Title

Studies of protein specificity

Author

Fayazmanesh, Nima

Publication Date

2008

STUDIES OF PROTEIN SPECIFICITY

by

NIMA FAYAZMANESH

THESIS

Submitted in partial satisfaction of the requirements for the degree of

MASTER OF SCIENCE

in

BIOPHYSICS

in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA, SAN FRANCISCO

[redacted]

Acknowledgements

I would like to acknowledge many people for helping me with this thesis. I would especially like to thank my advisor, Dr. Matthew Jacobson; my co-advisor, Dr. Patricia Babbitt; and my graduate advisor, Dr. Susan Miller. I also thank my research collaborators, Dr. Chakrapani Kalyanaraman, Dr. Margaret Glasner, Dr. Mohammed Sajid, Dr. Wei Zhou, and Dr. Lily Jan; my colleagues in the Jacobson and Babbitt labs at UCSF; Dr. Kevin Shannon and Ms. Jana Toutolmin of the UCSF Medical Scientist Training Program; Dr. David Agard and Ms. Rebecca Brown of the UCSF Biophysics Graduate Program; and my parents, Dr. Vida Samiian and Dr. Sasan Fayazmanesh, for helping to make this thesis possible.

The text of Chapter 1 is a reprint of the material as it appears in the *Journal of Molecular Biology*. The co-author listed in this publication, Dr. Margaret Glasner, directed and supervised the research that forms the basis for Chapter 1. My specific contributions included analysis of sequence conservation in the *o*-succinylbenzoate synthase/N-acylamino acid racemase (OSBS/NAAAR) family; comparison of structural differences in the active sites and overall structural differences within the OSBS/NAAAR family; comparison of root mean square deviation between pairs of structures; and comparison of capping domain orientation between pairs of structures. I thank Mr. Eric Pettersen and Dr. Matthew Jacobson for their contributions to analysis of domain orientation.

Chapter 2 is the product of collaboration with Dr. Mohammed Sajid, UCSF. A paper based on this work is in preparation to be submitted for publication, and a draft is included here as an appendix. My specific contributions were prediction of catalytic and

S1 subsite residues based on profile-profile alignment of metacaspase sequences to caspase sequences, and structural modeling of metacaspases from *Plasmodium berghei*, *Arabidopsis thaliana*, *Leishmania major* strain Friedlin, and *Trypanosoma brucei.*

Chapter 3 is the result of ongoing collaboration with Dr. Wei Zhou and Dr. Lily Jan, UCSF. My specific contributions included computational docking and rescoring of the KEGG database against the x-ray structure of Kvß and selection of potential substrates for experimental testing. I thank Dr. Chakrapani Kalyanaraman, UCSF for his assistance with physics-based rescoring of docking results.

The ideas, algorithm, and analysis in Chapter 4 are my own. I thank Dr. Chakrapani Kalyanaraman for providing me with the docking and rescoring results that were used to test the algorithm.

Abstract

A Study of Protein Specificity

by

Nima Fayazmanesh

Understanding how proteins bind substrate is a basic question in biology that has implications for protein function prediction, protein engineering and drug design. This thesis consists of four separate studies that deal with protein specificity, structure and function. Chapter 1 analyzes the evolution of structure and function in the *o*-succinylbenzoate synthase/N-acylamino acid racemase (OSBS/NAAAR) family of the enolase superfamily. Although all members of this family catalyze the OSBS reaction, some members, such as the *Amycolatopsis* OSBS/NAAAR, are promiscuous, catalyzing both dehydration and racemization. Evolutionary trace analysis demonstrated that all residues conserved in this family are also found in enolase superfamily members that have different functions, and structural analysis showed that the family exhibits surprising structural variations, including large differences in orientation between the two domains. Chapter 2 studies the role of metacaspase 1 in the life cycle of the malarial parasite *Plasmodium berghei.* The metacaspases are a family of caspase-related cysteine proteases that are found in plants, mammals, and protozoa. We constructed sequence alignments and structural models of *Plasmodium berghei* metacaspase 1 and other metacaspases, which suggest that these enzymes may have specificity for arginine or lysine at the P1 subsite. Chapter 3 attempts to predict the natural substrate of the Shaker family potassium channel ß subunit Kvß2. Kvß2 is an aldo-keto reductase that has been implicated in axonal targeting of Shaker channels. In order to better understand this process, a virtual library of small molecule metabolites was docked against the x-ray structure of Kvß2 from rat, and potential substrates were selected for experimental testing. In Chapter 4, we develop an algorithm that performs consensus scoring using docking or rescoring results from isofunctional proteins. We hypothesize that random error is reduced in the consensus, which should cause the rank of the native substrate to improve. The algorithm was tested on docking and rescoring results for 283 proteins from the enolase superfamily. Although the mean change in native substrate rank was negative, there were specific cases in which consensus scoring caused the rank of the native substrate to improve.

Table of Contents

List of Tables

List of Figures and Illustrations

Introduction

There is a substantial gap between the number of proteins that have been sequenced and the number of sequences with functional annotations that have been verified experimentally. In March 2008, there were ~5.2 million protein sequences and ~670 completely sequenced genomes in the RefSeq and Genome databases at the National Center for Biotechnology Information. However, only approximately 20%, 7%, 10% and 1% of annotated proteins in *Homo sapiens*, *Mus musculus*, *Drosophila melanogaster* and *Caenorhabditis elegans* have been experimentally characterized (Lee et al. 2007). In addition, there were only ~50,000 protein structures in the Protein Data Bank, which represents less than 1% of protein sequences.

Computational methods for protein function prediction attempt to fill this gap by using sequence and structural data to predict function. In this thesis, both sequence- and structure-based methods were used to study protein specificity and function. Sequence-based methods rely on inheritance through homology, which assumes that proteins with similar sequences carry out similar functions; the sequence-based methods used here include sequence-sequence comparisons such as BLAST, profile-based comparisons such as PSI-BLAST, multiple sequence alignment, phylogenetic analysis, evolutionary trace analysis, and genomic operon context. Structure-based methods that were used include molecular modeling, computational docking, and comparison of overall protein structures, active sites, and domain orientation.

Understanding how proteins bind substrate is a fundamental question in biology that has implications for protein function prediction, protein engineering and drug design. This thesis consists of four separate studies that deal with protein specificity, structure and function. Chapter 1, "Evolution of Structure and Function in the *o*-Succinylbenzoate Synthase/N-Acylamino Acid Racemase Family of the Enolase Superfamily," examines how proteins evolve to provide both exquisite specificity and proficient activity. To study this problem, we analyzed the evolution of structure and function in the *o*-succinylbenzoate synthase/N-acylamino acid racemase (OSBS/NAAAR) family, part of the mechanistically diverse enolase superfamily. Although all characterized members of the family catalyze the OSBS reaction, the family is extraordinarily divergent, with some members sharing <15% identity. In addition, a member of this family, *Amycolatopsis* OSBS/NAAAR, is promiscuous, catalyzing both dehydration and racemization. We used evolutionary trace analysis to determine whether there are any sequence motifs that are unique to the OSBS/NAAAR family. We also studied structural variation in this family by superimposing x-ray structures and developed a method to compute differences in orientation between the capping and barrel domains. This work was published in the *Journal of Molecular Biology*.

Chapter 2, "Molecular Modeling and S1 Subsite Prediction of Metacaspase Proteins," analyzes the sequences and structures of *Plasmodium berghei* metacaspase 1 (PbMC1) and other metacaspase proteins in order to better understand the substrate specificity of these proteins. The metacaspases are a family of caspase-related cysteine proteases found in plants, mammals, and protozoa. PbMC1 plays an important role in the life cycle of *Plasmodium berghei*, which causes malaria, by controlling parasite numbers in the mosquito. As such, this protein represents an attractive target for an anti-malarial vaccine. In order to better understand the substrate specificity of PbMC1, and to ultimately develop an inhibitor for this protein, we constructed a profile-profile alignment of metacaspase sequences to caspase sequences. The alignment was used to predict the catalytic and specificity determining residues of PbMC1 and other metacaspases. Structural models were constructed for metacaspases from *Plasmodium berghei*, *Arabidopsis thaliana*, *Leishmania major* strain Friedlin, and *Trypanosoma brucei*. A paper based on this work, which is the result of collaboration with Dr. Mohammed Sajid, UCSF, is in preparation to be submitted for publication.

Chapter 3, "Searching for the Natural Substrate of the Shaker Family K+ Channel ß Subunit," attempts to identify the native substrate of the voltage-gated potassium channel subunit Kvß2. Voltage-gated potassium channels regulate the flow of potassium through cell membranes in response to changes in membrane voltage. The archetypal voltage-gated potassium channel is the Shaker channel (Kv1) from *Drosophila melanogaster*. Mutations at the Shaker locus cause fruit flies to shake their legs under ether anesthesia, and mutations of the homologous gene in humans are associated with episodic ataxia type 1. The ß subunit of the Shaker channel (Kvß2) is an aldo-keto reductase that has been implicated in axonal targeting of Kv1. In order to better understand this process, we attempt to identify the natural substrate of Kvß2, which may provide the link between the oxidation-reduction potential of the cell and targeting of Kv1 to the axon. A virtual library of small molecule metabolites was screened against the x-ray structure of Kvß2 from rat, and potential substrates were selected for experimental validation in the laboratory of Dr. Lily Jan, UCSF.

Chapter 4, "Consensus of Docking Results from Isofunctional Proteins: A Method for Decreasing Random Error in Computational Docking," addresses a fundamental problem in computational docking: distinguishing true binders or native substrates from false positive compounds that do not bind. Consensus scoring methods have previously attempted to address this problem by combining scoring functions from different docking programs. However, these attempts have met with limited success, which may be due to bias in the individual scoring functions that were used. We developed a novel algorithm that creates a consensus by combining computational docking or rescoring results from isofunctional proteins. Isofunctional proteins include homologues from the same isofunctional family, as well as enzymes that are adjacent in a metabolic pathway and bind chemically similar substrates. We hypothesize that reduction of random error in the consensus should cause the rank of the native substrate to improve. The algorithm was tested using docking and rescoring results for 283 proteins from the muconate lactonizing enzyme subgroup of the enolase superfamily.

Chapter 1

Evolution of Structure and Function in the *o*-Succinylbenzoate Synthase/*N*-Acylamino Acid Racemase Family of the Enolase Superfamily

Summary

The text of this chapter is a reprint of the material as it appears in the *Journal of Molecular Biology*. The co-author listed in this publication, Dr. Margaret Glasner, directed and supervised the research that forms the basis for this chapter. My specific contributions were as follows:

1. Analysis of sequence conservation in the *o*-succinylbenzoate synthase/N-acylamino acid racemase (OSBS/NAAAR) family (Figure 1-6).
2. Comparison of structural differences in the active sites (Figure 1-7) and overall structural differences within the OSBS/NAAAR family (Figure 1-8).
3. Comparison of root mean square deviation between pairs of structures (Table 1-2).
4. Comparison of capping domain orientation between pairs of structures (Table 1-3).

The final submitted version of the manuscript is included below.

Evolution of Structure and Function in the *o*-Succinylbenzoate Synthase/*N*-Acylamino Acid Racemase Family of the Enolase Superfamily

Running Title: Evolution of the *o*-Succinylbenzoate Synthase Family

Margaret E. Glasner[1], Nima Fayazmanesh[2], Ranyee A. Chiang[1], Ayano Sakai[3], Matthew P. Jacobson[2], John A. Gerlt[3], and Patricia C. Babbitt[1,2]*

[1]Department of Biopharmaceutical Sciences
[2]Department of Pharmaceutical Chemistry
University of California, San Francisco, California 94143

[3]Departments of Biochemistry and Chemistry,
University of Illinois, Urbana, Illinois 61801

*Corresponding Author

Summary

Understanding how proteins evolve to provide both exquisite specificity and proficient activity is a fundamental problem in biology that has implications for protein function prediction and protein engineering. To study this problem, we analyzed the evolution of structure and function in the *o*-succinylbenzoate synthase/*N*-acylamino acid racemase (OSBS/NAAAR) family, part of the mechanistically diverse enolase superfamily. Although all characterized members of the family catalyze the OSBS reaction, this family is extraordinarily divergent, with some members sharing < 15% identity. In addition, a member of this family, *Amycolatopsis* OSBS/NAAAR, is promiscuous, catalyzing both dehydration and racemization. Although the OSBS/NAAAR family appears to have a single evolutionary origin, no sequence or structural motifs unique to this family could be identified—all residues conserved in the family are also found in enolase superfamily members that have different functions. Based on their species distribution, several uncharacterized proteins similar to *Amycolatopsis* OSBS/NAAAR appear to have been transmitted by lateral gene transfer. Like *Amycolatopsis* OSBS/NAAAR, these might have additional or alternative functions to OSBS because many are from organisms lacking the pathway in which OSBS is an intermediate. In addition to functional differences, the OSBS/NAAAR family exhibits surprising structural variations, including large differences in orientation between the two domains. These results offer several insights into protein evolution. First, orthologous proteins can exhibit significant structural variation, and specificity can be maintained with little conservation of ligand-contacting residues. Second, the discovery of a set of proteins similar to *Amycolatopsis* OSBS/NAAAR supports the hypothesis that new protein functions evolve through promiscuous intermediates. Finally, a combination of evolutionary, structural, and sequence analyses identified characteristics that might prime proteins, such as *Amycolatopsis* OSBS/NAAAR, for the evolution of new activities.

Keywords: enolase superfamily; protein evolution; mechanistically diverse superfamily; substrate specificity; functional promiscuity

Introduction

The evolution of new protein functions is a major puzzle in biochemistry. Given that closely related proteins can have different functions, and distantly related proteins can have the same function, what kinds of structural alterations are required or tolerated during protein evolution? In addition, what characteristics of a particular protein determine its degree of evolvability, or the likelihood that it will evolve a new function? Some previous work has indicated that evolution often proceeds through promiscuous intermediates[1-10] and that conformational flexibility of surface loops near the active site might contribute to promiscuous substrate binding and hence to the evolution of promiscuous functions[11]. Unfortunately, there are still few proteins whose evolution, structure, and function have been analyzed in enough detail to fully evaluate these hypotheses. With the advent of large-scale genomic sequencing we are poised to answer these questions. Understanding how proteins evolve will help address several longstanding problems in biochemistry, including how to redesign proteins in the laboratory and how to predict function from sequence and structure.

Studying protein evolution requires identification of homologous proteins that have evolved to perform different functions, such as those found in mechanistically diverse superfamilies. Mechanistically diverse superfamilies are defined as assemblies of homologous proteins which are unified by a common chemical attribute of catalysis, although overall reactions can be quite different[4]. Here, we focus on the enolase superfamily, which includes enzymes catalyzing at least fourteen different reactions[12]. All enolase superfamily enzymes utilize a common partial reaction in which a proton alpha to a carboxylate is abstracted by a base, leading to a metal-stabilized enolate anion intermediate. Apart from this conserved partial reaction, the overall reactions catalyzed by enzymes in this superfamily are quite divergent, including racemization, β-elimination, and cycloisomerization. Very few residues are required for the superfamily partial reaction; three metal-binding residues are well conserved across the superfamily, but the identity and position of the general base is not universally conserved.

Enolase superfamily proteins are composed of two domains, a ~200 amino acid C-terminal modified $(\beta/\alpha)_8$-barrel domain $[(\beta/\alpha)_7\beta)]$ and a ~100-150 amino acid $\alpha + \beta$ domain comprised of elements from both the N- and C- termini, which we call the capping domain. As with other $(\beta/\alpha)_8$-barrel domain proteins, the active site is nestled in a depression formed by the C-terminal ends of the β-strands of the barrel domain. The capping domain is structurally conserved among all members of the enolase superfamily and has not been found in combination with any other $(\beta/\alpha)_8$-barrel domain protein superfamily, with domains of other folds, or as a single domain protein. Thus, it appears that the two domains have been co-evolving since the origin of the enolase superfamily. The capping domain closes the active site and appears to play a role in determining substrate specificity and conformational changes that occur upon substrate binding. These functions are thought to be primarily mediated by two N-terminal loops, centered around positions 20 and 50 (numbering defined relative to *Escherichia coli* *o*-succinylbenzoate synthase, PDB identifier 1FHV), which will be referred to as the 20s and 50s loops. In most enolase superfamily members, the 20s loop is disordered in the absence of ligand, and ordering of this loop upon substrate binding results in interactions with the ligand and shields the active site from solvent[13-18].

The enolase superfamily has been divided into subgroups based on sequence clustering and the identity and position of the catalytic residues[19]. Based on currently available sequences, the most functionally diverse subgroup is the muconate lactonizing enzyme (MLE) subgroup, which includes enzymes catalyzing cycloisomerization (MLE), β-elimination (*o*-succinylbenzoate synthase), racemization (L-Ala-D/L-Glu epimerase), and probably other uncharacterized functions.

Understanding protein evolution is complicated by difficulties in assigning functions to superfamily members. Categorizing superfamily members into families, or groups of proteins sharing the same function, is often accomplished by establishing a sequence similarity threshhold[20-24]. However, families in the enolase superfamily, as in other superfamilies, have most likely diverged at different rates or at different times during evolutionary history, making it difficult to define a similarity score cutoff that separates different isofunctional families. The *o*-succinylbenzoate synthase (OSBS) family poses a particularly thorny problem. First, sequence similarity between some OSBSs barely exceeds random similarity scores expected between unrelated proteins, making it impossible to define a similarity score that encompasses all OSBSs but excludes proteins of other functions. Second, a promiscuous protein from *Amycolatopsis* sp. T-1-60 that shares 42% identity with the OSBS from *Bacillus subtilis* catalyzes both OSB synthesis and *N*-acylamino acid racemization[25]. Even experimental characterization does not adequately determine the physiological function of this enzyme, since it catalyzes OSB synthesis and racemization of *N*-succinylphenylglycine at equivalent rates[26]. Thus, the OSBS/N-acylamino acid racemase (NAAAR) family is an especially interesting subject for investigating protein evolution because it includes both extremely divergent enzymes having the same function and very similar enzymes having different functions.

In this paper we have studied the evolution of the OSBS/NAAAR family. This study begins to answer several questions about how function and structure evolve in extremely divergent protein families. First, what sequence and structural features must be conserved to maintain function in extremely divergent families? Second, by what mechanisms do proteins evolve new functions? And finally, what functional and structural characteristics of a protein make it more or less capable of evolving a new function? Our study of the OSBS/NAAAR family's evolution demonstrates that sequence, structure, and modes of substrate binding are surprisingly malleable. In addition, we have identified a number of proteins of unknown function whose experimental characterization would be valuable for understanding evolutionary relationships and structural determinants of catalysis in the enolase superfamily. We also determined that the accuracy and extent of functional annotation could be improved using rigorous phylogenetic reconstruction accompanied by analysis of genomic context. Lastly, our in depth analysis of the evolution, structure and function of the

OSBS/NAAAR family identified several characteristics of *Amycolatopsis* OSBS/NAAAR which might enhance its evolvability relative to other OSBSs.

