AFRL-SR-AR-TR-09-0102

REPORT DOCUMENTATION PAGE

Inactive

PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ORGANIZATION.

1. REPORT DATE (DD-MM-YYYY)	2. REPORT TYPE	3. DATES COVERED (From - To)
2/26/2009	final	9/1/05 – 8/31/08

4. TITLE AND SUBTITLE
Exploring Convergent Evolution to Provide a Foundation for Protein Engineering

5a. CONTRACT NUMBER

5b. GRANT NUMBER
FA9550-06-0014
06-1-0014

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Copley, Shelley D.

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
University of Colorado at Boulder
University of California, San Francisco

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Air Force Office of Scientific Research
875 North Randolph Street
Suite 325, Rm 3112
Arlington, VA 22203
Dr. Walter Kozumbo/NL

10. SPONSOR/MONITOR'S ACRONYM(S)
AFOSR

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Distribution A: approved for public release

13. SUPPLEMENTARY NOTES

20101110208

14. ABSTRACT

"Pathway holes" correspond to steps in a pathway for which a homologue of known enzymes cannot be identified even though other enzymes in the pathway are present. Pathway holes can occur when an organism has independently evolved an enzyme to provide a catalytic function supplied by a structurally and possibly mechanistically different enzyme in other organisms. We have approached the identification of convergently evolved enzymes that fill pathway holes using both bioinformatic and experimental approaches. We have developed algorithms for generating divergent sets of protein sequences (Divergent Set) and for clustering proteins according to the patterns of motifs found by MEME (Motif Cluster). These algorithms can be used to search the database for sequences that are annotated with the same function and to cluster the sequences into families, thus revealing groups of sequences that show little relationship to each other and may therefore have evolved independently. We have identified a "missing enzyme" in Thermoplasma volcanium and are tracking down another enzyme that is "missing" in most Archaea. We have also developed the framework and ontology for a database of convergently evolved enzymes and have begun to populate the database.

15. SUBJECT TERMS
protein engineering, convergent evolution

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Shelley D. Copley
U	U	U	UU		19b. TELEPHONE NUMBER (Include area code) 303-492-6328

Goals of the Project

Enzymes are superb catalysts. They accelerate reactions by up to 17 orders of magnitude. With few exceptions, they catalyze reactions with extremely high regioselectivity and enantioselectivity. In addition, they are environmentally benign. The conditions required to produce and to use enzymes are "green" – they do not require toxic organic solvents and/or extremely high temperatures and pressures that are costly in terms of energy usage. For over two decades, protein engineers have pursued the goal of designing enzymes to catalyze specific reactions that are not found in nature. In some cases, the goal is to produce a catalyst where there currently is none. In others, the goal is to improve upon a process that requires toxic chemicals and/or high energy inputs.

Studies of protein structures and sequences have revealed that nature has often used different structural scaffolds, and sometime entirely different mechanistic strategies, to catalyze particular reactions. An understanding of the range of scaffolds that can catalyze a certain reaction will aid design efforts by providing protein engineers with the possibility of multiple starting places for catalyst design. In some cases, a particular mechanistic approach may be more suitable for certain applications, perhaps because of differences in cofactor requirement or effects of pH on catalytic efficiency. Furthermore, one of a set of convergently evolved enzymes might be more stable or more easily expressed. The goal of this project was to employ both experimental and bioinformatics methods to improve our knowledge of the number of reactions for which catalysis can be provided by multiple protein scaffolds.

