1 was used for conjugation. Strains used are listed in Supplementary Table 7. To study the role of uHyp-B in osmoprotection, cultures were grown in minimal medium with glucose, in the presence or absence of 500 mM NaCl. *E. coli* cultures were grown at 37 °C in Luria–Bertani medium. Antibiotics were used at the following concentrations (in μg ml⁻¹): kanamycin sulphate, 50; chloramphenicol, 35; sodium ampicillin, 100.

**Construction of disruption mutants.** Gene inactivation mutants were generated in *P. denitrificans* by homologous recombination or electroporation. Molecular biology protocols. Chromosomal DNA was isolated from 3–5 ml of *P. denitrificans* PD1222 cell cultures with a DNeasy Blood and Tissue Kit (Qiagen) or a Wizard Genomic DNA Purification Kit (Promega). Restriction enzymes, DNA polymerases and T4 DNA ligases were purchased from New England Biolabs, Fermentas, Invitrogen or Promega. Plasmids were prepared from *E. coli* TOP10 cells with a Plasmid Mini Kit (Qiagen).

**Gene disruption plasmids.** Most plasmids for the construction of gene disruptions were obtained by a standard protocol in which the appropriate chromosomal segments were amplified from *P. denitrificans* PD1222 genomic DNA using *Pfu* polymerase followed by insertion of the PCR products into the pGEM T Easy vector (Promega). The resulting plasmids were then digested with EcoRV (pRK1), NraI (pRK2), BsmBI (pRK4) and ligated to a 900-base-pair (bp) fragment blunt-ended chromamphenicol resistance (*cat*) cassette. These plasmids were then used as PCR templates with the same primers and the products were ligated to vector pSUP202, which had been digested with EcoRI and treated with the Klenow fragment of DNA polymerase I, plus the four dNTPs to give the plasmids used for gene disruption. Primer sets P1 + P2, P3 + P4 and P7 + P8 gave rise to plasmids pRK1, pRK2 and pRK4, respectively. Plasmid pRK3 was obtained similarly by using primers P9 and P10 except that the original PCR product was inserted into vector pCR2.1-TOPO (Invitrogen) and the *cat* cassette was inserted into the HinclI site. Plasmid pRK3 was obtained similarly by using primers P5 and P6 except that a 1,400-bp kanamycin resistance cassette was inserted into the BsmBI site of the intermediate pGEM T Easy construct. For plasmids pRK6 and pRK8, the products of PCR (primers P11 + P12 and P15 + P16, respectively) amplification from chromosomal DNA were inserted into pETBlue-1 (Novagen) and the kanamycin resistance cassette was inserted into the SflI site of this plasmid. Plasmid pRK7 was obtained by the same manipulations with primers 13 and 14 except that the chloramphenicol cassette was inserted into the SföI site.

**Expression plasmids.** These plasmids were constructed as above except that the PCR products obtained from chromosomal DNA contained the promoter and coding sequences and were directly ligated to vector pSUP402.4 that had been digested with EcoRI and treated with the Klenow fragment of DNA polymerase I plus the four dNTPs. Primers P19 + P2, P20 + P4 and P21 + P10 gave rise to plasmids pRK10, pRK11 and pRK12, respectively. High-level protein expression plasmids pRK9 and pRK13 were obtained by insertion of PCR products obtained from chromosomal DNA into pGEM T Easy. Plasmid pRK9 was obtained by ligation of the Ndel and BgII hypF fragment of the intermediate plasmid into pET15b digested with Ndel and BamHI, whereas pRK13 resulted from ligation of the hypO Ndel–BglI fragment of the pGEM T Easy intermediate plasmid into pET15b cut with the same enzymes.