Results

Identification of OSBS enzymes

To understand the evolution of the OSBS/NAAAR family, we began by identifying species which must have OSBS activity. OSBS is an intermediate in the menaquinone (Vitamin K_2) biosynthesis pathway, which is essential for electron transport in a wide variety of prokaryotes[27]. Because characterized OSBSs are highly divergent, the presence of proteins catalyzing other steps in the menaquinone pathway was used as a marker for species encoding OSBS (Figure 1). First, proteins sharing > 40% identity with a characterized menaquinone pathway enzyme were annotated as having that function if the alignment covered > 90% of their lengths. This corresponds to BLAST E-values of $\sim10^{-40} - 10^{-95}$ using the NCBI non-redundant (nr) protein database. Although using this threshold is expected to produce some error[23,24], homologs of most menaquinone pathway proteins could only be identified in a few of the species which are known to produce menaquinone[28], suggesting that this threshold is fairly stringent. However, the menB protein, the most highly conserved protein in the pathway (average percent identity of 58%), could be identified in many genomes and served as a marker to identify species likely to encode the menaquinone operon. More distantly related menaquinone pathway proteins were identified by sequence similarity (BLAST E-values $< 10^{-20}$ relative to a reliably annotated menaquinone pathway protein using the nr database) and proximity to other menaquinone pathway genes (≤ 5 non-pathway genes intervening between a pair of menaquinone pathway genes). The combination of sequence similarity and genomic context is expected to identify orthologs, because consecutive enzymes in a pathway are rarely recruited together to function in a different pathway[29]. An important exception is that menB and menE both have homologs in carnitine metabolism; however, the homologs of menE in carnitine metabolism were too divergent to meet our criteria.

Using these criteria, we identified 127 strains in which at least five of the eight menaquinone pathway genes could be identified (Figure 2, Supplemental Figure 1). In organisms in which most menaquinone pathway genes were identified, some or all are colocalized in the genome and are likely to be coregulated as operons. Gene order is fairly well conserved among the γ-Proteobacteria (the phylum which includes *E. coli*), Bacteroidetes, and Firmicutes (the phylum which includes *B. subtilis*). In particular, the *menF* and *menD* genes, whose proteins catalyze the first three steps of the pathway, are adjacent in most species within these groups. Among some Cyanobacteria, the menaquinone operon has been almost completely fragmented. Gene order in the Actinobacteria is the least similar to other organisms, and most menaquinone pathway genes are separated by intervening genes; thus it is unclear whether they are coregulated.

In a number of genomes, only one or two possible menaquinone biosynthesis proteins could be identified. UbiE, a methyltransferase which functions in both menaquinone and ubiquinone synthesis, was identified in many genomes. Of species in which other homologs of menaquinone pathway proteins were found, four were in draft genomes which might encode the menaquinone pathway, and in six a close homolog of menB was found which may be more likely to function in benzoate degradation or fatty acid metabolism. This suggestion is based on the observation that no other menaquinone pathway proteins can be identified in these genomes, adjacent genes are annotated as being in these pathways, and homologs of menB which are difficult to distinguish from menB have been characterized and shown to function in these pathways. The remaining strains encode OSBS homologs, but no other menaquinone pathway proteins could be identified. These OSBS homologs share > 40% identity with *B. subtilis* OSBS or *Amycolatopsis* NAAAR, raising interesting questions about the evolution of new enzyme functions, as discussed below.

In species in which most but not all of the menaquinone pathway proteins were found, the most difficult proteins to identify were OSBS, menH (which is not fully characterized), and ubiE, which catalyzes the final step in the pathway and might not be required in all species[30]. As expected from previous work, OSBS is not well conserved, and its gene was difficult to identify in many species, as shown by hollow arrows in Figure 2. However, possible OSBSs could be identified in all finished and most unfinished genomes in which the menaquinone pathway was identified. In most genomes, a gene encoding an enolase superfamily member is adjacent to a menaquinone pathway gene and is likely to encode OSBS; however, outside of the Firmicutes or γ-Proteobacteria, the sequence similarity of the putative OSBS rarely met our criteria for annotation. Also, in a few genomes there was no OSBS candidate near a menaquinone pathway gene, but a gene encoding an MLE subgroup enzyme of unknown function could be identified elsewhere in the genome. The only genomes in which an OSBS candidate gene could not be identified were unfinished (*Bacillus cereus* ATCC 14579, *Haemophilus influenzae* 86028NP, and *Salmonella enteritidis*); as these species are closely related to *E. coli* or *B. subtilis*, it will be surprising if they do not encode an *E. coli*- or *B. subtilis*-like OSBS, respectively.

Phylogeny of the OSBS/NAAAR family

The difficulty of unequivocally identifying OSBSs based on sequence similarity and genome context is in agreement with the observation of Palmer et al. that OSBSs are extremely divergent and can share < 15% identity[25]. In fact, some putative OSBSs are barely recognizable as enolase superfamily members. For instance, sequence similarity searches using the OSBS from *Bdellovibrio bacteriovorus* as a query identifies another very divergent, putative OSBS as the best match, but the E-value (0.05) is barely significant. Thus, we speculated that OSBS activity might have evolved multiple times within the enolase superfamily. To investigate this hypothesis and to understand how the NAAAR-like proteins from organisms lacking menaquinone are related to OSBS, we examined the phylogeny of a subset of the enolase superfamily comprised of 288 sequences which includes all OSBS candidates, the rest of the MLE subgroup, and any other enolase superfamily members which could not be assigned to a subgroup or family by Hidden Markov Models (HMMs) created to describe OSBS and other enolase superfamily members in the Structure-Function Linkage Database (SFLD)[31,32]. Contrary to our hypothesis, the phylogenetic tree of a representative subset of these sequences demonstrated that all OSBSs and NAAAR-like proteins are included in a single clade (Figure 3)[*]. Although the resolution at many interior nodes is low, the branch confidence

[*] Phylogenetic trees for the MLE subgroup and the OSBS/NAAAR family were also constructed using only the capping or barrel domain (data not shown). For the whole

value separating the OSBS/NAAAR family from the rest of the MLE subgroup is 1.00. This result confirms that the OSBSs identified by sequence similarity and genomic context, including those that are too divergent to match the MLE subgroup HMM and those that are not encoded near other menaquinone pathway genes, belong to the OSBS/NAAAR family. In addition, this result strongly suggests that this family had a single evolutionary origin, because rooting the tree with MLE or AEE, the closest known paralogs of the OSBS/NAAAR family[19], leaves the family as a monophyletic group.

The other characterized proteins included in the MLE subgroup phylogenetic tree are MLE and L-Ala-D/L-Glu epimerase (AEE). The characterized AEEs from *B. subtilis* (aee.Bacsu) and *E. coli* (aee.Escco) are part of a large clade encompassing proteins from a diverse set of species, suggesting that these proteins are all AEEs. However, because the branch support for this clade is not high (0.76) and the genomic context of these proteins has not been thoroughly examined, determining their functions requires more study. Finally, a number of proteins on the MLE subgroup tree do not cluster with the

MLE subgroup, trees built using only the barrel domain were nearly identical to trees built using the entire protein, although the resolution was somewhat lower. In contrast, trees built using the capping domain, which is shorter and more divergent than the barrel domain, were highly multifurcating. For the OSBS/NAAAR family, trees constructed using only the capping or barrel domain were nearly identical to trees built using the entire protein, although the resolution was somewhat lower. This data is consistent with the notion that domain shuffling is not likely to have occurred among different members of the MLE subgroup or OSBS/NAAAR family, although it cannot be completely ruled out.

OSBS/NAAAR, MLE, or AEE families, suggesting that there are several more catalytic activities within the MLE subgroup remaining to be discovered.

If the OSBS/NAAAR family has a single evolutionary origin (i.e. all family members are orthologous), we expected its phylogenetic tree to be similar to trees built using other proteins or methods. The main difficulty with comparing phylogenetic trees encompassing all menaquinone-producing organisms is that it has not been possible to generate fully resolved prokaryotic evolutionary trees because of extensive lateral gene transfer, variable evolutionary rates, mutational saturation, and other factors that limit the statistical consistency and resolving power of phylogenetic mehods[33-36]. However, comparisons of more well-resolved branches might provide insight into whether lateral gene transfer or inclusion of paralogous proteins contribute to differences between the OSBS/NAAAR family tree and other trees.

We compared the phylogeny of the OSBS/NAAAR family (Figure 4A) to those of the menB and enolase families, which are much more highly conserved (Figure 4B, Supplemental Figure 2). In spite of the greater divergence of the OSBS/NAAAR family, all three trees had similar topologies and resolution. With a few exceptions, well-resolved branches are in agreement with published prokaryotic phylogenies, but higher level clustering of phyla (such as the reported *Deinococcus*/Cyanobacteria/Actinobacteria group) is absent, which is not unexpected since these groups become apparent only when multiple genes or genome characteristics are used for tree construction[37-43].

While certain differences among the OSBS/NAAAR, enolase, menB, and species trees might be artifacts of phylogenetic reconstruction, the unusual phylogenetic positions of some proteins appear to have more biologically interesting explanations. For instance, the δ-Proteobacteria *Desulfotalea psychrophila* (osbs.Desps) groups with Bacteroidetes in both the OSBS/NAAAR and menB trees, but not in the enolase tree (Supplemental Figure 2). Inspection of the sequences demonstrates that the *D. psychrophila* OSBS and menB are much more similar to the Bacteroidetes proteins (percent identity > 40% and > 75%, respectively) than Proteobacteria (percent identity ≤ 26% and ≤ 62%, respectively). Thus, *D. psychrophila* OSBS and menB appear to be correctly positioned on the phylogenetic trees, suggesting that the menaquinone operons of *D. psychrophila* and Bacteroidetes are related by lateral transfer. Another unusual feature of the OSBS/NAAAR tree is that the Archaea do not cluster together. In fact, the only Archaea that have a menaquinone operon are the two Halobacteria (osbs.Halma and osbs.Hal). Although they cluster with the Actinobacteria in both the OSBS/NAAAR and menB trees, the menaquinone operon structures of the two groups are very different; thus, it is possible that clustering of the Actinobacteria and Halobacteria is an artifact of phylogenetic reconstruction. In either case, it is likely that the Halobacteria attained the menaquinone operon by lateral transfer, since no other archaeon has this pathway.

The most striking feature of the OSBS/NAAAR tree is the placement and taxonomic distribution of the NAAAR-like proteins. This cluster of proteins encompasses not only the Firmicute OSBSs, but also proteins from a number of taxonomic groups, including Deinococcus-Thermus, Actinobacteria, Cyanobacteria, and Archaea. Because the resolution of this part of the tree is low, we constructed a phylogenetic tree using only sequences in this group, hoping to see a clear distinction between OSBS and NAAAR (Figure 5). The topology of the Firmicute OSBS/NAAAR subfamily tree differs slightly from its topology in the whole OSBS/NAAAR tree, but the branch confidence values are higher, suggesting that this tree might be a better representation of the subfamily's phylogeny.

The Firmicute OSBS/NAAAR subfamily tree contains several surprises. First, a few species have a NAAAR-like protein but do not appear to encode the other menaquinone pathway proteins, suggesting that the NAAAR-like proteins have a physiological role distinct from menaquinone synthesis in these species. Second, *Erwinia carotovora* (unk.Erwca and osbs.Erwca) and *Thermobifida fusca* (unk.Thefu and osbs.Thefu) encode both a NAAAR-like protein (which is not encoded in the menaquinone operon) and an OSBS (which is encoded in the menaquinone operon and, unlike their NAAAR-like proteins, clusters with OSBSs of species in the same phyla as these two organisms). This also suggests that these NAAAR-like proteins have a physiological function distinct from OSBS. Third, several species do not encode OSBS in their menaquinone operons but have a NAAAR-like protein encoded elsewhere. Conceivably, these function physiologically as OSBS, but they might have an additional function as well. This seems particularly likely for *Oceanobacillus iheyensis* (unk.Oceih) and *Geobacillus kaustophilus* (unk.Geoka), which cluster with *E. carotovora* NAAAR. Likewise, the NAAAR-like proteins from *Crocosphaera watsonii* (unk.Crowa) and *Chloroflexus aurantiacus* (unk.Chlau), which are Cyanobacteria and Chloroflexi, respectively, cluster most closely with NAAARs from species that appear to lack the menaquinone pathway (albeit with mediocre branch confidence values), suggesting that these NAAAR-like proteins are required for their OSBS activity and have replaced the original Cyanobacteria or Chloroflexi OSBS.

Given the high sequence similarity of the NAAAR-like proteins (most share > 40% identity with *B. subtilis* OSBS) and the fact that they are found in very distantly related species, it seems likely that this protein has been transmitted by multiple lateral transfer events. For instance, the *E. carotovora* NAAAR might derive from an ancestor of *O. iheyensis* and *G. kaustophilus*, and there may have been separate transfer events to the two groups of Archaea—the euryarchaeote *Thermoplasma* clade (unk.Theac, unk.Thevo, unk.Ferac, and unk.Picto) and the crenarchaeote *Aeropyrum pernix* (unk.Aerpe). In summary, the phylogeny of the Firmicute OSBS/NAAAR subfamily does not clearly differentiate between apparently monofunctional OSBSs such as that of *B. subtilis* and promiscuous OSBS/NAAARs such as that of *Amycolatopsis.* The presence of NAAAR-like proteins in species lacking the menaquinone pathway suggests that they have an unknown function, perhaps amino acid racemization.

Diversity of the OSBS/NAAAR family

Having performed a comprehensive survey of the distribution of the OSBS/NAAAR family, we were interested in reevaluating the family's diversity to discover whether it is unusually divergent compared to other protein families, as suggested previously[25]. Initially, we compared lengths of OSBS/NAAAR family trees to tree lengths of other families in the menaquinone pathway or enolase superfamily. Tree length (measured as substitutions per site) is expected to be the most accurate measure of sequence divergence, because it corrects for multiple substitutions per site. In comparisons of trees built using sequences from the same set of species, the length of OSBS/NAAAR trees were usually at least twice as long as those of other protein families, indicating that the OSBS/NAAAR family has indeed evolved at a much faster rate (data not shown). However, the topology of the OSBS/NAAAR tree was similar but rarely identical to the topology of trees built using other families, even when using subsets of the OSBS/NAAAR family that are well resolved on the phylogenetic tree.

Because the significance of comparing lengths of trees that have different topologies is uncertain, we also calculated pairwise percent sequence identities, even though these are a more approximate measure of evolutionary distance. Comparison of OSBSs and menBs from a wide taxonomic distribution agree well with those previously reported, with menB proteins generally sharing > 40% identity while OSBSs from the same set of species generally share < 30% identity[25]. To gain a better perspective concerning the divergence of the OSBS family, we compared minimum and average percent identities of the OSBS family to other families in the enolase superfamily and menaquinone pathway (Table 1, Figure 1). For each comparison, the set of OSBSs and the set of proteins from the compared family were taken from the same set of species. Compared to other families in the enolase superfamily, the OSBS family is unusually divergent. However, comparison to other proteins in the menaquinone pathway reveals a different picture. Although MenB is extremely well-conserved, the sequence divergence of MenD and MenE is more similar to OSBS. On average, the OSBS family is slightly more divergent than the MenD or MenE families, but because percent identity is only a rough approximation of evolutionary distance, it is unclear whether the OSBS family is significantly more divergent than these proteins. Thus, although the OSBS family is unusually divergent for the enolase superfamily, it is less extraordinary compared to other proteins in its pathway.

In addition to being more divergent than other families in the enolase superfamily, the OSBS/NAAAR family is unusual in that it includes proteins catalyzing at least two different reactions. Surprisingly, the NAAAR-like proteins are not among the more divergent proteins in the family, but are closely related to proteins identified as OSBS based on genomic context and experimental evidence[25]. As shown above, phylogenetic analysis failed to separate the NAAAR-like proteins into a separate clade. In fact, most NAAAR-like proteins which are not encoded in menaquinone operons share > 40% identity with *B. subtilis* OSBS. Only the genomic position of the genes encoding NAAAR-like proteins hints that their function might differ from the menaquinone operon-encoded OSBSs.

Conservation of sequence and structure in the OSBS/NAAAR family

Despite the high sequence divergence of the OSBS/NAAAR family, all proteins in the family form a single clade in the MLE subgroup phylogenetic tree, indicating that there must be conserved sequence information that differentiates this family from the rest of the MLE subgroup. To identify conserved residues specific to the OSBS/NAAAR family, we compared the pattern of sequence conservation among the OSBS/NAAAR, MLE, and AEE families. For this analysis, the OSBS/NAAAR family was treated as a single unit or divided into subfamilies representing clades containing at least five sequences (γ-Proteobacteria, Cyanobacteria, Bacteroidetes, Actinobacteria, and Firmicutes/NAAAR-like proteins), as indicated in Figure 4A. Except for unk.Thefu (gi23018694 from *Thermobifida fusca*, discussed below), the NAAAR-like proteins were included with the Firmicute OSBSs because they could not be cleanly separated based on phylogeny or the presence of the menaquinone operon. In addition, the AEEs were divided into two groups comprised of close relatives of characterized *E. coli* or *B. subtilis* epimerases because the clade including both groups had poor statistical support on the MLE subgroup phylogenetic tree (Figure 3).

The pattern of sequence conservation is summarized in Figure 6, in which residues conserved in > 90% of subfamily members are highlighted in magenta, and residues conserved in both > 90% of the subfamily and > 90% of the entire MLE subgroup are highlighted in black. The only residues conserved throughout the entire MLE subgroup are the catalytic residues in the barrel domain, except for the lysine on barrel domain strand β6 (Bar-β6) which is replaced by tyrosine or arginine in some MLE subgroup members, including one branch of the Cyanobacteria OSBS subfamily. For these Cyanobacteria OSBSs, an arginine at this position might have little effect on catalysis, because the lysine at this position in *E. coli* OSBS appears to stabilize the enediolate intermediate rather than act as a general acid/base catalyst[44]. The other highly conserved residues in the MLE subgroup appear to be involved in maintaining the structure. For instance, the conserved elements of capping domain strand β3 and helix α3 (Cap-β3 and Cap-α3) are adjacent and probably important for capping domain structure, and the glycine before Bar-β6 is located in a tight turn. Other than these residues, the pattern of sequence conservation is somewhat variable. Although some groups appear to have greater numbers of conserved residues, this is mostly because these groups are small (e.g. the Bacteroidetes group) or include sequences of limited diversity (e.g. MLE and AEE groups, in which sequences share > 40% identity). In comparison, the Firmicutes/NAAAR-like subfamily includes more divergent sequences; it should be noted that the most divergent sequences in this group (osbs.Staau, osbs.Staep, osbs.Lacla, osbs.Desha, osbs.Leume, and osbs.Exi) are menaquinone operon-encoded OSBSs, not NAAAR-like proteins.

Surprisingly, the results of this analysis indicate that there are no conserved residues shared by all five OSBS/NAAAR subfamilies, other than residues also shared with the rest of the MLE subgroup. Conserved residues within subfamilies are most likely to fall in regions near the active site, either on two loops of the capping domain or on the strands or loops of the barrel domain. Although one or more OSBS/NAAAR subfamilies often has conserved residues at the same position, the identities of those residues are rarely the same. In cases where the residue identity is conserved, the same residue is often present in the MLE or AEE families. Thus, although the OSBS/NAAAR family is phylogenetically unified and most, if not all (including characterized NAAAR-like proteins) catalyze the OSBS reaction, there are no unique OSBS/NAAAR family motifs to differentiate them from other MLE subgroup members.