Development of the DivergentSet and MotifCluster Algorithms

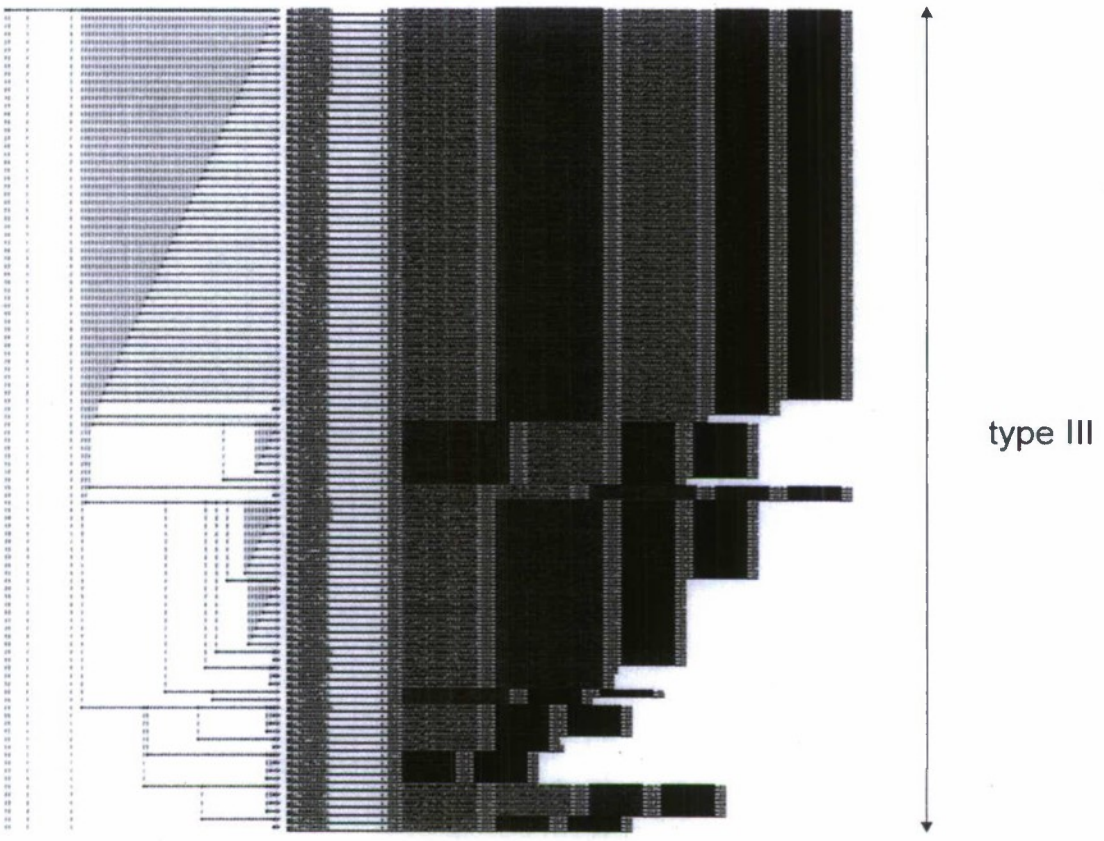
Using support from this grant, we developed two software packages that provide key infrastructure for identification of "missing enzymes". "Missing enzymes" are enzymes that catalyze reactions that are known to occur in an organism, but cannot be identified by homology methods because they are either too divergent to be recognized or because they have evolved convergently and therefore are not homologous to known enzymes. We are particularly interested in the latter class of missing enzymes because we want to discover novel scaffolds and mechanisms for catalyzing chemical reactions.

The first algorithm we developed is DivergentSet,[i] which allows rapid selection of sets of protein sequences that share no more than a user-selected level of identity (typically 40 – 60%). Divergent sets that represent the diversity of protein sequences in a family are important for identification of motifs (highly conserved sequences) that have been conserved because they contribute to structure and/or function. Multiple sequence alignments and motif analysis of closely related proteins are not informative because regions that are conserved because they play key roles in structure and/or function cannot be distinguished from those that are conserved simply because of recent common ancestry. Before the development of DivergentSet, divergent sets of proteins were selected by a laborious manual procedure. DivergentSet collects a divergent set of proteins in a short period of time, and its speed allows repeated runs for selection of multiple divergent sets. The availability of multiple divergent sets allows analysis of how the motifs identified by motif finding algorithms depend on the exact identity of the sequences included in the set. Using multiple divergent sets, we found that the results of

a widely used motif finder, MEME, are unstable to the choice of sequences used as input for MEME analysis, a finding with far-reaching implications for motif identification. The paper describing DivergentSet was published in Molecular and Cellular Proteomics, the best specialty journal in proteomics, which has an impact factor comparable to that of PNAS.