**Cell preparation for gene expression analysis.** *P. denitrificans* PD1222 wild-type or mutant cultures were grown in 10 ml of minimal medium with 20 mM glucose as carbon source to a Do₅₀ of 0.4. The cells were pelleted by centrifugation (4,000 g for 10 min at 4 °C). The cell pellet was washed twice and resuspended in 10 ml of minimal medium lacking a carbon source. The cultures were divided into two 5-ml aliquots. One of these was added to one aliquot, and the second was incubated for 15 min at 30 °C before being transferred to ice. After verification of the concentrated cell density by multiple 50-fold dilutions into minimal medium (calculated Do₅₀ 267 ± 0.2, 1-ml aliquots of cell suspensions with a Do₅₀ of 6 were prepared in Eppendorf tubes on ice in minimal medium. Hyp-B (20 mM) was added quickly to half of the samples before incubation in a 30 °C water bath; at time points of 0, 2, 5 and 15 min, samples were transferred on ice to a centrifuge at 4 °C for pelleting at 16,000g for 2 min. The supernatant was then removed and the cell pellets were flash-frozen in liquid nitrogen. The process of collecting time-point samples from 30 °C to liquid nitrogen took about 3 min. Samples were stored at −80 °C before analysis.

**To ascertain the effect of osmotic stress on the levels of the metabolites from tHyp-B catabolism, *P. denitrificans* was grown as described above except that a replicate culture with 0.5 M NaCl was also prepared. After these cultures reached a Do₅₀ of about 0.5, each was centrifuged at 10,000 g, and resuspended in minimal medium without carbon source. The cell suspension was then depleted of catabolic metabolites by incubation for 15 min at 30 °C before being transferred to ice. After verification of the concentrated cell density by multiple 50-fold dilutions into minimal medium (calculated Do₅₀ 267 ± 0.2, 1-ml aliquots of cell suspensions with a Do₅₀ of 6 were prepared in Eppendorf tubes on ice in minimal medium. Hyp-B (20 mM) was added quickly to half of the samples before incubation in a 30 °C water bath; at time points of 0, 2, 5 and 15 min, samples were transferred on ice to a centrifuge at 4 °C for pelleting at 16,000g for 2 min. The supernatant was then removed and the cell pellets were flash-frozen in liquid nitrogen.**