To understand how substrate specificity is conserved with so little sequence conservation, we compared the structures of *E. coli* OSBS bound to the substrate or OSB (1FHV and 1R6W), *Amycolatopsis* OSBS/NAAAR bound to OSB (1SJB), and *B. bacteriovorus* OSBS bound to OSB (coordinates generously provided by Alexander Fedorov, Elena Fedorov and Dr. Steven Almo, Albert Einstein College of Medicine)[17,44,45]. In all three structures, residues lining the active site pocket are in homologous positions, and these residues tend to be more highly conserved within and between subfamilies than regions distant from the active site (Figure 6). The structures exhibit similar hydrophobic interactions between the benzene ring of OSB and the 50s loop, in which at least one of the residues interacting with ligand is aromatic. Most members of the OSBS/NAAAR family (and many other members of the MLE subgroup) have aromatic residues at one or both positions, suggesting that this hydrophobic pocket is important for ligand binding.

In contrast to these similarities, there are also some striking differences in active site structure, which might contribute to differences in function and inherent evolvability. As previously reported, the conformation of OSB differs in the *Amycolatopsis* and *E. coli* enzymes[45]. In *Amycolatopsis*, the succinyl tail of OSB is extended, while it is bent in *E. coli* and *B. bacteriovorus* (Figure 7A). Likewise, the succinyl or acetyl moieties of *N*-acylamino acid substrates also lie in extended conformations in *Amycolatopsis* OSBS/NAAAR. For *N*-succinyl-methionine, this conformation provides suitable hydrogen bond donors and acceptors, which are unavailable in *E. coli* OSBS, accounting for the inability of *E. coli* OSBS to racemize this substrate[45].

The second major difference among these structures is the position of the 20s loop (Figure 7B, top). In spite of its proximity to the active site, the 20s loop is poorly conserved within and between different subfamilies. The lack of conservation might be explained by the necessity of compensatory mutations to accommodate other structural changes, such as shifts in the orientation between the two domains, although there might also be consequences for the catalytic activity (see below). In *Amycolatopsis* OSBS/NAAAR bound to OSB, the 20s loop contacts the catalytic lysine that acts as a general base (the second lysine in the KXK motif), sandwiching it between the loop and the barrel and orienting it appropriately for proton abstraction. In contrast, the 20s loop of *E. coli* OSBS bound to either substrate or product does not contact the barrel, leaving the active site slightly open and the catalytic lysine disordered and solvent accessible. Similarly, the catalytic lysine is also solvent accessible in *B. bacteriovorus* OSBS, although the 20s loop is disordered, even when OSB is bound (data not shown).

In addition to comparing active site structure, we analyzed overall structural differences. Examination of all pairwise superpositions of the three members of the OSBS/NAAAR family and MLE I from *Pseudomonas putida* (1MUC) revealed significant structural differences (Figure 8, Tables 2,3). Intriguingly, *Amycolatopsis* OSBS/NAAAR is more similar to MLE I than it is to the other OSBSs. While this correlates with the central positions of MLE I and *Amycolatopsis* OSBS/NAAAR in the MLE subgroup phylogenetic tree (Figure 3), it is remarkable that structural differences are more pronounced between enzymes catalyzing the same reaction than between those catalyzing different reactions. One obvious structural difference among the enzymes is that the capping domains of the OSBSs are poorly aligned. To investigate this difference further, the capping and barrel domains were aligned separately. This resulted in better alignments of both domains, in which structural differences tend to be located at the surfaces of the proteins (Figure 8). Thus, orientation between the domains differs among the OSBS/NAAAR enzymes.

To quantify these differences in domain orientation, we measured the angle of rotation of the capping domain between pairs of structures in which the barrel domain had been superposed. This was accomplished by using sets of structurally aligned residues in the capping domains to define planes representing capping domain orientation and measuring the dihedral angle between these planes. Because defining planes in this manner depends on which residues are used, the calculation was repeated several times with different sets of residues, revealing similar results (data not shown).

Table 3 shows differences in capping domain orientation between pairs of structures in which the capping or barrel domains are superposed. A slight rotation (3-4°) between the two domains is observed when comparing structures of *E. coli* and possibly *B. bacteriovorus* OSBSs with and without ligand. A somewhat higher degree of rotation is observed between *Amycolatopsis* OSBS/NAAAR and MLE I and between *B. bacteriovorus* OSBS and MLE I. Because no ligand is bound to MLE I, rotations of this magnitude might reflect conformational differences due to ligand binding as well as slight structural differences between different proteins. In the remaining comparisons, the rotation of the capping domain is significantly higher than that observed for liganded versus unliganded structures of the same protein. In particular, the capping domain of *E. coli* OSBS is rotated 13.3° or 17.7° relative to those of MLE I and *Amycolatopsis* OSBS/NAAAR, respectively. We hypothesize that these structural differences might contribute to differences in binding specificity and catalysis among these enzymes, as well as to their capacities to evolve new functions, as discussed below.

In order to understand the consequences of domain orientation on the structure of the active site and the function of the enzymes, we analyzed the effect of twisting the *E. coli* OSBS capping domain to match the orientation of the *Amycolatopsis* OSBS/NAAAR capping domain (Figure 7B, bottom). To do this, the capping and barrel domains were superimposed separately on the *Amycolatopsis* enzyme. Twisting the *E. coli* capping domain shifts the 20s and 50s loops ~ 6 Å down toward Bar-β2. As a result, the 20s loop is no longer in contact with the ligand. Instead, it now approaches the catalytic lysine of the KXK motif, which is disordered in the *E. coli* structures. Having the 20s loop in this position would prevent this lysine from adopting an extended conformation, possibly forcing it into the active site toward the substrate. When the converse experiment is performed and the *Amycolatopsis* capping domain is twisted to match that of *E. coli*, the 20s and 50s loops shift ~ 6 Å away from the barrel so that the 50s loop is no longer in contact with the ligand. In this position, the 20s loop barely contacts the second lysine of the KXK motif, leaving it mostly exposed to solvent outside the active site.

Although we have only shifted the orientations of the two domains and have not refined the models to ameliorate steric hindrances or reposition loop residues into more favorable conformations, these results suggest that proper orientation of the capping and barrel domains is required for positioning the catalytic lysine for catalysis in *Amycolatopsis* OSBS/NAAAR. For *E. coli* OSBS, these results suggest two possibilities. First, perhaps the flexible lysine is resident in the active site often or long enough for catalysis. Second, it is also conceivable that the crystal structures of *E. coli* OSBS bound to either substrate or product do not capture the structure of the enzyme in the transition state. As in *Amycolatopsis* OSBS/NAAAR, repositioning the 20s loop through domain rotation or other conformation changes might be required in order to correctly position the lysine for catalysis. The fact that the 20s loop is disordered in *B. bacteriovorus* OSBS in the presence of ligand provides some support for the latter possibility.

Discussion

Changes in protein structure during evolution

Investigating the evolutionary relationships among the OSBS and NAAAR-like proteins of the enolase superfamily uncovered several surprising observations. The most remarkable are that these proteins exhibit significant structural variation and that sequence motifs unique to the OSBS/NAAAR family which distinguish it from other families in the enolase superfamily could not be identified, in spite of the fact that OSBS activity has been conserved and the family appears to have a single evolutionary origin.

This raises the question of how enzyme specificity can be maintained over the course of evolution. Some structural differences would be expected between *Amycolatopsis* OSBS/NAAAR and the other two OSBSs, since the *Amycolatopsis* enzyme has an additional activity. However, structural differences as exemplified by both RMSD and domain orientation are at least as great between *E. coli* and *B. bacteriovorus* OSBSs. One way in which specificity might be maintained during evolution is through compensatory mutations and structural flexibility of surface loops that close the active site[11]. In the three OSBS/NAAAR family structures, the function of the 50s loop appears to be conserved, since it is structurally well-aligned and forms a hydrophobic binding pocket for the benzene ring (Figure 6). The ring is anchored at one end by the carboxyl group binding to the metal ion and by the 50s loop at the other. Mutations that affect the orientation of the benzene ring could be accommodated by structural reorganization and mutations of the 50s loop, such as the small insertion observed in the *Amycolatopsis* enzyme.

The 20s loop is also likely to play an important role in maintaining, and perhaps altering enzyme specificity. In most enolase superfamily members, this loop is disordered in the absence of ligand[13-18]. In addition to being less well-conserved than the 50s loop, the 20s loop is not well-aligned in the structures of *Amycolatopsis* OSBS/NAAAR and *E. coli* OSBS bound to OSB, and it is disordered in *B. bacteriovorus* OSBS bound to OSB. The flexibility and apparent mutability of this loop suggest that it could have coevolved with other sequence and structure elements (such as those determining domain orientation) to maintain substrate binding. In addition, the flexibility of this loop might allow promiscuous binding and reactions with new substrates without impairing OSBS activity, leading to the evolution of new protein functions, such as NAAAR activity[10,11].

While the role of flexible loops in maintaining OSBS activity is somewhat speculative, it has also been proposed that structural requirements for catalysis are relatively permissive because the OSBS reaction is highly exergonic and can proceed uncatalyzed at significant rates[25,46]. In all three OSBS/NAAAR family structures, interactions with OSB are largely hydrophobic, and most hydrogen bonds are formed with water or residues conserved in the whole MLE subgroup (Alexander Fedorov, Elena Fedorov and Dr. Steven Almo, unpublished)[17,45]. Thus, it appears that interaction with subgroup-conserved residues is sufficient for correctly orienting the substrate for catalysis, and the only additional requirement is a hydrophobic cavity of an appropriate size and shape. Additional evidence for this is supplied by single point mutations in *Pseudomonas* sp. P51 MLE II and *E. coli* AEE which confer OSBS activity on these enzymes[6]. These mutations are located at the same position in Bar-β8 and exchange an aspartate or glutamate for a glycine, creating space to accommodate the succinyl tail of OSB if it is bound in the same conformation as in *E. coli* OSBS (Figures 6 and 7A).

The differences in substrate binding and overall structure in only three of the divergent OSBS/NAAAR subfamilies raises the question of how many strategies for substrate binding there might be and whether or not there are additional promiscuous and perhaps biologically relevant activities catalyzed by members of this family. All OSBS/NAAAR subfamilies exhibit variation in length. The regions preceding and following Cap-α3 are especially variable throughout the family, and the entire region, including Cap-α3, is deleted from all members of the Actinobacteria subfamily, except *Tropheryma whipplei* (osbs.Trowh) (Figure 6, represented by *Mycobacterium tuberculosis*). In *Amycolatopsis* OSBS/NAAAR, these regions include helical sections and are at the oligomeric interface of the octamer[45]. In contrast, these regions are shorter and not helical in *E. coli* and *B. bacteriovorus* OSBSs, which are monomers (A.S. and J.A.G., unpublished data)[17]. Thus, much of the length variation in these enzymes appears to have altered their oligomeric structure.

The structural differences among members of the OSBS/NAAAR family appear surprisingly large, but it is unknown whether such differences are typical among orthologous proteins. Although a number of orthologous proteins from distantly related species have been structurally characterized, detailed structural comparisons have not always been performed. The structures of several enolases from both eukaryotes and prokaryotes have been solved, and they share much more similarity than the OSBS/NAAAR family proteins do, which correlates with their much higher sequence conservation (RMSD = 0.5 – 0.6 Å when 265 atoms are aligned between enolases from five different species; see Tables 1 and 2 for comparison)[13,47-51]. In contrast, the ribonucleotide reductase family is more mechanistically and structurally divergent than the OSBS/NAAAR family. Whereas the mechanism for OSBS synthesis has been conserved in the OSBS/NAAAR family, the ribonucleotide reductases can be divided into three classes which employ different means of radical generation, have slightly different substrate preferences, and share < 10% sequence identity even though they catalyze the same general reaction utilizing a conserved cysteine radical[52,53]. These proteins have a common 10-stranded α/β barrel, but each class has different insertions and deletions resulting in larger structural variations than observed in the OSBS/NAAAR family, which has limited numbers of insertions and a conserved bidomain structure. A more thorough analysis of allowable structural variation in other protein families will be required to determine to what degree structural variations reflect functional divergence in different protein families.

Changes in protein function during evolution

In addition to the surprising structural variations in the OSBS/NAAAR family, we have discovered that at least one new function has apparently evolved within the family and has been transmitted by lateral transfer to diverse species (see below). Recently, we have begun characterizing other NAAAR-like proteins and have discovered that those from *Deinococcus radiodurans*, *Thermus thermophilus*, and *Geobacillus kaustophilus* are also promiscuous (A.S. and J.A.G., unpublished results). In addition, the genes for these proteins are adjacent to succinyltransferase genes, suggesting that racemization of *N*-succinylamino acids is their biological function[54]. While *D. radiodurans* and *T. thermophilus* do not appear to have the menaquinone pathway, a gene encoding OSBS is missing from *G. kaustophilus*'s menaquinone operon, suggesting that its NAAAR-like protein might also function in the menaquinone pathway. Other NAAAR-like proteins found in organisms whose menaquinone operons are missing an OSBS gene might also have two biologically relevant activities. Some of these, such as the NAAAR-like proteins from *E. carotovora* and *O. iheyensis*, are found in species that appear to lack succinyltransferases, suggesting that they might function only as OSBSs or have additional, unknown functions.

The identification of a whole set of related, promiscuous proteins strongly supports the role of promiscuity in protein evolution and the idea that new activities can evolve prior to gene duplication[1-10]. The fact that several NAAAR-like proteins are promiscuous and the strong statistical support for their position in the phylogenetic tree support a scenario in which racemization activity arose in a Firmicute ancestor. The converse hypothesis that OSBS activity evolved in an ancestral racemase seems less likely given that OSBS is more widespread, more divergent and plays an essential metabolic role in many prokaryotes. The possibility that the ancestor of the entire OSBS/NAAAR family was promiscuous for the two activities cannot be ruled out, however.

Although sequence and phylogenetic analysis could not separate most NAAAR-like proteins from operon-encoded OSBSs, one sequence does stand out. This protein (unk.Thefu, gi23018694) is found in the Actinobacterium *T. fusca*, which has both this NAAAR-like protein and an operon-encoded Actinobacteria-like OSBS. What makes this protein unique is that, although it shares 35% identity with *Amycolatopsis* OSBS/NAAAR, its catalytic residues differ from all other members of the MLE subgroup. Instead of the conserved KXK motif on Bar-β2, this protein has RLH; DGG replaces DXN on Bar-β3, and there is an arginine instead of the conserved lysine on Bar-β6. Because arginine is expected to be a poor general base, it seems unlikely that this enzyme is a racemase, which would require a base on both sides of the active site. The presence of histidine instead of lysine on Bar-β2 also suggests that this protein might have a different activity. Thus, it appears that at least three activities may have evolved within the OSBS/NAAAR family.

Evolvability of the NAAAR-like proteins

Consideration of the structural and functional variation between the Firmicute/NAAAR-like subfamily and other apparently monofunctional OSBSs prompts the question of what structural differences contribute to the ostensible functional evolvability of the Firmicute/NAAAR-like subfamily. While it is possible that any of the other OSBSs might have uncharacterized, promiscuous activities, one or more new activities appear to have evolved in the Firmicute/NAAAR-like subfamily. In addition, the stronger structural similarity between *Amycolatopsis* OSBS/NAAAR and MLE compared to *E. coli* OSBS raises the possibility that the structural configuration of the *Amycolatopsis* enzyme could be more suitable for evolving new functions. Alternatively, its similarity to MLE might also reflect similarities in their quaternary structures, since both *Amycolatopsis* OSBS/NAAAR and MLE are octamers, while *E. coli* and *B. bacteriovorus* OSBSs are monomers (A.S. and J.A.G., unpublished data)[17,45,55].

If its higher structural similarity to proteins outside the OSBS/NAAAR family reflects its suitability as a scaffold for evolving new functions, what structural features of *Amycolatopsis* OSBS/NAAAR might facilitate this? The major structural difference between *Amycolatopsis* OSBS/NAAAR and *E. coli* OSBS is the orientation of the capping domain and its effect on the position of the 20s loop. Whereas the 20s loop contacts and orients the catalytic lysine in *Amycolatopsis* OSBS/NAAAR, it is more distant from the ligand in *E. coli* OSBS, leaving the catalytic lysine disordered and exposed to solvent. This structural difference might affect the kinetic constants of the two enzymes. The k_{cat} of *Amycolatopsis* OSBS/NAAAR is 10-fold higher than that of *E. coli* OSBS, possibly because the catalytic lysine is held in a more appropriate position. However, k_{cat}/K_m of the *Amycolatopsis* enzyme is 6-fold lower[25]. Inasmuch as K_m reflects the strength of substrate binding, this suggests that substrate affinity is as much as 40-fold higher for *E. coli* OSBS. Thus, *E. coli* OSBS might be evolutionarily optimized to maximize the strength of substrate binding; mutations that alter substrate binding might be more deleterious relative to similar mutations in the *Amycolatopsis* enzyme, rendering *E. coli* OSBS less likely to gain promiscuous functions during evolution and therefore less likely to be a source for novel enzymatic activities. In contrast, the position of the 20s loop in *Amycolatopsis* OSBS/NAAAR might be optimized to maximize k_{cat} by clamping down on the catalytic lysine to hold it in position and more effectively close the active site. This could result in relaxation of constraints on substrate binding such that decreases in substrate affinity would remain within physiological tolerances. As a result, the interior size and shape of the active site could evolve to allow promiscuous binding and catalysis, leading to the evolution of new protein functions. Thus, the relatively minor alterations in the size, shape and potential hydrogen bond donors and acceptors between *Amycolatopsis* OSBS/NAAAR and *E. coli* OSBS may not have been sufficient to allow productive binding and catalysis of *N*-acylamino acids in the *Amycolatopsis* enzyme by themselves. Instead, other aspects of the protein structure including domain orientation and the position of the 20s loop may have also been required in order to produce an “evolvable” environment which could tolerate these mutations. This argument is speculative and assumes that the conformation of the capping domain in the *E. coli* OSBS structures is not an artifact of crystallization, but reflects either the catalytically competent form of the enzyme or a more stable form of the enzyme which requires a conformation change to reach the transition state, resulting in a lower apparent k_{cat} relative to that of *Amycolatopsis* OSBS/NAAAR. It might be possible to test this hypothesis by experimental evolution and by identifying mutations that affect substrate binding or the position of the 20s loop and studying their effects on catalysis.

Although some of the characteristics suggesting that the Firmicute/NAAAR subfamily might be particularly evolvable are specific to this subfamily, a number of them may be more generally applicable to other protein families. First, more highly evolvable proteins are expected to share more similarities with functionally different proteins in their superfamily than less evolvable proteins, as manifested by a central location in their superfamily’s phylogenetic tree and structural similarities with functionally different proteins in their superfamily. Second, there might be enzymatic traits such as highly optimized k_{cat}s, which enhance the likelihood that highly evolvable proteins are promiscuous. Examining these characteristics in the OSBS/NAAAR family as well as expanding these ideas to other superfamilies will be necessary to test these hypotheses. In addition, such detailed analysis of structure-function relationships might be extremely valuable for identifying scaffolds that are particularly amenable to protein engineering[56,57].