The second software package we developed, MotifCluster,[ii] provides a novel way of detecting distantly related homologs, one of the key aims of the proposal. Unlike alignment-based methods, which require a reasonably good seed alignment, MotifCluster takes a motif-based approach in which the presence of shared motifs in a divergent set of sequences is used to infer relationships among the sequences and to cluster sequences according to patterns of shared motifs. Using a previously published set of gold- and silver-standard protein families,[iii] we discovered that two or more shared motifs are diagnostic of an evolutionary relationship, but that one motif might be shared by chance. MotifCluster was able to assign sequences to the correct families with 0.17% false positives and no false negatives. MotifCluster also provides a number of user interface innovations that greatly assist in analysis of protein families and superfamilies, including the ability to conveniently visualize motifs on a multiple sequence alignment and to map motifs onto the structures of the proteins (where available). It also allows the user to highlight differences within motifs that differentiate members of a protein family, or different protein families, often providing insight into divergence of mechanism within families.

An example of the use of these algorithms for identification of convergently evolved enzymes is shown in Figure 1. This was a test case in which we analyzed sequences annotated as pantothenate kinases. Three unrelated classes of pantothenate kinases have been identified. Using known examples of each class, we assembled a divergent set representing all three classes and analyzed motifs using MEME.[iv] The motifs were then clustered using MotifCluster (see Figure 1). The results show that MotifCluster was able to cluster the sequences into the three known classes. Most Type III sequences share 4-5 motifs, although some have as few as two of these. Most Type II sequences share 2-5 motifs. Type I sequences are the most divergent, and thus have fewer shared motifs, but these sequences still share at least one motif and lack motifs characteristic of the other two Types. This example illustrates the utility of MotifCluster for separating proteins that are annotated with a common function and have two or more shared motifs, and are therefore evolutionarily related, from those that do not share common motifs and may have arrived at a common function by convergent evolution.