**RNA analysis.** Analysis of these samples followed the procedure from ref. 48. The cell pellets were extracted with 0.375 ml of 10 mM ammonium bicarbonate (pH 9.2) in 90% acetonitrile by pipetting followed by 15 min of vortex-mixing at 22 °C. The extraction was cleared of cell debris by two rounds of centrifugation at 16,000g before analysis. Samples were applied to a custom 11-T liquid-trap quadrupole Fourier transform mass spectrometer (Thermo–Fisher Scientific) with an Agilent 1200 high-performance liquid chromatography (HPLC) system equipped with a Sequest Zic-HILIC column (2.1 mm × 150 mm) previously equilibrated with the extraction buffer (solvent B). Solvent A was 10 mM ammonium bicarbonate pH 9.2. Each extracted sample was injected in 100 μl for three separate analyses. The flow rates were eluted at a flow rate of 200 μl min⁻¹ with the following elution profile: 100% B for 17 min, a linear gradient from 100% B to 40% B over 3 min, and another linear gradient from 40% to 100% B over 15 min. Data were collected at a resolution of 50,000 with full scan set to m/z 100–1,000, and duplicate samples were analysed individually in either positive or negative mode. Data analysis was performed with the QualBrowser application of Xcalibur (Thermo–Fisher Scientific) (Supplementary Figs 10 and 11). For metabolites, tHyp-B, N-methyl tHyp, cHyp, and t-ketogluconate, standards were run to verify retention time. Unfortunately, the *P. denitrificans* samples seemed to damage the Zic-HILIC column over time, because after dozens of runs many of the peaks broadened and had retention times. Although HPLC analyses of betaine derivatives of biological origin have been reported56–61, the metabolite analysis of bacteria with the use of betaines has not. It is assumed that the extremely large concentrations of tHyp-B accumulated in *P. denitrificans* cells were to blame as a result of overloading the column.
31. Barber, A. E. & Babitt, P. C. Pythoscape: a framework for generation of large protein similarity networks. Bioinformatics 28, 2845–2846 (2012).
32. Katoh, K., Kuma, K., Toh, H. & Miyata, T. MAFFT version 5: improvement in accuracy of multiple sequence alignment. Nucleic Acids Res. 33, 511–518 (2005).
33. Suite, S. 2009 Protein Preparation Wizard; Epik version 2.0; Impact version 5.5; Prime version 2.1 (Schrodinger LLC, 2009).
34. Suite, S. 2009 LigPrep, version 2.3 (Schrodinger LLC, 2009).
35. Kalyanaraman, C., Bernacki, K. & Jacobson, M. P. Virtual screening against highly charged active sites: identifying substrates of α-β barrel enzymes. Biochemistry 44, 2059–2071 (2005).
36. Friesner, R. A. et al. Extra precision glide: docking and scoring incorporating a model of hydrophobic enclosure for protein–ligand complexes. J. Med. Chem. 49, 6177–6196 10.1021/jm051256o (2006).
37. Sauder, M. J. et al. High throughput protein production and crystallization at NYSGXRC. Methods Mol. Biol. 426, 561–575 (2008).
38. Fox, B. G. & Blommel, P. G. Autoinduction of protein expression. Curr. Protocols Protein Sci. 5.23.1–5.23.18 (2009).
39. Studier, F. W. Protein production by auto-induction in high density shaking cultures. Protein Expr. Purif. 41, 207–234 (2005).
40. Leslie, A. G. The integration of macromolecular diffraction data. Acta Crystallogr. D Biol. Crystallogr. 62, 48–57 (2006).
41. Evans, P. Scaling and assessment of data quality. Acta Crystallogr. D Biol. Crystallogr. 62, 72–82 (2006).
42. McCoy, A. J. et al. Phaser crystallographic software. J. Appl. Cryst. 40, 658–674 (2007).
43. Zwart, P. H. et al. Automated structure solution with the PHENIX suite. Methods Mol. Biol. 426, 419–435 (2008).
44. Emsley, P. & Cowtan, K. Coot: model-building tools for molecular graphics. Acta Crystallogr. D Biol. Crystallogr. D60, 2126–2132 (2004).
45. Schuttelkopf, A. W. & van Aalten, D. M. PRODRG: a tool for high-throughput crystallography of protein–ligand complexes. Acta Crystallogr. D Biol. Crystallogr. 60, 1355–1363 (2004).
46. Simon, R., Priefer, U. & Puhler, A. A broad host range mobilization system for in vivo genetic engineering: transposon mutagenesis in Gram negative bacteria. Nature Biotechnol. 1, 784–791 (1983).
47. Livak, K. J. & Schmittgen, T. D. Analysis of relative gene expression data using real-time quantitative PCR and the 2 –ΔΔCT method. Methods 25, 402–408 (2001).
48. Erb, T. J. et al. A RubisCO-like protein links SAM metabolism with isoprenoid biosynthesis. Nature Chem. Biol. 8, 926–932 (2012).
49. Chambers, S. T. & Kunin, C. M. Isolation of glycine betaine and proline betaine from human urine. Assessment of their role as osmoprotective agents for bacteria and the kidney. J. Clin. Invest. 79, 731–737 (1987).
50. Oufir, M. et al. Simultaneous measurement of proline and related compounds in oak leaves by high-performance ligand-exchange chromatography and electrospray ionization mass spectrometry for environmental stress studies. J. Chromatogr. A 1216, 1094–1099 (2009).
51. Li, C., Hill, R. W. & Jones, A. D. Determination of betaine metabolites and dimethylsulfoniopropionate in coral tissues using liquid chromatography–time-of-flight mass spectrometry and stable isotope-labeled internal standards. J. Chromatogr. B Analty. Technol. Biomed. Life Sci. 878, 1809–1816 (2010).