Lateral gene transfer in the OSBS/NAAAR family

Phylogenetic comparison of the OSBS/NAAAR, menB, and enolase families revealed several probable instances of lateral gene transfer. Lateral gene transfer is a major driving force of prokaryotic genome evolution, accounting for the origin of as much as 15% of the genes in some species[58-63]. Evidence for lateral gene transfer has been inferred from nucleotide composition, codon bias, unusual species distributions of genes, sequence similarity, and phylogenetic analysis, which is considered the most robust method[61,62]. Although extensive statistical comparisons of phylogenetic trees are beyond the scope of this paper, several instances of lateral gene transfer in the OSBS/NAAAR family are supported by phylogeny, species distributions of genes, and sequence similarity. For example, the Halobacteria are the only Archaea which encode the menaquinone operon, and the gene order of the operon resembles that of menaquinone operons in other species (Figures 2, 4A). Lateral transfer of this operon in *Halobacterium* sp. NRC-1 (osbs.Hal) was previously detected by sequence similarity to bacterial proteins and anomalous nucleotide composition[64]. The menaquinone operon of the δ-Proteobacteria *D. psychrophila* also appears to have been the donor or recipient of lateral transfer, since both its OSBS and menB cluster with the Bacteroidetes.

The most compelling examples of lateral gene transfer are among the NAAAR-like proteins. These cluster with the Firmicute OSBSs but are found in extremely distant species. In addition, the Firmicute/NAAAR-like subfamily exhibits relatively high levels of sequence identity, averaging 36% identity (41% excluding OSBSs encoded in menaquinone operons), compared to 26% for the whole OSBS/NAAAR family. The NAAAR-like proteins are found in γ-Proteobacteria, Cyanobacteria, Actinobacteria, Deinococcus/Thermus, Chloroflexi, and Archaea, suggesting that this protein has been transferred multiple times. The fact that the archaeal sequences do not cluster together suggests that the proteins were transferred in separate events.

Characterized NAAAR-like proteins exhibit both NAAAR and OSBS activity (A.S. and J.A.G., unpublished results), and some of these might be required to perform both functions in vivo. For instance, the cyanobacterium *Crocosphaera watsonii* does not encode a cyanobacterial OSBS, but it does encode a NAAAR-like protein, which may be required for both NAAAR and OSBS activities and could have replaced the original cyanobacterial OSBS. Two other species, the actinobacterium *Thermobifida fusca* and the γ-proteobacterium *Erwinia carotovora,* have complete menaquinone operons encoding actinobacterial or γ-proteobacterial OSBSs, respectively, in addition to a NAAAR-like protein. While the NAAAR-like protein of *E. carotovora* might have OSBS activity, like other characterized NAAAR-like proteins, it is possible that the *T. fusca* NAAAR-like protein lacks OSBS activity because of mutations in the catalytic residues discussed above.

Rejecting lateral gene transfer as an explanation for these observations would imply that a NAAAR-like protein was present in the common ancestor of Archaea and Bacteria. Explaining the current distribution of this protein would require an enormous number of gene loss events (suggesting that the protein is not essential in most environments), in spite of the seemingly contradictory observation that the sequences are well-conserved (suggesting that selection has acted to maintain the protein's function). Thus, lateral gene transfer is the more parsimonious explanation[61].

Ramifications for structure and function prediction in genomics

Two important contributions of genomics are to correctly annotate protein functions and identify proteins of unknown structure and function whose characterization will enhance biological understanding. As noted previously and shown here, simple sequence metrics are often inadequate for predicting protein function[23,24]. Perusal of GenBank annotations of the OSBS/NAAAR family reveals that only 60% are correctly annotated (43% excluding proteins misleadingly annotated as "*o*-succinylbenzoate-CoA synthases")[†]. While only 7% of these annotations are completely incorrect, the remainder are incomplete or somewhat misleading, often assigning OSBS/NAAAR proteins to the wrong family or subgroup of the enolase superfamily. For example, several proteins are incorrectly annotated as muconate or chloromuconate cycloisomerases. Many others are annotated as "COG4948: L-alanine-DL-glutamate epimerase and related enzymes of enolase superfamily", which correctly relates them to the MLE subgroup but also implies an incorrect function.

Functional annotation of the OSBS/NAAAR family is difficult for two reasons. First, some members of the family are so divergent that sequence similarity cannot be used to distinguish them. Outliers such as the *B. bacteriovorus* OSBS could only be identified using a combination of genomic context, phylogenetic analyses, and ultimately experimental validation. Second, the NAAAR-like proteins could not be separated from the OSBSs based on sequence similarity or position in the phylogenetic tree. Instead,

[†] A detailed and systematic study of misannotation in the enolase and other superfamilies is currently underway in our laboratory, and the corrected annotations will be incorporated into the Structure-Function Linkage Database (SFLD)[32].

their main characteristics are that they are closely related to *Amycolatopsis* OSBS/NAAAR and they are not encoded in menaquinone operons.

Given such complexities, it is not surprising that automated annotation methods have had so much difficulty with this family. The orthogonal information furnished by phylogenetic reconstruction and analysis of genome context not only provides stronger confidence in functional annotation, but it is also invaluable for identifying proteins whose functions cannot be predicted with certainty. Similarly rigorous application of these methods will probably be required for accurate annotation of other protein families which exhibit high sequence, structural, and functional divergence.

Detailed studies of the sort undertaken here are also useful for identifying candidates for experimental characterization and structural genomics projects. Not only is there significant functional diversity in the OSBS/NAAAR family, but we also discovered significant structural variation among the family's three crystallized members. As discussed above, it is expected that several other subfamilies, especially the Actinobacteria subfamily, also exhibit structural variations. Solving the three-dimensional structures of representatives of other subfamilies will be valuable for understanding allowable variations in protein-substrate interactions in isofunctional proteins. In addition, our current and future studies of the structure and function of the NAAAR-like proteins will help elucidate how new protein functions evolve. Although our strategy is more labor-intensive than purely automated methods of target selection for structural genomics projects, it provides more context for understanding structure-function relationships and evolutionary mechanisms.

Concluding remarks

Our analysis of the OSBS/NAAAR family revealed several insights into how protein function and structure evolve. First, highly divergent protein families can exhibit significant structural variations. Second, enzyme specificity can be maintained in spite of limited sequence conservation among ligand-contacting residues. Third, new activities can evolve through promiscuous intermediates, and there might be structural features of proteins that make them more or less prone to evolve promiscuous activities. Few analyses of protein structure, function, and evolution have been performed in this depth; thus, extending these studies to other protein families will be important for testing the generality of these conclusions.

Materials and Methods

Identification of menaquinone pathway genes

Menaquinone biosynthesis genes were identified in complete and incomplete genomes using the Seed Annotation and Analysis Tool from the Fellowship for Interpretation of Genomes (FIG)[65]. Genes were initially annotated as menaquinone pathway genes if the percent identity of a pairwise protein alignment covering > 90% of the length of a characterized menaquinone pathway protein was > 40%. Experimentally characterized menaquinone pathway proteins include all pathway proteins from *E. coli*; menB, menC, menD, menE, and menF from *B. subtilis*; ubiE from *Geobacillus stearothermophilus*; and menA and menB from *Synechocystis* sp. PCC 6803 (Figure 1)[25,27,66-71]. As a second criterion, genes were annotated as encoding a menaquinone pathway protein if they were 5 or fewer genes distant from another menaquinone pathway gene and their proteins had BLAST expectation values $< 10^{-20}$ relative to reliably annotated menaquinone pathway proteins when searching the nr database. Most of the remaining genes were provisionally assigned functions if their proteins share ~25-40% identity with a characterized menaquinone pathway protein and nearly all proteins identified as being similar (BLAST E-values $< 10^{-5}$ using the nr database) are annotated as having that function.

Identification of MLE subgroup members

The initial enolase superfamily data set was downloaded from the Structure-Function Linkage Database (SFLD)[31,32]. Additional superfamily members were identified using a subset of the superfamily filtered to include only proteins sharing < 35% identity as input for Shotgun[72]. This program performs a BLAST search[73,74] of each input sequence and outputs a score indicating the number of input sequences that find a given BLAST hit, allowing homologs which have barely significant BLAST E-value scores to be identified. These sequences were then manually screened to remove fragments and to verify that they contained the canonical catalytic residues of the enolase superfamily. The final enolase superfamily data set was compared to HMMs from the SFLD to classify sequences into subgroups and isofunctional families. All further analyses were performed using protein sequences matching the MLE subgroup HMM with expectation values $< 10^{-18}$ and any other enolase superfamily sequences which could not be classified into a subgroup or family by the HMMs.

Phylogenetic analysis

The MLE subgroup and outlying enolase superfamily members were aligned using Muscle v.3.52[75]. The initial alignment was manually refined using structural alignments of muconate lactonizing enzyme (1MUC), L-Ala-D/L-Glu epimerase (1JPM and 1JPD), N-acylamino acid racemase (1SJB and 1XS2), and OSBS (1FHV and *B. bacteriovorus* OSBS). Structural alignments were generated by MinRMS[76] and the structure matching and alignment feature of UCSF Chimera from the Resource for Biocomputing, Visualization, and Informatics at the University of California, San Francisco (supported by NIH P41 RR-01081)[77]. Phylogenetic reconstruction was performed using Bayesian and distance methods. Bayesian trees were constructed with MrBayes v3.1.1[78,79] under the WAG amino acid substitution model[80] using a gamma distribution to approximate rate variation among sites. Distance trees were constructed using the NEIGHBOR program in PHYLIP[81] under the JTT amino acid substitution model[82] and a gamma distribution of rate variation among sites using the alpha parameter estimated in the Bayesian analysis. Trees produced by the two methods were similar, although the Bayesian method produced trees with higher resolution and branch confidence values. Accession numbers of sequences and species abbreviations used for phylogenetic analysis are listed in Supplementary Tables 1-4. In general, species names are abbreviated using the first three letters of the genus and first two letters of the species. The strain is indicated if multiple strains of the same species were used in the analysis, and *Bacteroides* is abbreviated with "Bct" to avoid confusion with *Bacillus*.

Sequence and structural analysis

Sequence conservation was analyzed by comparing the aligned OSBS/NAAAR, MLE, and AEE families. Family assignments of MLE and AEE proteins were taken from the SFLD, which uses HMMs and information from the literature to assign proteins to families. Conserved positions were defined as those in which > 90% of family or subfamily members have the same amino acid residue. Phenylalanine and tyrosine or aspartate and glutamate were treated as equivalent. Conserved residues were mapped onto the structures of 1FHV (*E. coli* OSBS) and 1SJB (*Amycolatopsis* NAAAR) in Chimera[77].

Structural superpositions of the whole proteins, capping domains, and barrel domains of 1SJB, 1FHV, *B. bacteriovorus* OSBS, and 1MUC were generated from the structure-based sequence alignment of the MLE subgroup using the Match feature of Chimera[77] or Combinatorial Extension (CE)[83]. To quantify differences in the orientation between the barrel and capping domains, barrel domains of each structure were first superimposed relative to 1SJB. Then, a plane was fit to each capping domain using the alpha carbons of specific sets of residues that are closely aligned in structural superpositions of the capping domains. The dihedral angles between these planes were calculated in Chimera to measure differences in relative rotation between the capping and barrel domains of each pair of superimposed structures.

Acknowledgements

We thank Alexander Fedorov, Elena Fedorov, and Dr. Steven Almo for providing unpublished structure coordinates and Dr. Scott Pegg, Kai-Yeung Lau, Dr. Elaine Meng and Eric Pettersen for their assistance. This work was supported by National Institutes of

Health Grants GM60595 to P.C.B. and P01071790 and GM52594 to J.A.G. M.P.J. acknowledges start-up funds provided by HHMI Biomedical Research Support Program grant #5300246 to the UCSF School of Medicine. M.E.G. is supported by a Postdoctoral Fellowship in Informatics from the Pharmaceutical Researchers and Manufacturers of America.

Table 1. Relative divergence of the OSBS family.

Family for Comparison	Number of species[a]	Compared Family[b]		OSBS[b]	
		Average % identity	Minimum % identity	Average % identity	Minimum % identity
Enolase[c]	66	56%	27%	26%	15%
Galactonate dehydratase[c]	8	55%	32%	31%	20%
Glucarate dehydratase[c,d]	11	78%	66%	45%	20%
AEE[c]	30	38%	24%	33%	18%
MenB	67	58%	35%	26%	14%
MenD	66	32%	21%	26%	14%
MenE	67	27%	14%	26%	14%

[a] OSBSs were compared to proteins from a second family which were taken from the same set of species as the OSBSs.

[b] Percentage identities were calculated as number identical/length of the longer sequence from pairwise alignments generated by ALIGN[84].

[c] Some NAAAR-like proteins not encoded in menaquinone operons are included in the OSBS family.

[d] Glucarate dehydratase related protein, which has an unknown function was excluded.

Table 2. Comparison of root mean square deviation (RMSD)[a] between pairs of structures[b].

	Whole Structure (Å)	Barrel Domain (Å)	Capping Domain (Å)
1FHV vs. 1SJB	3.4 (1.4)	2.4 (1.6)	1.8 (1.0)
1FHV vs. 1MUC	3.3 (1.7)	2.2 (1.6)	1.9 (1.2)
1FHV vs. BDEBA	4.1 (1.4)	3.2 (1.7)	4.7 (1.5)
BDEBA vs. 1SJB	2.7 (1.2)	3.4 (1.6)	2.1 (1.5)
BDEBA vs. 1MUC	2.8 (1.3)	2.6 (1.7)	2.0 (1.4)
1SJB vs. 1MUC	1.3 (0.7)	1.3 (1.0)	0.7 (0.5)

[a] Calculated using the same number of atoms in each comparison (264 for the whole structure, 163 for the barrel domain, and 98 for the capping domain). Similar trends are observed using fewer atoms to calculate the RMSD (125 for the whole structure, 121 for the barrel domain, and 65 for the capping domain), as shown in parentheses.

[b] Abbreviations: 1FHV = *E. coli* OSBS bound to OSB, BDEBA = *B. bacteriovorus* OSBS bound to OSB, 1SJB = *Amycolatopsis* OSBS/NAAAR bound to OSB, and 1MUC = MLE I without ligand.

Table 3. Comparison of capping domain orientation between pairs of structures[a].

	Angle with Barrel Domains Aligned	Angle with Capping Domains Aligned[b]
1FHV vs. 1FHU	3.3°	1.1°
BDEBA-OSB vs. BDEBA-apo	1.0° (3.7°)[c]	0.5° (0.9°)
1SJB vs. 1MUC	5.5°	0.3°
BDEBA-OSB vs. 1MUC	4.2°	1.2°
BDEBA-apo vs. 1MUC	5.2° (7.8°)	1.2° (1.2°)
BDEBA-OSB vs. 1SJB	8.3°	1.3°
1FHV vs. BDEBA-OSB	9.5°	1.1°
1FHU vs. 1MUC	11.8°	0.3°
1FHV vs. 1MUC	13.3°	1.4°
1FHV vs. 1SJB	17.7°	1.7°

[a]Abbreviations: 1FHV = *E. coli* OSBS bound to OSB, 1FHU = *E. coli* OSBS without ligand, BDEBA-OSB = *B. bacteriovorus* OSBS bound to OSB, BDEBA-apo = *B. bacteriovorus* OSBS without ligand, 1SJB = *Amycolatopsis* OSBS/NAAAR bound to OSB, and 1MUC = MLE I without ligand.

[b]As a control for defining planes that reflect domain orientation and not other structural differences, we calculated dihedral angles between planes from superpositions in which only the capping domains were superposed. Since dihedral angles between these planes

should approach zero if they reflect domain orientation alone, the set of residues from all structures that minimizes this angle was used to calculate dihedral angles when the barrel domain was aligned. Dihedral angles between aligned capping domains of different proteins were < 2° for the residue set that minimizes this angle. These values are comparable to those obtained by comparing liganded versus unliganded structures of *E. coli* and *B. bacteriovorus* OSBS's, suggesting that structural differences other than domain rotation marginally affect this measurement.

[c]Because differences in domain orientation could be an artifact of crystal packing, dihedral angles were calculated for all available chains. For 1MUC and 1SJB, this resulted in differences of < 0.7° (data not shown). However, the dihedral angle between the two chains of the BDEBA-apo structure was 2.7° when the barrel domains were aligned. Thus, dihedral angles are given for both chains, with those of chain A in parentheses.

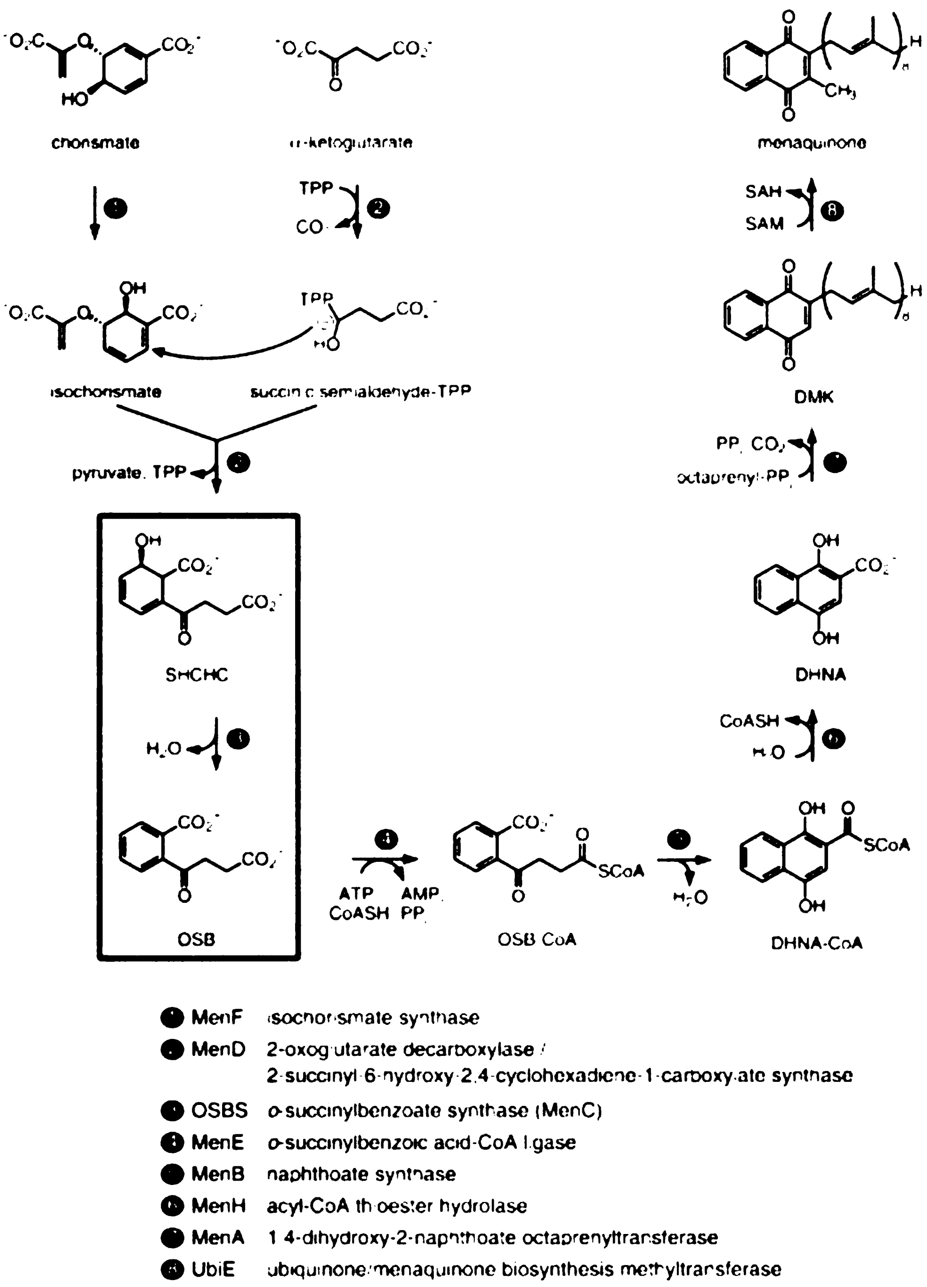

chorismate
α-ketoglutarate
menaquinone
TPP
CO₂
SAH
SAM
isochorismate
succinic semialdehyde-TPP
DMK
PPi, CO₂
octaprenyl-PPi
pyruvate, TPP
SHCHC
DHNA
H₂O
CoASH
H₂O
OSB
ATP AMP
CoASH PPi
OSB CoA
H₂O
DHNA-CoA
● MenF isochorismate synthase
● MenD 2-oxoglutarate decarboxylase / 2-succinyl-6-hydroxy-2,4-cyclohexadiene-1-carboxylate synthase
● OSBS o-succinylbenzoate synthase (MenC)
● MenE o-succinylbenzoic acid-CoA ligase
● MenB naphthoate synthase
● MenH acyl-CoA thioester hydrolase
● MenA 1,4-dihydroxy-2-naphthoate octaprenyltransferase
● UbiE ubiquinone/menaquinone biosynthesis methyltransferase

Figure 1. Menaquinone biosynthesis pathway of *E. coli*[27]. The OSBS reaction is boxed. Compounds are abbreviated as follows: TPP = thiamine pyrophosphate; SHCHC = 2-succinyl-6-hydroxy-2,4-cyclohexadiene-1-carboxylate; OSB = *o*-succinylbenzoate; CoASH = coenzyme A; DHNA = 1,4-dihydroxy-2-naphthoate; DMK = demethylmenaquinone; SAM = *S*-adenosylmethionine; and SAH = *S*-adenosylhomocysteine.