type III

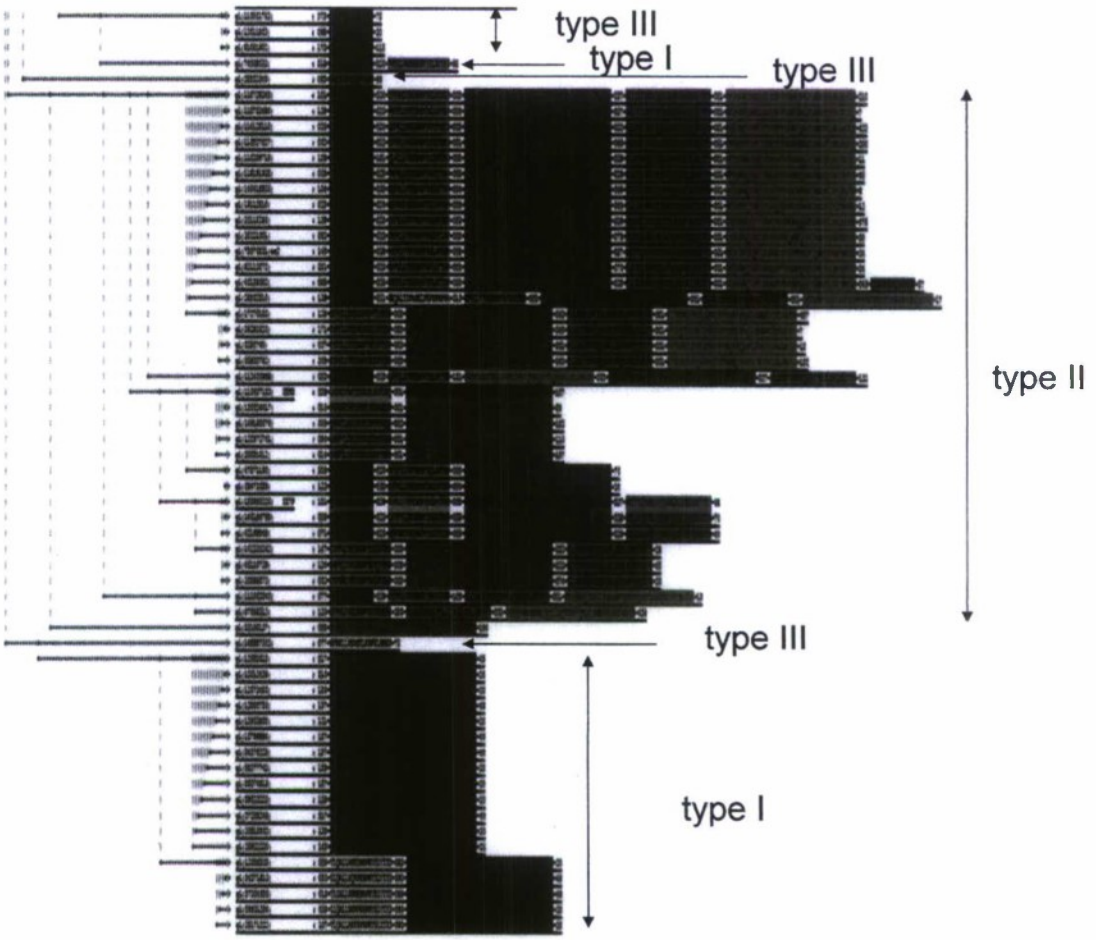


Figure 1. MotifCluster clusters pantothenate kinases into three classes based upon shared motifs found by MEME.

DivergentSet and Motifcluster are available as publicly available web servers (at http://bmf.colorado.edu/divergentset and http://bmf.colorado.edu/motifcluster, respectively), and the source code is available as open-source projects for other investigators to build on.

Experimental searches for "missing enzymes"

We have searched for two "missing enzymes". The branched chain aminotransferase enzyme in *Thermoplasma volcanium* is a "missing enzyme". We identified this enzyme (TVN0402) by generating a library of *T. volcanium* genomic DNA and transforming it into a strain of *E. coli* that lacks IlvE (an aminotransferase involved in biosynthesis of valine, leucine, and isoleucine, see Figure 2). Plasmids were recovered from colonies that were able to grow on M9/glucose and the genomic insert was sequenced. TVN0402 has branched-chain aminotransferase activity *in vitro*. The genomic insert that complemented the lack of IlvE contained two full genes (TVN0402, annotated as a predicted transcription regulator, and TVN0403, annotated as a

phosphopantetheine adenylyltransferase),as well as two partial genes. Sub-cloning of individual genes showed that TVN0402 encodes a protein with branched-chain aminotransferase activity. We still need to measure kinetic parameters for the enzyme before publishing this result.

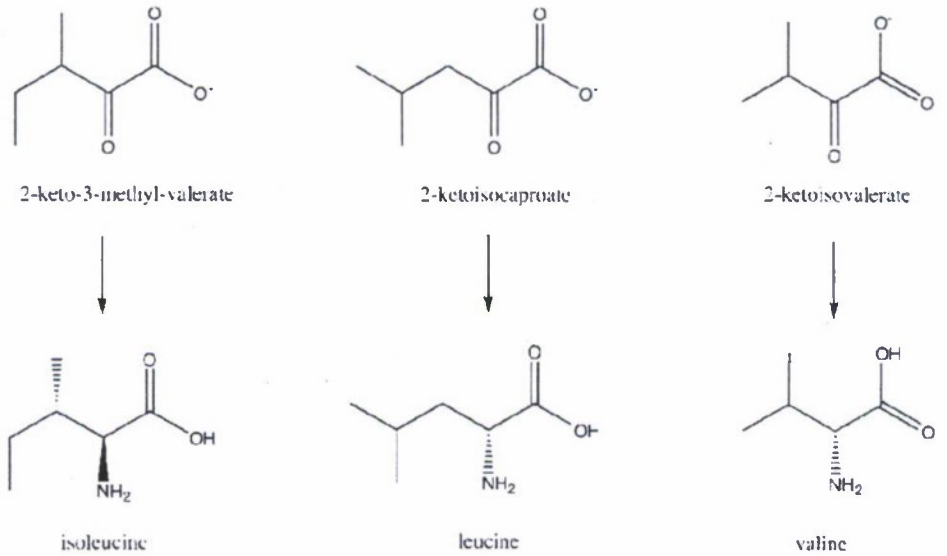


Figure 2. Reactions catalyzed by IlvE in *E. coli*. The comparable enzyme is missing in *Thermoplasma volcanium*.