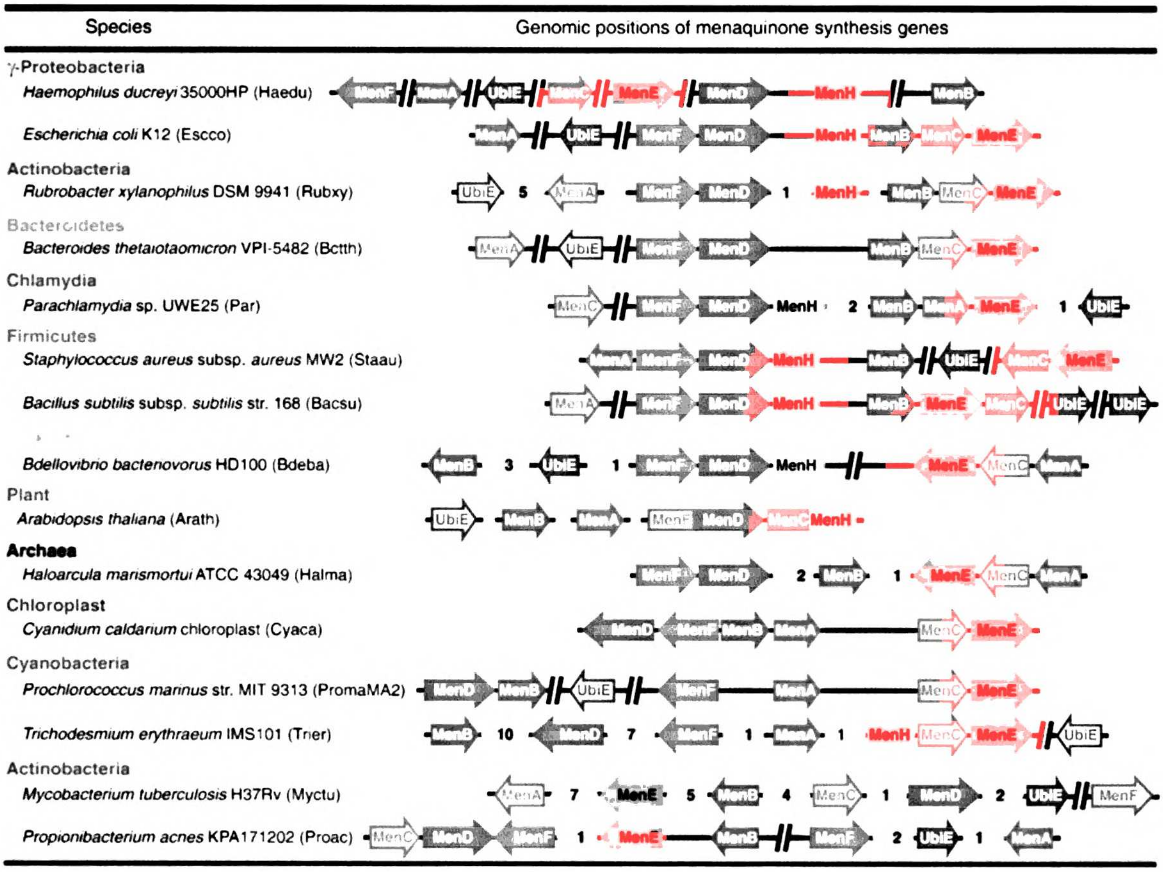

Species
Genomic positions of menaquinone synthesis genes
γ-Proteobacteria
Haemophilus ducreyi 35000HP (Haedu)
Escherichia coli K12 (Escco)
Actinobacteria
Rubrobacter xylanophilus DSM 9941 (Rubxy)
Bacteroidetes
Bacteroides thetaiotaomicron VPI-5482 (Bctth)
Chlamydia
Parachlamydia sp. UWE25 (Par)
Firmicutes
Staphylococcus aureus subsp. aureus MW2 (Staau)
Bacillus subtilis subsp. subtilis str. 168 (Bacsu)
Bdellovibrio bacteriovorus HD100 (Bdeba)
Plant
Arabidopsis thaliana (Arath)
Archaea
Haloarcula marismortui ATCC 43049 (Halma)
Chloroplast
Cyanidium caldarium chloroplast (Cyaca)
Cyanobacteria
Prochlorococcus marinus str. MIT 9313 (PromaMA2)
Trichodesmium erythraeum IMS101 (Trier)
Actinobacteria
Mycobacterium tuberculosis H37Rv (Myctu)
Propionibacterium acnes KPA171202 (Proac)
MenA
MenB
MenC
MenD
MenE
MenF
MenH
UbiE

Figure 2. Genomic context of menaquinone biosynthesis genes from representative species. All identified menaquinone synthesis genes are shown as arrows; hollow arrows indicate provisional assignments, as defined in Materials and Methods. Menaquinone synthesis genes have been aligned to show similarities in gene order; as a result, spaces between genes are not proportional to the length of the DNA separating the genes. Each horizontal segment indicates a contiguous DNA segment. The genomes of some species have multiple chromosomes or have not been completely assembled, as indicated by gaps between segments. Hash marks indicate an intervening region encoding > 40 genes. Smaller intervening regions are shown as light grey arrows with the number of intervening genes and their orientation on the chromosome indicated. Although gene neighborhood in plants does not suggest transcriptional coregulation as it does in most prokaryotes, genome locations of menaquinone synthesis genes in *Arabidopsis thaliana* are shown because two pairs of genes (MenF/MenD and MenC/MenH) are predicted to be gene fusions. Intriguingly, these two fusions are adjacent in the genome, and the gene order resembles that found in many bacteria, suggesting that this locus could be a remnant of DNA that was transferred from the mitochondrial or chloroplast genome to the nucleus. For the complete list of species used in phylogenetic analysis and their menaquinone operons, see Supplemental Figure 1.

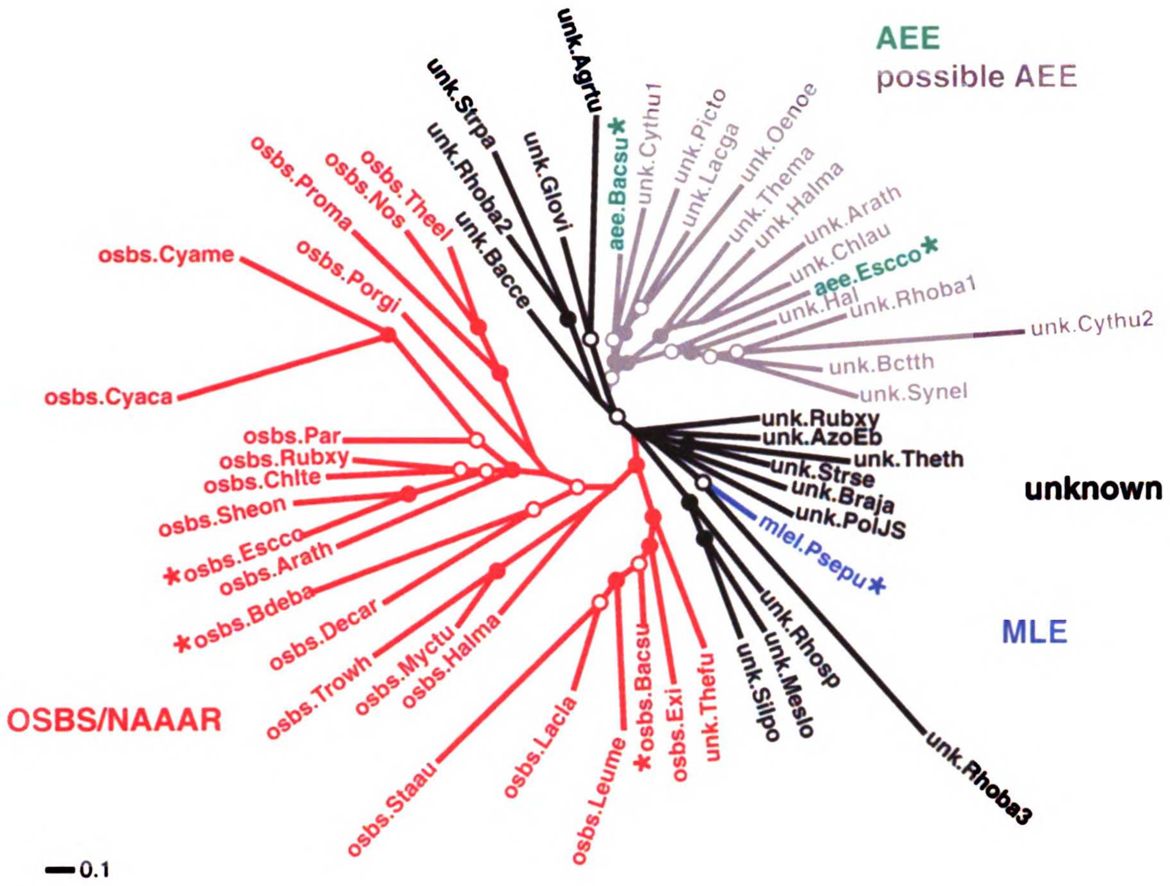
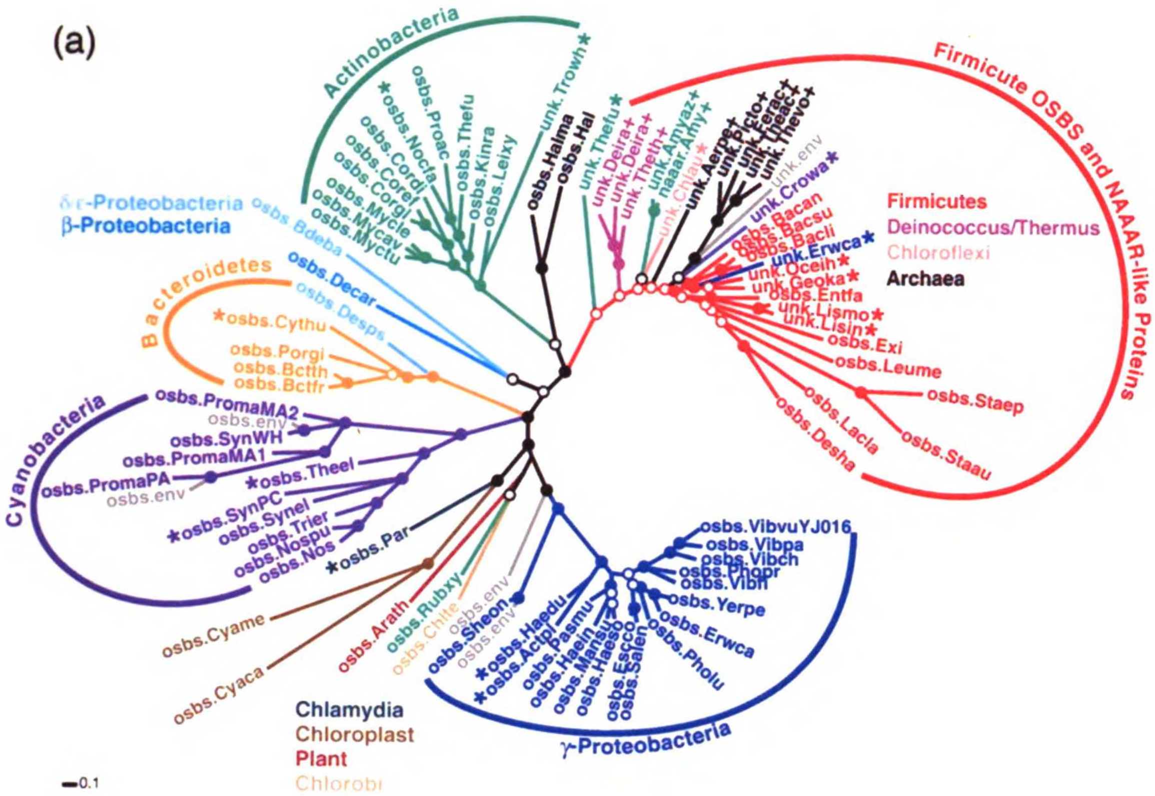


Figure 3. Bayesian phylogenetic tree of the proteins in the MLE subgroup. A representative set of 54 proteins was selected from the 288-protein subgroup by using only proteins sharing < 40% identity. The predicted or verified function is indicated by the prefix "osbs", "aee", or "mleI", and characterized proteins are indicated with an asterisk (*). Proteins of unknown function are prefixed by "unk". OSBS/NAAAR family members are shown in red, characterized AEEs are in green, and MLE I is in blue. Other possible AEEs are in gray, but they cluster with the characterized AEEs with only moderate statistical support. Proteins of unknown function are in black. Branch confidence values are indicated as solid circles (≥ 0.95), hollow circles (0.7-0.94), or no indication (0.5-0.7).


(a)
Actinobacteria
*osbs.Nocfa
osbs.Proac
osbs.Thefu
osbs.Kinra
osbs.Leixy
osbs.Cordi
osbs.Coref
osbs.Corgl
osbs.Mycle
osbs.Mycav
osbs.Myctu
unk.Trowh*
osbs.Halma
osbs.Hal
unk.Thefu*
unk.Deira+
unk.Deira+
unk.Theth+
unk.Amyaz+
naaar.Amy+
unk.Chlau*
unk.Aerpe+
unk.Picto+
unk.Ferac+
unk.Theac+
unk.Thevo+
unk.env
unk.Crowa*
osbs.Bacan
osbs.Bacsu
osbs.Bacli
unk.Erwca*
unk.Oceih*
unk.Geoka*
osbs.Entfa
unk.Lismo*
unk.Lisin*
osbs.Exi
osbs.Leume
osbs.Staep
osbs.Lacla
osbs.Staau
osbs.Desha
Firmicute OSBS and NAAAR-like Proteins
Firmicutes
Deinococcus/Thermus
Chloroflexi
Archaea
δ/ε-Proteobacteria
β-Proteobacteria
osbs.Bdeba
osbs.Decar
osbs.Desps
Bacteroidetes
*osbs.Cythu
osbs.Porgi
osbs.Bctth
osbs.Bctfr
Cyanobacteria
osbs.PromaMA2
osbs.env
osbs.SynWH
osbs.PromaMA1
osbs.PromaPA
osbs.env
*osbs.Theel
*osbs.SynPC
osbs.Synel
osbs.Trier
osbs.Nospu
osbs.Nos
*osbs.Par
osbs.Cyame
osbs.Cyaca
osbs.Arath
osbs.Rubxy
osbs.Chlte
osbs.env
osbs.Sheon
osbs.env
*osbs.Haedu
*osbs.Actpl
osbs.Pasmu
osbs.Haein
osbs.Mansu
osbs.Haeso
osbs.Escco
osbs.Salen
osbs.Pholu
osbs.Erwca
osbs.Yerpe
osbs.Vibfi
osbs.Phopr
osbs.Vibch
osbs.Vibpa
osbs.VibvuYJ016
γ-Proteobacteria
Chlamydia
Chloroplast
Plant
Chlorobi
0.1

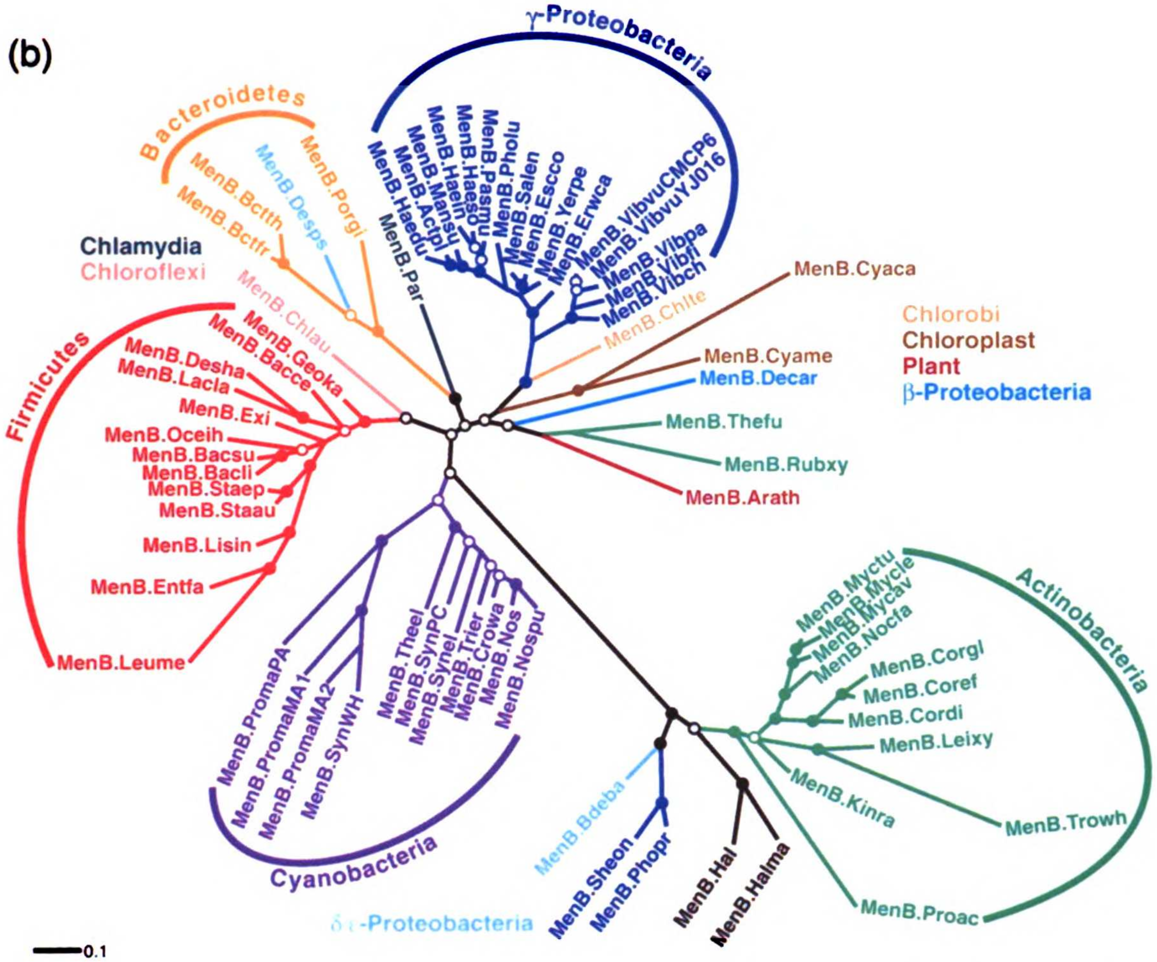


Figure 4. Bayesian phylogenetic tree of the proteins in the OSBS/NAAAR and menB families. Branch confidence values are shown as in Figure 3. A) The OSBS/NAAAR family. To build the tree, the full set of OSBS/NAAAR proteins was filtered to remove proteins sharing > 94% identity with any other in the set. Proteins are colored according to phylum[37], and arcs indicate the main subfamilies. Proteins in gray are environmental sequences derived from the Sargasso Sea data set[85]. A plus sign (+) indicates NAAAR-like proteins found in strains in which menaquinone synthesis genes could not be identified. An asterisk (*) identifies proteins which are not encoded in menaquinone operons but are found in strains which have the menaquinone pathway. Two of these species (*Erwinia carotovora* (Erwca) and *Thermobifida fusca* (Thefu)), encode both an

OSBS in the menaquinone operon and a NAAAR-like protein elsewhere in the genome.

B) The menB family. Proteins are colored as in part A.

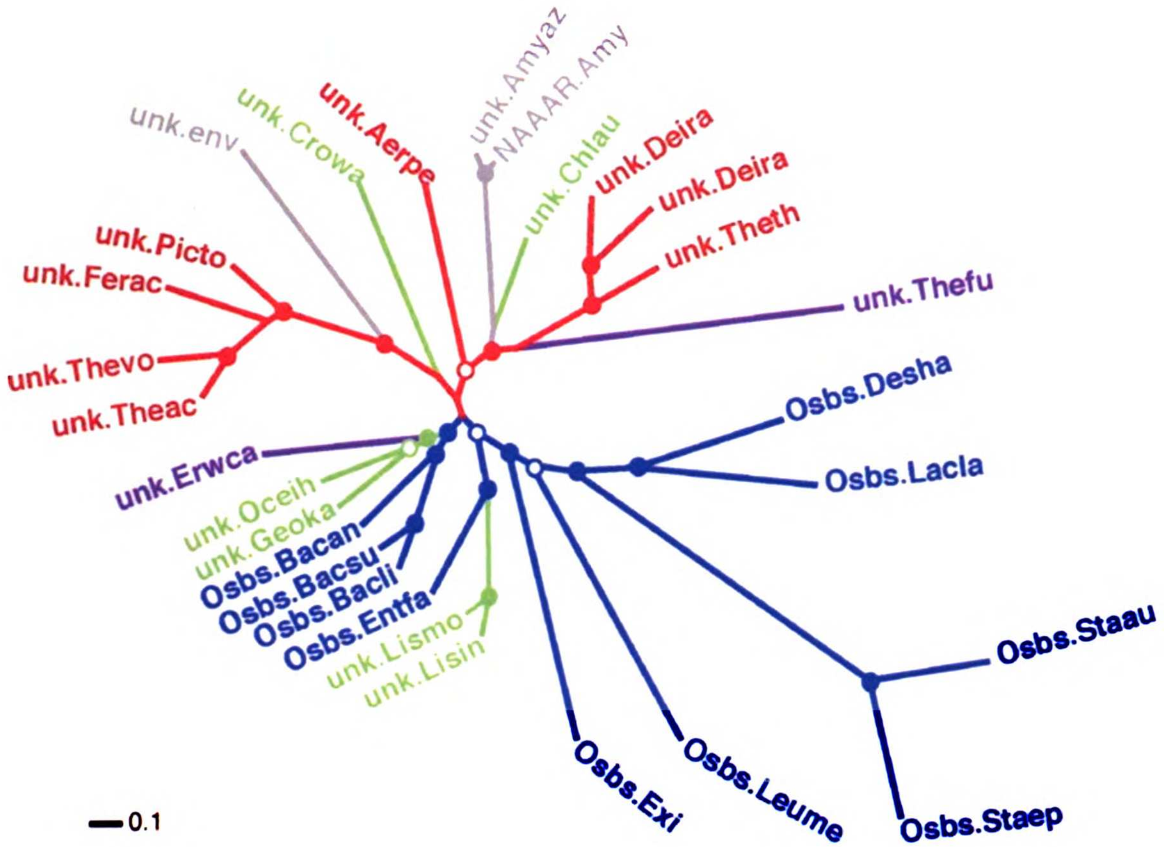


Figure 5. Bayesian phylogenetic tree of the proteins in the Firmicute OSBS/NAAAR subgroup. Blue indicates that the OSBS is encoded in a menaquinone operon; green indicates that the OSBS/NAAAR protein is not encoded in a menaquinone operon, but the species has the menaquinone pathway; purple indicates that the NAAAR-like protein is not encoded in the menaquinone operon, but a different OSBS is encoded by the menaquinone operon; red indicates that there is no menaquinone operon detected in the species; gray indicates that the genome sequence is unavailable. Branch confidence values are indicated as in Figure 3.

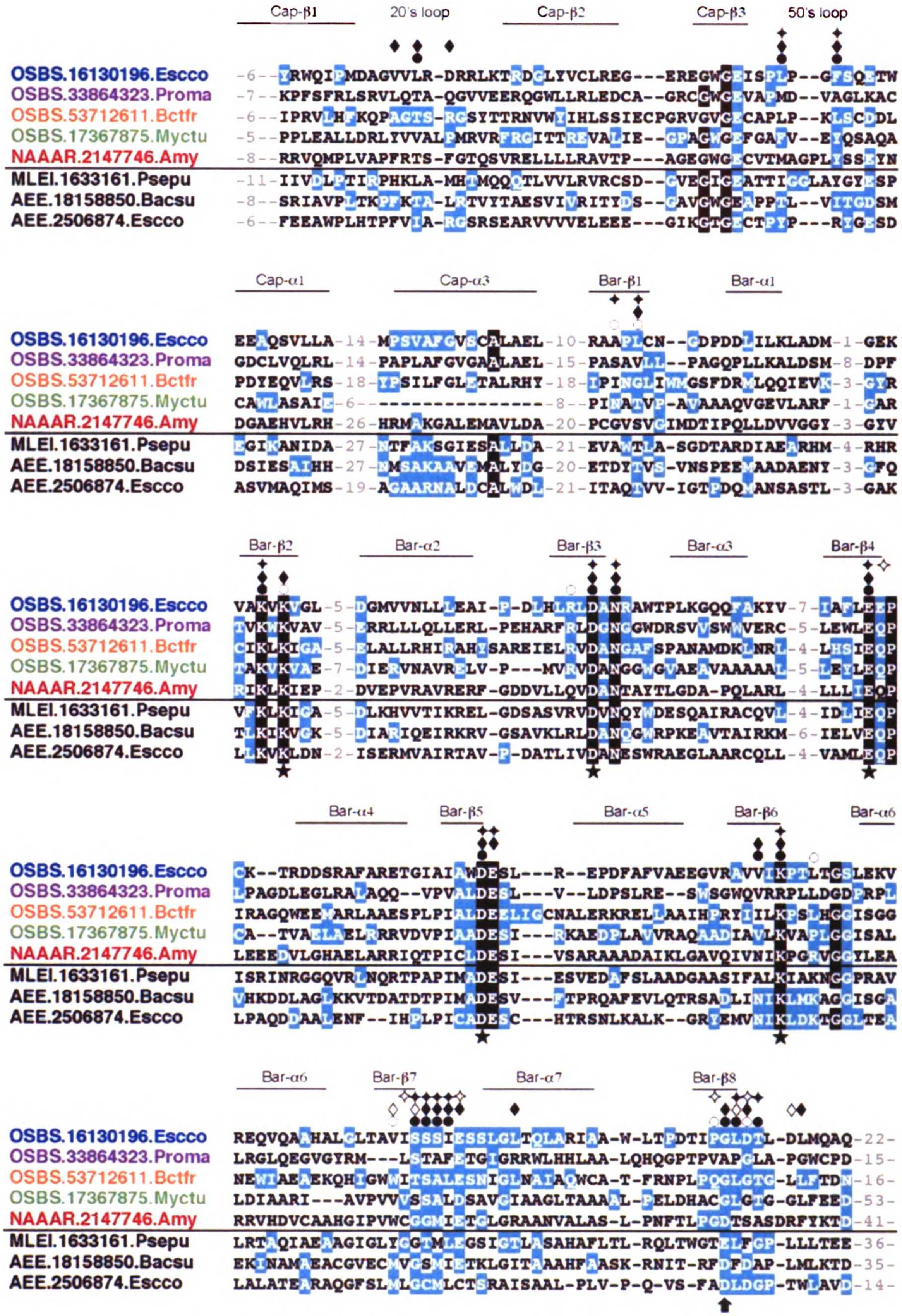
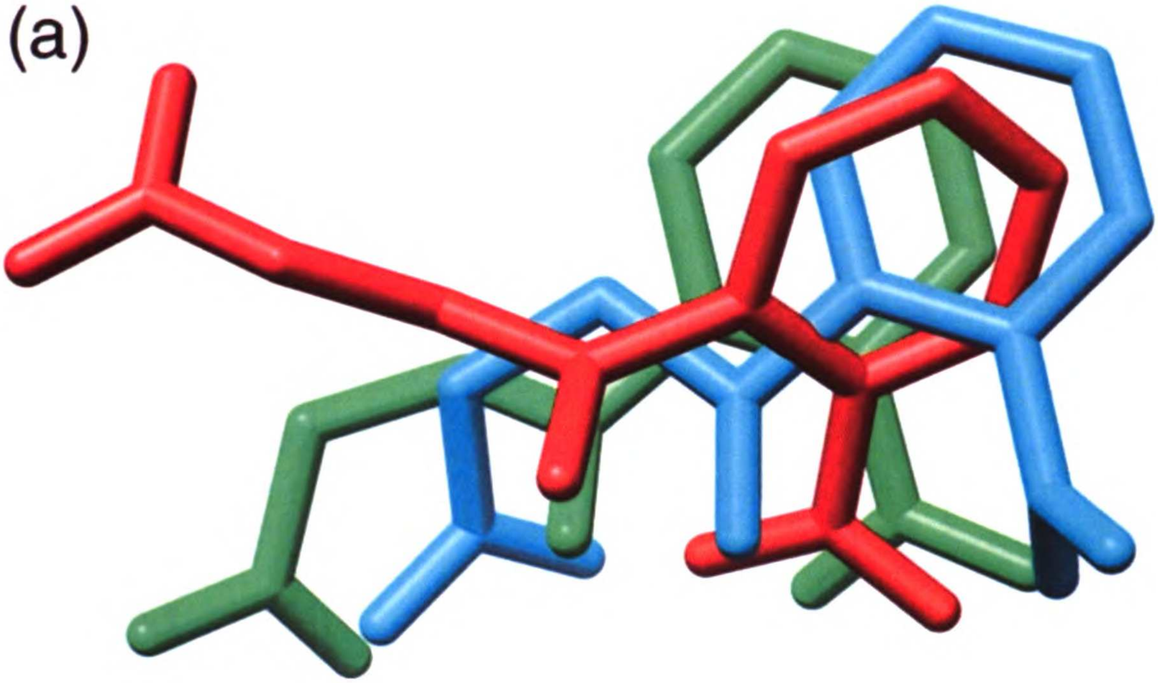

Cap-β1
20's loop
Cap-β2
Cap-β3
50's loop
Cap-α1
Cap-α3
Bar-β1
Bar-α1
Bar-β2
Bar-α2
Bar-β3
Bar-α3
Bar-β4
Bar-α4
Bar-β5
Bar-α5
Bar-β6
Bar-α6
Bar-β7
Bar-α7
Bar-β8
OSBS.16130196.Escco
OSBS.33864323.Proma
OSBS.53712611.Bctfr
OSBS.17367875.Myctu
NAAAR.2147746.Amy
MLEI.1633161.Psepu
AEE.18158850.Bacsu
AEE.2506874.Escco

Figure 6. Analysis of sequence conservation in the OSBS/NAAAR family. The sequence alignment shows representatives of each of the five OSBS/NAAAR subfamilies, the MLE family, and two AEE subfamilies. The membership of each OSBS/NAAAR subfamily is shown in Figure 4A, as indicated by the arcs, with the exception that the NAAAR-like *T. fusca* protein (unk.Thefu) was not included in this analysis. γ-Proteobacteria is represented by OSBS.16130196.Escco, Cyanobacteria by OSBS.33864323.Proma, Bacteroidetes by OSBS.53712611.Bctfr, Actinobacteria by OSBS.17367875.Myctu, and the Firmicute/NAAAR-like protein subfamily by NAAAR.2147746.Amy. The membership of the AEE subfamilies and the MLE family consists of proteins sharing > 40% identity with each sequence that is shown. Magenta residues indicate conservation in > 90% of subfamily members, and black residues indicate conservation in both > 90% of the subfamily and > 90% of the entire MLE subgroup. Gray numbers indicate the length of segments that are not shown. Secondary structure of the capping and barrel domains are indicated by Cap- and Bar-, respectively. Catalytic residues are indicated by a five-pointed star below the sequences (★). Positions of residues lining the active site pocket are indicated for *E. coli* OSBS (●), *Amycolatopsis* OSBS/NAAAR (♦), and *B. bacteriovorus* OSBS (✦, sequence not shown). Solid symbols represent residues < 5 Å away from bound OSB, and open symbols indicate residues 5-6 Å away from the ligand. The arrow indicates the position of the glutamate or aspartate to glycine mutation that confers OSBS activity on *E. coli* AEE or *Pseudomonas* sp. P51 MLE II[6].


(a)

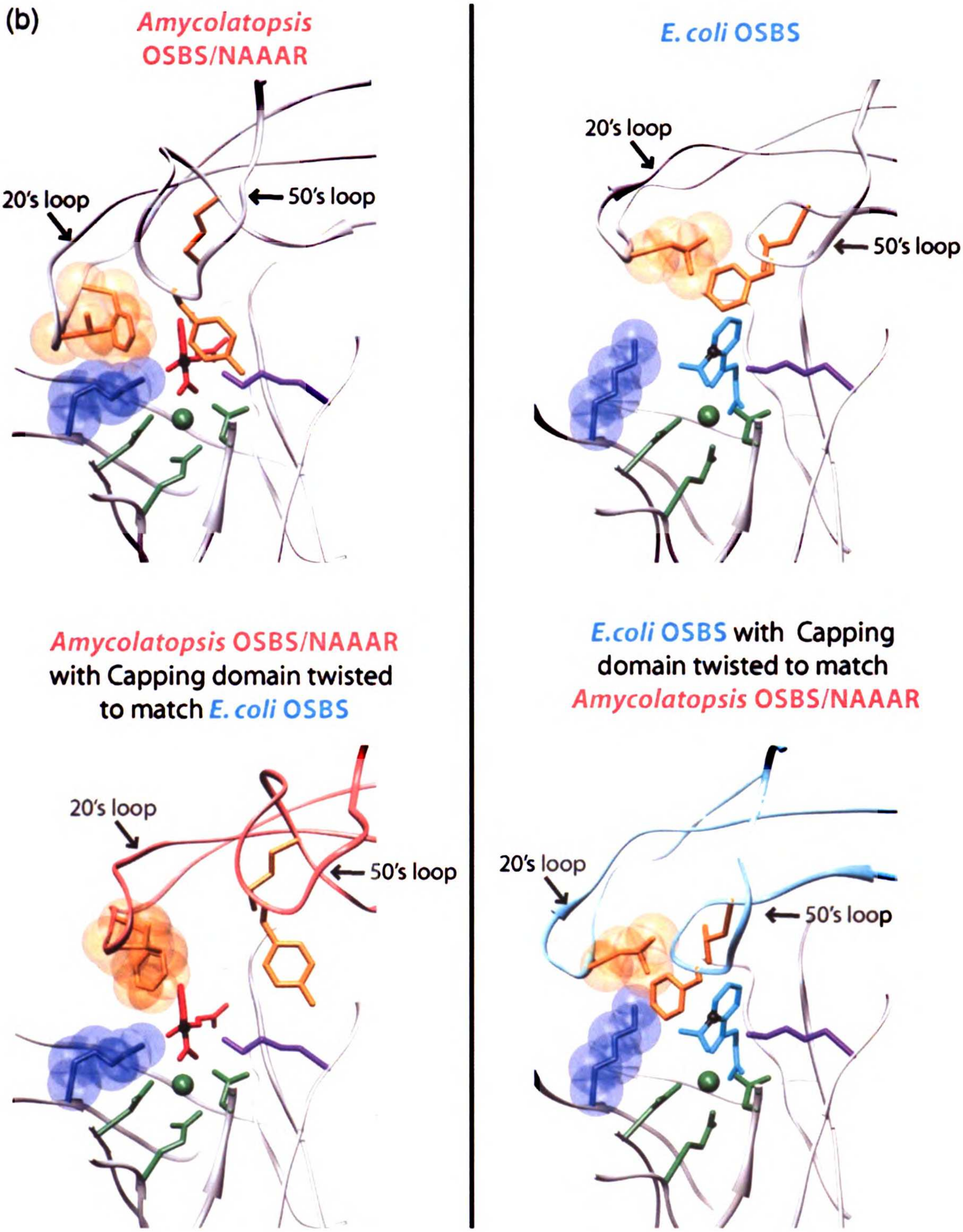

(b)
Amycolatopsis OSBS/NAAAR
20's loop
50's loop
E. coli OSBS
20's loop
50's loop
Amycolatopsis OSBS/NAAAR with Capping domain twisted to match E. coli OSBS
20's loop
50's loop
E.coli OSBS with Capping domain twisted to match Amycolatopsis OSBS/NAAAR
20's loop
50's loop

Figure 7. Structural differences in the active sites of OSBS/NAAAR family proteins. A) Comparison of OSB binding orientation. *Amycolatopsis* OSBS/NAAAR (1SJB) is red, *E. coli* OSBS (1FHV) is cyan, and *B. bacteriovorus* OSBS is green. B) Comparison of the

20s and 50s loop positions in *E. coli* OSBS and *Amycolatopsis* OSBS/NAAAR. The native structures are shown in the top panels. In the bottom panels, the capping domain of *Amycolatopsis* OSBS/NAAAR has been rotated to match the position of the *E. coli* OSBS capping domain (left), and the *E. coli* OSBS capping domain has been rotated to match the *Amycolatopsis* OSBS/NAAAR capping domain (right). Metal binding residues and the metal ion are shown in green, the Bar-β2 lysine that acts as the general base is shown in blue, the Bar-β6 lysine required for catalysis is purple, and residues on the 20s and 50s loops that contact the ligand are in orange. The carbon from which the proton is abstracted is shown in black.