Pantothenate kinase, the first enzyme in the pathway for synthesis of Coenzyme A (see Figure 3) is a "missing enzyme" in Archaea. As described above, three different types of pantothenate kinases have been discovered. However, most Archaea lack homologs of any of these pantothenate kinases even though they have homologs of other enzymes in the pathway and certainly synthesize CoA. We have cloned genes from *Methanococcus jannaschii* and *Pyrococcus horikoshii* that may encode pantothenate kinases based upon their proximity to other genes encoding other enzymes in the CoA biosynthetic pathway and a distant relationship to mevalonate and homoserine kinases that suggests that these genes may encode kinases with an as yet unidentified specificity. We are expressing these proteins in preparation for assays of pantothenate kinase activity. We are also considering the possibility that Archaea may have convergently evolved a pathway for synthesis of Coenzyme A that differs from that found in other organisms. Solving this puzzle may provide a fourth example of a structural scaffold that can support phosphorylation of pantothenate. This would be an unusual situation in which Nature has utilized multiple scaffolds and possibly mechanistic strategies to catalyze a reaction required for all life. Alternatively, if we find a different pathway, we will discover enzymes that have different substrate specificities from those in known CoA biosynthesis pathways. These enzymes could be useful for biotechnological purposes.

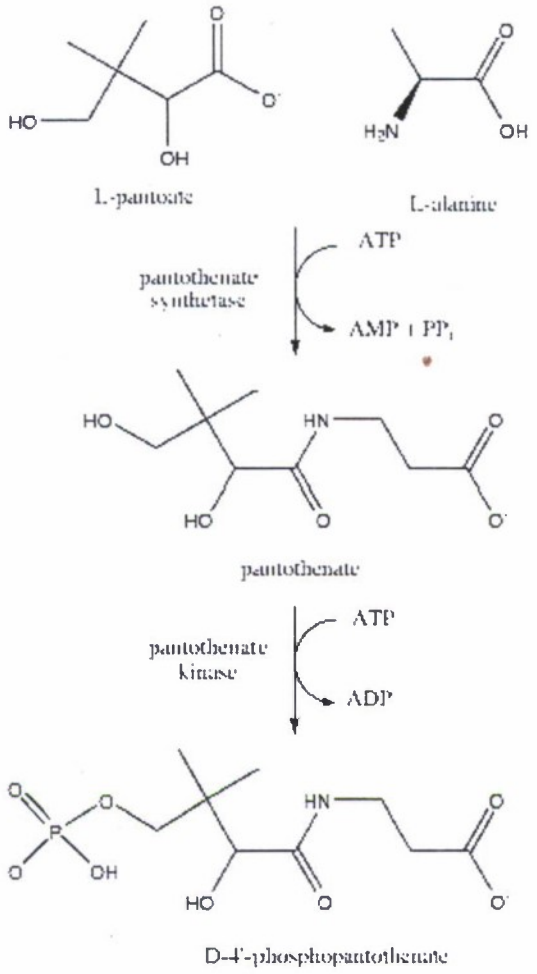


Figure 3. Initial steps in the synthesis of coenzyme A. Pantothenate kinase is a "missing enzyme" in most Archaea.

Progress toward development of a database of convergently evolved enzymes

The role of the UCSF sub-contract effort on this grant was to develop computational infrastructure to incorporate convergently evolved enzymes into our Structure-Function Linkage Database (SFLD).[v] Use of the SFLD is ideal for this purpose as it allows results associated with convergently evolved enzymes to be examined in the context of the so-called canonical enzymes that perform the same functions. Schema elements were developed in MySQL to structure the data and allow development of a graphical user interface (GUI) to serve it. The schema and GUI were based on that of the SFLD and take advantage of its tools for searching for overall or partial reactions using tools similar to those provided by the commonly used ChemDraw software or SMILES strings. Matches to either canonical enzymes or "missing enzymes" identified in this project allow a user to link to those SFLD pages. Specialized enzyme pages were developed for this project that show the reaction and constituent partial reactions, accompanied by an explanation of the different structural classes of enzymes that catalyze a given function. A screen shot of the page under development for pantothenate

kinases is shown in Figure 4. Each page includes information as available for the name of the reaction, a listing of available PDB structures, quaternary structure, its phylogenetic distribution, metal ion, co-factor, co-substrates, and inhibitors. Links are provided to additional pertinent information from source databases including BRENDA,[vi] MetaCyc,[vii] EzCatDB, [viii] and MaCIE.[ix] New tools have recently been developed to query the metabolic context of these and other enzymes in our database. These tools generate potential operons using sequence-based matches of genes of target enzymes and nearby genes to those in both the SEED resource[x] and the MicrobesOnline resource.[xi] Additionally, we have adapted protein similarity networks, a major approach used in the Babbitt lab,[xii] to query relationships among sequence families in the context of their common membership in specific metabolic pathways. Access to information about multiple structural strategies to catalyze specific reactions or constituent partial reactions provides a foundation for exploring the range of enzyme structures likely to be most fruitful for use as starting scaffolds for directed evolution of enzymes capable of catalyzing new reactions.