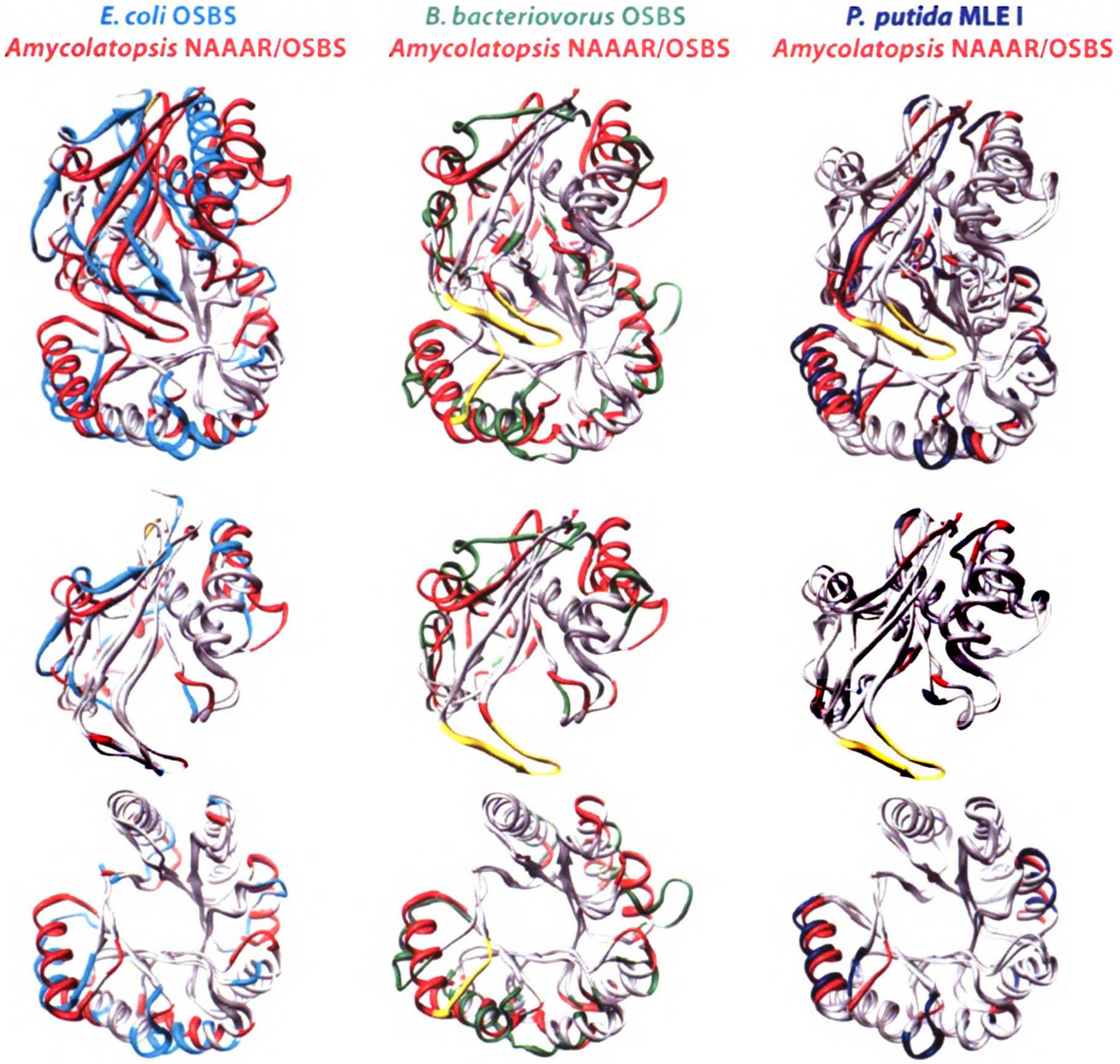


Figure 8. Overall structural differences within the OSBS/NAAAR family. Superpositions of the whole protein are shown at the top, superpositions of the capping domain are shown in the middle, and superpositions of the barrel domain are shown at the bottom. Colored segments show regions where the aligned alpha carbons are > 3 Å apart or where there is an insertion in one sequence relative to the other. Segments in yellow correspond to disordered regions in the other structure. *Amycolatopsis* OSBS/NAAAR (1SJB) is red, *E. coli* OSBS (1FHV) is cyan, *B. bacteriovorus* OSBS is green, and *P. putida* MLE I (1MUC) is blue.

Chapter 2

Molecular Modeling and S1 Subsite Prediction of Metacaspase Proteins

Summary

In the paper "*Plasmodium berghei* metacaspase 1 is involved in controlling parasite numbers in the mosquito," my specific contributions were as follows:

1. Predicting the catalytic and specificity determining residues of 9 metacaspase proteins, shown as a table in the *Results* section. This is based on a profile-profile alignment of the metacaspases to the caspase family of sequences, as described in the *Methods* section.
2. Molecular modeling of metacaspase proteins from *Plasmodium berghei*, *Arabidopsis thaliana*, *Leishmania major* strain Friedlin, and *Trypanosoma brucei*. The active site of each protein is shown as a figure in the *Results* section.

These contributions are included here (Table 2-1 and Figure 2-1) and a draft of the paper follows (Appendix A). The co-author listed in this publication, Dr. Mohammed Sajid, directed and supervised the research that forms the basis for this paper. The paper is currently in preparation to be submitted for publication.

Methods

Molecular modeling and S1 subsite prediction of PbMC1, AtMC1, LmMC, and TbMC

Alignment of PbMC1 with known metacaspases and caspases was used to identify the core enzymatic domain of PbMC1. Using the protein structure prediction program PRIME (Schrödinger, Inc.), one hundred sequences with structures in the Protein Data Bank were identified as having significant homology to the predicted secondary structure of PbMC1. Human caspase 3 (PDB ID: 1CP3) was selected as the template for modeling PbMC1 due to its high PRIME threading rank (number 3) and because it is also a clan CD enzyme. A BLAST search of the NCBI non-redundant sequence database identified sequences with expectation values better than $1e^{-20}$ to 1CP3, PbMC1, AtMC1, LmMC, and TbMC. The sequences were filtered at 90% identity to remove redundant sequences, and the multiple sequence alignment program MUSCLE was used to construct a profile-profile alignment of the caspases to the metacaspases. Homology models of PbMC1, AtMC1, LmMC, and TbMC were generated using the Protein Local Optimization Program (Dr. Matthew P. Jacobson, UCSF). The p17 and p12 heavy and light chains of caspase 3 were modeled as one polypeptide.

Results

Table 2-1. Predicted catalytic and S1 subsite residues of metacaspase proteins.

Predictions are based on profile-profile alignment of metacaspase sequences to caspases.

Human caspase 3 is shown for reference.

Protein	Annotation	GI	Organism	Cat 1	Cat 2	1st S1	2nd S1	3rd S1
1CP3	Caspase 3	2780971	*Homo sapiens*	H121	C163	R64	Q161	R207
PbMC1	Metacaspase 1	29788140	*Plasmodium berghei*	H408	C464	L338	D463	E511
PbMC2	Metacaspase 2	29788142	*Plasmodium berghei*	Y1277	T1330	L1207	D1328	-
PbMC3	Metacaspase 3	-	*Plasmodium berghei*	K218	S277	L75	D275	-
AtMC1	Metacaspase 1	30678252	*Arabidopsis thaliana*	H164	C220	L96	D218	D274
AtMC6	Metacaspase 6	15219342	*Arabidopsis thaliana*	H86	C139	L19	D137	D344
LmMC	Metacaspase	72547606	*Leishmania major* strain Friedlin	H147	C202	L78	D200	D255
LdMC1	Metacaspase 1	87116787	*Leishmania donovani*	H147	C202	L78	D200	D255
LdMC2	Metacaspase 2	87116789	*Leishmania donovani*	H147	C202	L78	D200	D255
TbMCA2	Metacaspase MCA2	72389843	*Trypanosoma brucei*	H158	C213	L89	D211	D266

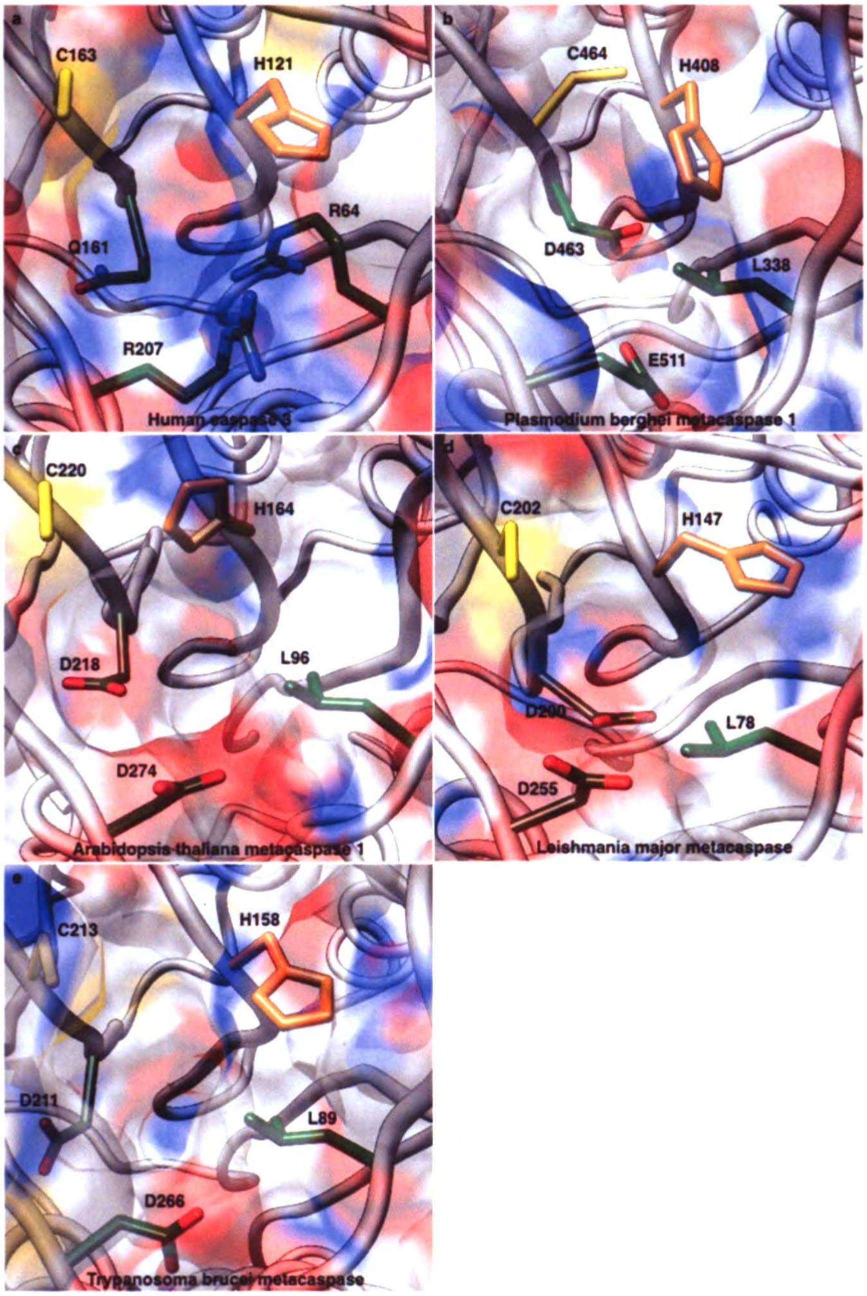

a
C163
H121
R64
Q161
R207
Human caspase 3
b
C464
H408
D463
L338
E511
Plasmodium berghei metacaspase 1
c
C220
H164
D218
L96
D274
Arabidopsis thaliana metacaspase 1
d
C202
H147
D200
L78
D255
Leishmania major metacaspase
e
C213
H158
D211
L89
D266
Trypanosoma brucei metacaspase

Figure 2-1. Structural models showing predicted catalytic and S1 subsite residues of metacaspases. The metacaspases that were modeled include (b) PbMC1, (c) AtMC1, (d) LmMC, and (e) TbMCA2. (a) The crystal structure of the template, 1CP3, is shown for reference. The residue colors are as follows: S1 subsite residues, green; histidine, orange; cysteine, yellow.

Chapter 3

Searching for the Natural Substrate of the *Shaker* Family K+ Channel ß Subunit

Summary

In this study, we attempt to identify the natural substrate of the *Shaker* family (Kv1) potassium channel ß subunit (Kvß2). In order to better understand ß-mediated axonal targeting of Kv1, a computational docking program was used to screen a virtual library of small molecule metabolites against the x-ray structure of Kvß2 from rat. The docking results were rescored using a physics based scoring function, and a subset of the top scoring compounds was selected for experimental validation.

Introduction

Voltage-gated potassium channels regulate the flow of potassium through the plasma membrane in response to changes in membrane voltage. The archetypal voltage-gated potassium channel is the *Shaker* channel from *Drosophila melanogaster* (Tempel et al. 1987). The *Shaker* channel was the first potassium channel to be cloned; it is an A-type potassium channel and carries the I_A current, which functions in the repolarization of action potentials (Hille 1992). Mutations at the *Shaker* locus on the X chromosome cause fruit flies to shake their legs under ether anesthesia. Homologues of the *Shaker* channel have been described in animals, plants, fungi, and prokaryotes. In humans, mutations of the homologous gene, KCNA1, are associated with episodic ataxia type 1 (Gulbis et al. 1999).

The *Shaker* channel consists of four alpha subunits that form the transmembrane channel and four beta subunits that attach to the cytosolic face of the channel to form a macromolecular complex (Weng et al. 2006). It has been demonstrated that the ß subunit Kvß2 is responsible for axonal targeting of the *Shaker* channel (Gu et al. 2006). Kvß2 is a member of the aldo-keto reductase family that reduces small molecule aldehydes to alcohols by oxidizing bound NADPH cofactor. The substrate specificity of Kvß2 is broad; artificial substrates that turn over slowly have been identified and include the benzaldehyde derivatives 4-carboxybenzaldehyde and 4-cyanobenzaldehyde (Weng et al. 2006). However, the native substrate of Kvß2, which may provide the link between the oxidation-reduction potential of the cell and targeting of Kv1 to the axon, has not been identified. In order to better understand ß-mediated axonal targeting of Kv1, we aim to identify the natural substrate of Kvß2.

Methods

Using the docking program Glide (Schrödinger, Inc.), the KEGG database of 60,000 compounds was docked against 1QRQ, the crystal structure of Kvß2 from *Rattus norvegicus*. The ligand-sampling grid was defined by superimposing 1AFS, the crystal structure of testosterone and 3 α-hydroxysteroid dehydrogenase, onto 1QRQ and computing the centroid of the ligand testosterone. Two docking runs were performed using the oxidized (NADP+) and reduced (NADPH) forms of the cofactor.

The docking results were rescored using the Protein Local Optimization Program (Dr. Matthew Jacobson, UCSF), which employs a physics based scoring function to compute the binding energy of each ligand. The poses of the top 200 compounds in each list were examined to identify substrates for experimental validation. A cutoff of 5 Å was applied to the distance from the carbonyl oxygen of each substrate to C4 of the cofactor's nicotinamide ring, which donates a hydride.

Results and Discussion

Proteins in the aldo-keto reductase superfamily metabolize a diverse range of substrates including aliphatic and aromatic aldehydes, neurotransmitter aldehydes, lipid-derived aldehydes, monosaccharides, glucuronate, steroids, prostaglandins, polycyclic aromatic hydrocarbons, flavonoids, and xenobiotics (Jez et al. 1997). The top ranked compounds after rescoring with cofactors NADP+ and NADPH are shown in Table 3-1 and Table 3-2, respectively. Although there was no bias in the computations toward any class of substrates, a large number of ketones are represented among the top hits. The top hits contain many flavonoids, monosaccharides, and disaccharides. There are also polycyclic aromatic hydrocarbons and aminoglycoside antibiotics such as streptomycin. 15 potential substrates of Kvß2 were selected for experimental validation (Table 3-3). These potential substrates, which include the monosaccharides glucose, galactose, arabinose, xylulose, and threose, will be tested by our experimental collaborators in the laboratory of Dr. Lily Jan, UCSF.

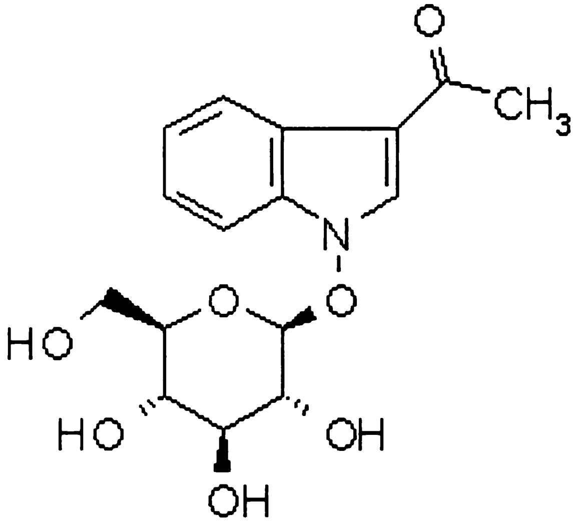
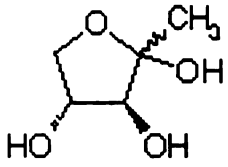
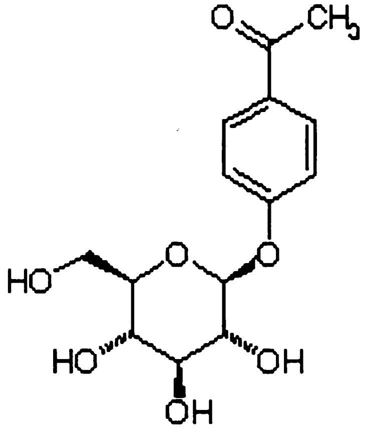
Table 3-1. Top compounds after rescoring (NADP+ cofactor).