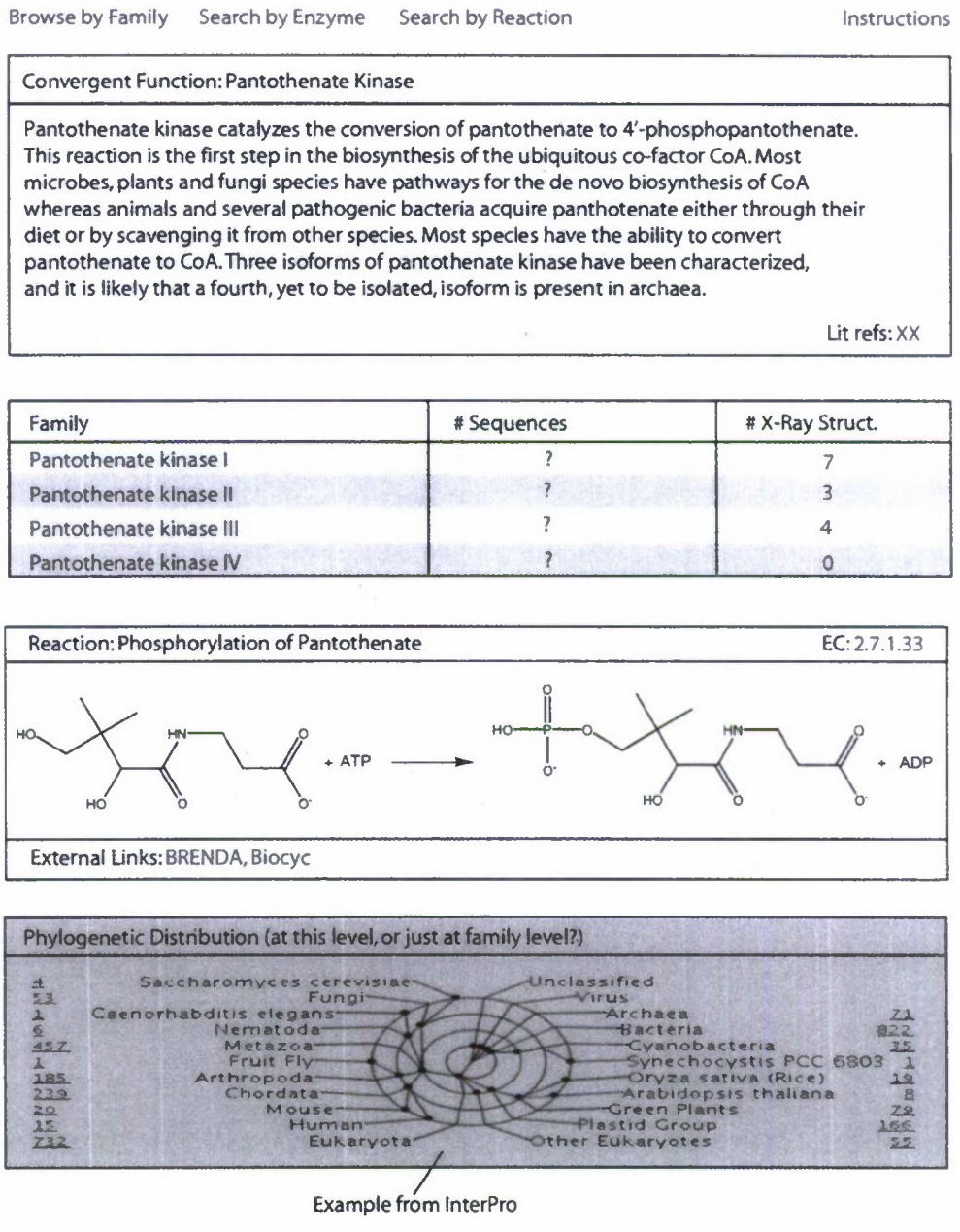


Figure 4. Screen shot of one page of the entry for pantothenate kinases in the database of convergently evolved enzymes we are in the process of populating.

Publications

Widmann, J., Hamady, M., and Knight, R. (2006) DivergentSet, a tool for picking non-redundant

sequences from large sequence collections. *Mol. Cell. Proteomics 5.8*, 1520-1532.

Hamady, M., Widmann, J., Shelley D. Copley, S. D., and Knight, R. "MotifCluster: An Interactive Online Tool for Clustering and Visualizing Sequences Using Shared Motifs", *Genome Biology*, 2008, **9**, R128. (designated as "highly accessed" by the journal)

References

---

[i] Widmann, J., Hamady, M., and Knight, R. (2006) DivergentSet, a tool for picking non-redundant sequences from large sequence collections. *Mol. Cell. Proteomics 5.8*, 1520-1532.

[ii] Hamady, M., Widmann, J., Copley, S. D., and Knight, R. (2008) MotifCluster: an interactive online tool for clustering and visualizing sequences using shared motifs. *Genome Biol 9*, R128.

[iii] Brown, S. D., Gerlt, J. A., Seffernick, J. L., and Babbitt, P. C. (2006) A gold standard set of mechanistically diverse enzyme superfamilies. *Genome Biol 7*, R8.

[iv] Bailey, T. L., and Elkan, C. (1994) in *Proceedings of the Second International Congress on Intelligent Systems for Molecular Biology.* pp 28-36, AAAI Press.

[v] Pegg, S. C.-H., Brown, S., Ojha, S., Huang, C. C., Ferrin, T. E., and Babbitt, P. C. (2005) Representing structure-function relationships in mechanistically diverse enzyme superfamilies. *Pac. Symp. Biocomput. 2005*, 358-369.

[vi] Chang, A., Scheer, M., Grote, A., Schomburg, I., and Schomburg, D. (2009) BRENDA, AMENDA and FRENDA the enzyme information system: new content and tools in 2009. *Nucleic Acids Res 37*, D588-92.

[vii] Caspi, R., Foerster, H., Fulcher, C. A., Kaipa, P., Krummenacker, M., Latendresse, M., Paley, S., Rhee, S. Y., Shearer, A. G., Tissier, C., Walk, T. C., Zhang, P., and Karp, P. D. (2008) The MetaCyc Database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. *Nucleic Acids Res 36*, D623-31.

[viii] Nagano, N. (2005) EzCatDB: the Enzyme Catalytic-mechanism Database. *Nucleic Acids Res 33*, D407-12.

[ix] Holliday, G. L., Almonacid, D. E., Bartlett, G. J., O'Boyle, N. M., Torrance, J. W., Murray-Rust, P., Mitchell, J. B., and Thornton, J. M. (2007) MACiE (Mechanism, Annotation and Classification in Enzymes): novel tools for searching catalytic mechanisms. *Nucleic*

Acids Res 35, D515-20.

[x] Overbeek, R., Begley, T., Butler, R. M., Choudhuri, J. V., Chuang, H. Y., Cohoon, M., de Crecy-Lagard, V., Diaz, N., Disz, T., Edwards, R., Fonstein, M., Frank, E. D., Gerdes, S., Glass, E. M., Goesmann, A., Hanson, A., Iwata-Reuyl, D., Jensen, R., Jamshidi, N., Krause, L., Kubal, M., Larsen, N., Linke, B., McHardy, A. C., Meyer, F., Neuweger, H., Olsen, G., Olson, R., Osterman, A., Portnoy, V., Pusch, G. D., Rodionov, D. A., Ruckert, C., Steiner, J., Stevens, R., Thiele, I., Vassieva, O., Ye, Y., Zagnitko, O., and Vonstein, V. (2005) The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. *Nucleic Acids Res 33*, 5691-702.

[xi] Alm, E. J., Huang, K. H., Price, M. N., Koche, R. P., Keller, K., Dubchak, I. L., and Arkin, A. P. (2005) The MicrobesOnline Web site for comparative genomics. *Genome Res 15*, 1015-22.

[xii] Atkinson, H. J., Morris, J. H., Ferrin, T. E., and Babbitt, P. C. (2009) Using sequence similarity networks for visualization of relationships across diverse protein superfamilies. *PLoS ONE 4*, e4345.