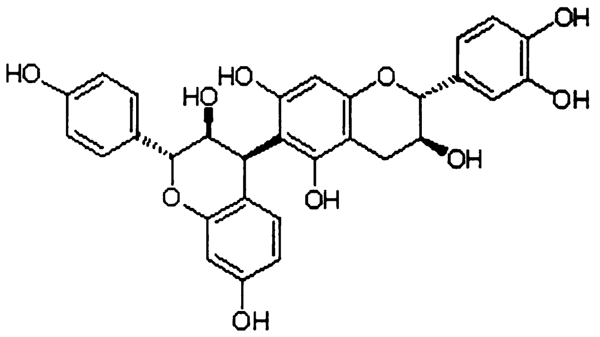
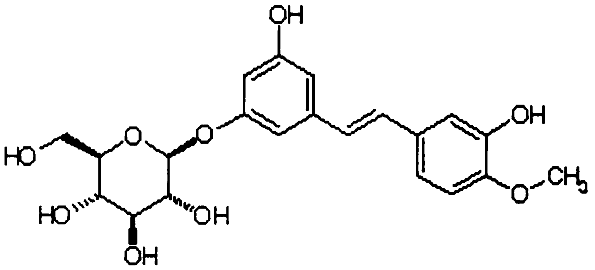
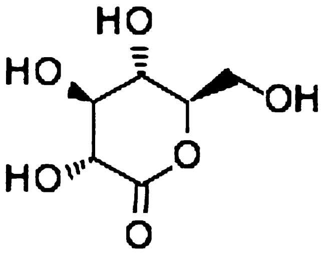
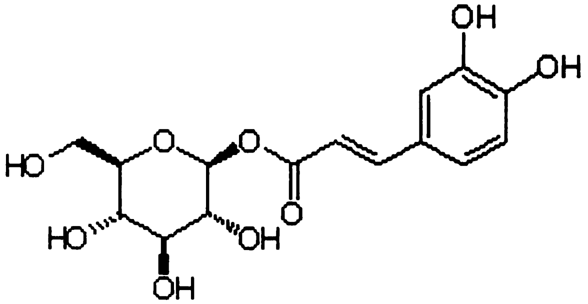
Rank	KEGG ID	Name
1	C12638	Quercetin 3-O-(6-O-malonyl-beta-D-glucoside); Quercetin 3-O-malonylglucoside; Quercetin-3-O-(6"-malonylglucoside); [PubChem:583028]
2	C08620	Cyanidin 3-O-rutinoside; Cyanidin 3-O-rhamnosylglucoside; [PubChem:10813]
3	C01130	-
4	C09803	Neoastilbin; (2S,3S)-Taxifolin 3-rhamnoside; [PubChem:11991]
5	C06449	Myxochlin B; [PubChem:8682]
6	C00299	Uridine; [PubChem:3593] [ChEBI:16704] [CCD:URI]
7	C08649	Isobutrin; [PubChem:10842]
8	C10108	Myricitrin; Myricetin 3-O-rhamnoside; [PubChem:12294]
9	C09763	Manniflavanone; [PubChem:11951]
10	C04452	4-Nitrophenol-alpha-D-galactopyranoside; [PubChem:7079] [CCD:147]
11	C06449	Myxochlin B; [PubChem:8682]
12	C08307	Hordatine A; [PubChem:10505]
13	C03946	Flavonol 3-O-beta-D-glucoside; Flavonol 3-O-D-glucoside; [PubChem:6669]
14	C12404	Urdamycinone B; [PubChem:582794]
15	C12472	Novclobiocin 105; [PubChem:582862]
16	C05839	cis-beta-D-Glucosyl-2-hydroxycinnamate; beta-D-Glucosyl-2-coumarinate; [PubChem:8132]
17	C01750	Quercitrin; [PubChem:4883] [ChEBI:17558]
18	C10173	Swainsonine; [PubChem:12359] [CCD:SWA]
19	C10108	Myricitrin; Myricetin 3-O-rhamnoside; [PubChem:12294]
20	C12404	Urdamycinone B; [PubChem:582794]
21	C06257	1-Deoxy-D-xylulose; [PubChem:8496] [ChEBI:28354]
22	C04124	Indole-3-acetyl-beta-1-D-glucose; [PubChem:6810]
23	C12404	Urdamycinone B; [PubChem:582794]
24	C08330	p-Glucosyloxymandelonitrile; [PubChem:10528]
25	C06721	cis-1,2-Dihydroxy-1,2-dihydrodibenzothiophene; [PubChem:8944] [ChEBI:16941]
26	C03946	Flavonol 3-O-beta-D-glucoside; Flavonol 3-O-D-glucoside; [PubChem:6669]
27	C09126	Genistein 7-O-beta-D-glucoside; Genistin; [PubChem:11318]
28	C05855	4-Hydroxycinnamyl alcohol 4-D-glucoside; p-Coumaryl alcohol 4-O-glucoside; [PubChem:8148]
29	C10108	Myricitrin; Myricetin 3-O-rhamnoside; [PubChem:12294]
30	C06205	1,2-Dihydronaphthalene-1,2-diol; [PubChem:8455] [ChEBI:28516]
31	C04514	(1S,2S)-1,2-Dihydronaphthalene-1,2-diol; trans-1,2-Dihydronaphthalene-1,2-diol; [PubChem:7127] [ChEBI:28809]
32	C10720	Picein; [PubChem:12903]
33	C04314	cis-1,2-Dihydronaphthalene-1,2-diol; (1R,2S)-1,2-Dihydronaphthalene-1,2-diol; [PubChem:6972] [ChEBI:15561] [CCD:NDH]
34	C10141	Deoxymannojirimycin; DMJ; [PubChem:12327] [CCD:DMJ]
35	C12407	104-2; [PubChem:582797]
36	C10231	Guibourtinidol-(4alpha->6)-catechin; [PubChem:12417]
37	C10456	Forsythiaside; [PubChem:12639]
38	C10057	Hinokiflavone; [PubChem:12243]
39	C07730	Transferred to D00632; [PubChem:9932]
40	C11611	Phenyl beta-D-glucopyranoside; Phenylglucoside; [PubChem:13776]

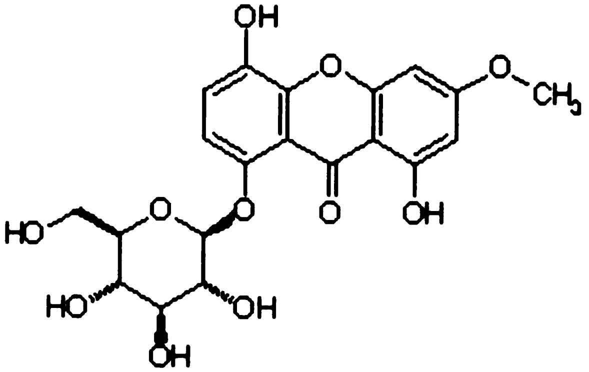
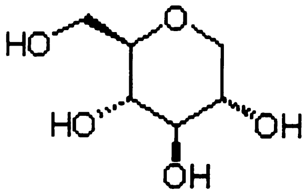
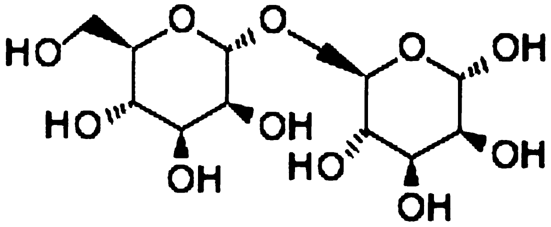
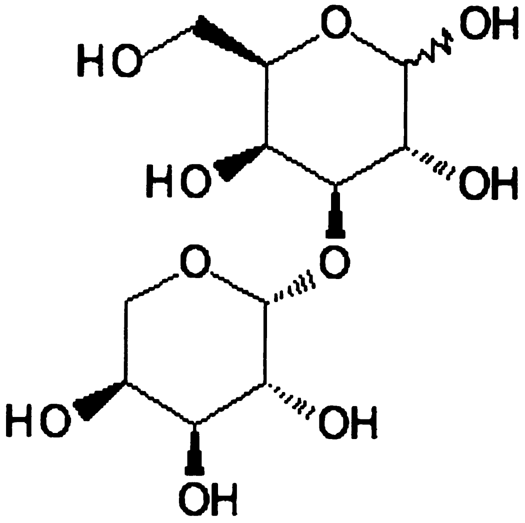
Table 3-2. Top compounds after rescoring (NADPH cofactor).

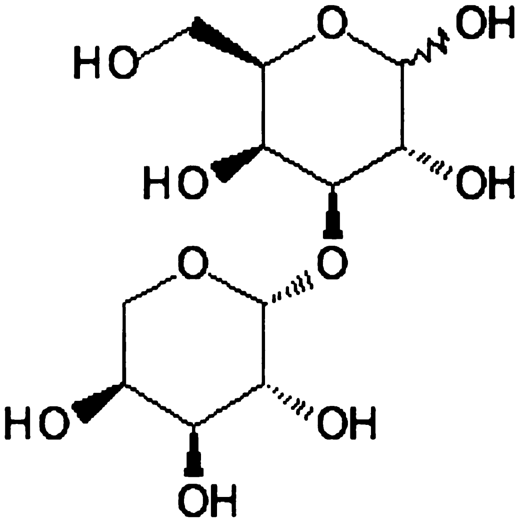
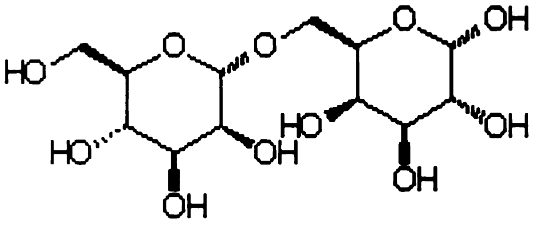
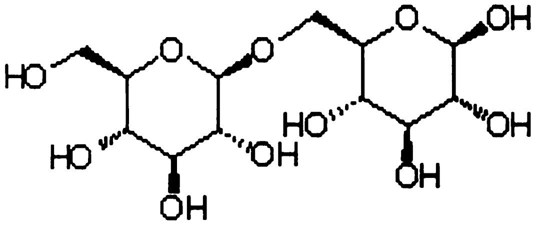
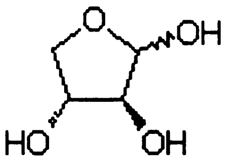
Rank	KEGG ID	Name
1	C09803	Neoastilbin; (2S,3S)-Taxifolin 3-rhamnoside; [PubChem:11991]
2	C03946	Flavonol 3-O-beta-D-glucoside; Flavonol 3-O-D-glucoside; [PubChem:6669]
3	C04124	Indole-3-acetyl-beta-1-D-glucose; [PubChem:6810]
4	C09803	Neoastilbin; (2S,3S)-Taxifolin 3-rhamnoside; [PubChem:11991]
5	C03946	Flavonol 3-O-beta-D-glucoside; Flavonol 3-O-D-glucoside; [PubChem:6669]
6	C12721	Transferred to D01291; [PubChem:583110]
7	C10108	Myricitrin; Myricetin 3-O-rhamnoside; [PubChem:12294]
8	C10216	Daidzin; Daidzein 7-O-glucoside; [PubChem:12402] [ChEBI:4307] [CCD:DZN]
9	C03503	Glucosyloxyanthraquinone; [PubChem:6315]
10	C10073	Hyperin; Quercetin 3-galactoside; [PubChem:12259]
11	C06342	Quinolin-2,8-diol; [PubChem:8578] [ChEBI:17715] [3DMET:B00945]
12	C10216	Daidzin; Daidzein 7-O-glucoside; [PubChem:12402] [ChEBI:4307] [CCD:DZN]
13	C08330	p-Glucosyloxymandelonitrile; [PubChem:10528]
14	C12404	Urdamycinone B; [PubChem:582794]
15	C10155	Lentiginosine; [PubChem:12341]
16	C09794	Oleuropein; [PubChem:11982]
17	C01604	Phlorizin; Phlorhizin; Phloridzin; [PubChem:4758]
18	C08649	Isobutrin; [PubChem:10842]
19	C08308	Hordatine B; [PubChem:10506]
20	C06721	cis-1,2-Dihydroxy-1,2-dihydrodibenzothiophene; [PubChem:8944] [ChEBI:16941]
21	C10073	Hyperin; Quercetin 3-galactoside; [PubChem:12259]
22	C12475	Desmethyldescarbamoylnovobiocin; [PubChem:582865]
23	C12404	Urdamycinone B; [PubChem:582794]
24	C09805	Neoeriocitrin; Eriodictyol 7-O-neohesperidoside; [PubChem:11993]
25	C04167	(+/-)-trans-Acenaphthene-1,2-diol; [PubChem:6849] [ChEBI:28395]
26	C10720	Picein; [PubChem:12903]
27	C12095	Cyanidin 3-O-(6-O-p-coumaroyl)glucoside; [PubChem:14243]
28	C08330	p-Glucosyloxymandelonitrile; [PubChem:10528]
29	C05839	cis-beta-D-Glucosyl-2-hydroxycinnamate; beta-D-Glucosyl-2-coumarinate; [PubChem:8132]
30	C10216	Daidzin; Daidzein 7-O-glucoside; [PubChem:12402] [ChEBI:4307] [CCD:DZN]
31	C08481	Indican; Indican, plant; [PubChem:10674] [ChEBI:16700]
32	C07349	3'-Demethylstaurosporine; [PubChem:9556] [ChEBI:15692]
33	C01421	Daphnin; [PubChem:4608] [ChEBI:17989]
34	C08222	Transferred to D01019
35	C12404	Urdamycinone B; [PubChem:582794]
36	C10885	Simplexoside; [PubChem:13068]
37	C05855	4-Hydroxycinnamyl alcohol 4-D-glucoside; p-Coumaryl alcohol 4-O-glucoside; [PubChem:8148]
38	C06344	3-Methyl-quinolin-2,8-diol; [PubChem:8580] [3DMET:B00947]
39	C06205	1,2-Dihydronaphthalene-1,2-diol; [PubChem:8455] [ChEBI:28516]
40	C01130	-

Table 3-3. Potential Kvß substrates selected for testing.

Rank	KEGG ID	Name	Structure
3	C04124	Indole-3-acetyl-beta-1-D-glucose	 C04124
21	C06257	D-Xylulose (1-Deoxy-D-xylulose)	 C06257
26	C10720	Picein	 C10720

36	C10231	Guibourtinidol-(4alpha->6)-catechin	 C10231
69	C10288	Rhaponticin; Rhapontin	 C10288
81	C00198	D-Glucose (D-Glucono-1,5-lactone)	 C00198
91	C10433	1-Caffeoyl-beta-D-glucose	 C10433

95	C10093	Swertianolin; Bellidifolin-8-O-glucoside	 OH O O CH₃ HO O O O OH HO OH OHC10093
134	C07326	Sorbitol (1,5-Anhydro-D-glucitol; 1,5-Anhydro-D-sorbitol; 1,5-Anhydroglucitol)	 C07326
143	C01728	Mannobiose	 C01728
146	C07285	D-Galactose (Arabino-galactose)	 C07285

146	C07285	D-Arabinose (Arabino-galactose)	 C07285
154	C05400	Epimelibiose	 C05400
163	C08240	Gentiobiose	 C08240
211	C06463	D-Threose; D-threo-Tetrose	 C06463

Chapter 4

Consensus of Docking Results from Isofunctional Proteins: A Method for Decreasing Random Error in Computational Docking

Summary

In this study, we develop an algorithm to perform consensus scoring using computational docking and rescoring results from isofunctional proteins. Isofunctional proteins include homologues from the same isofunctional family as well as enzymes that are adjacent in a metabolic pathway and bind chemically similar substrates. We hypothesize that random error will be reduced in the consensus, leading to an increase in the rank of the native substrate. The algorithm is tested using docking and rescoring results for 283 proteins from the dipeptide epimerase, muconate lactonizing enzyme (MLE) I, and MLE II families of the enolase superfamily, a mechanistically diverse superfamily of enzymes that are related by their ability to catalyze the abstraction of a proton alpha to a carboxylic acid group to form an enolic intermediate.

Specific Aims

This study aims to:

1. Show that using consensus methods to combine docking or rescoring results from isofunctional proteins increases signal to noise in the consensus.
2. Ultimately apply such methods to the prediction of ligand binding in proteins of unknown function.

Introduction

A major problem in computational docking is distinguishing true binders or native substrates from false positive compounds that do not bind. Consensus scoring methods have previously attempted to address this problem by combining scoring functions from different computational docking programs (Feher 2006). The limited success of these attempts may be due to bias in the individual scoring functions that were used.

We hypothesize that by using consensus methods to combine docking or rescoring results from isofunctional proteins, random error in the consensus will be reduced and true binders may be separated more effectively from false positive compounds that do not bind. A basic underlying assumption here is that random error is a significant source of error in docking results.

In this study, we use consensus methods to combine docking and rescoring results of homologous "neighbor" sequences from the enolase superfamily, where neighbors are defined by sequence similarity cutoffs. These methods could also be applied to proteins that are adjacent in the same metabolic pathway, where the product of the protein that is upstream in the pathway is the substrate of the downstream protein. Adjacent pathway proteins may have different folds, but they have evolved to bind chemically similar substrates.

We selected the enolase superfamily as a test case because the superfamily is mechanistically diverse and has been characterized extensively by our collaborators in the laboratory of Dr. Patricia Babbitt, UCSF (Babbitt et al. 1996). Enzymes in the enolase superfamily are related by their ability to catalyze abstraction of a proton alpha to a carboxylic acid group to form an enolic intermediate. Although all of the enzymes in the superfamily share this common partial reaction, their overall reactions, which include cycloisomerization, ß-elimination of water, ß-elimination of ammonia, and racemization, are quite diverse.

Methods

283 proteins from the dipeptide epimerase, MLE I, and MLE II families of the MLE subgroup of the enolase superfamily were modeled using the Protein Local Optimization Program (PLOP, Dr. Matthew Jacobson, UCSF). Using the docking program Glide (Schrödinger, Inc.), a computational library of 500 small molecule metabolites including dipeptides, N-succinylated amino acids, mono- and dicarboxylic acids, deoxyacids and uronic acids was docked against the models. The docking results were rescored using PLOP, which uses a physics based scoring function, to compute the binding energy of each ligand. The docking and rescoring results in this study were provided courtesy of Dr. Chakrapani Kalyanaraman, UCSF.

BLAST 2 Sequences was used to construct a matrix of expectation values for all sequence pairs. Homologous "neighbor" sequences were selected for each sequence at BLAST E-value thresholds of 1e^{-20}, 1e^{-30}, 1e^{-40}, 1e^{-50}, 1e^{-60}, 1e^{-70}, 1e^{-80}, 1e^{-90}, 1e^{-100}, 1e^{-110}, 1e^{-120}, 1e^{-130}, 1e^{-140}, 1e^{-150}, 1e^{-160}, 1e^{-170}, 1e^{-180}, and 1e^{-190}.

The docking or rescoring results for each sequence and its neighbors were combined to create a consensus using two methods, rank-by-rank, in which ligands are ranked according to their average rank across the docking hit lists of isofunctional proteins, and rank-by-number, in which ligands are ranked according to the average of their normalized docking scores. A measure of the similarity between rank lists, Spearman's rank correlation, was used to estimate the strength of the consensus.

In the rank-by-rank method, a ligand's mean rank $\bar{r}$ is computed as

$$\bar{r} = \frac{1}{n}\sum_{i=1}^{n} r_i ,$$

where n is the number of lists in which the ligand is ranked and r_i is the ligand's rank in each list. The ligands are re-ranked according to their mean ranks to create a consensus.

In the rank-by-number method, the raw docking or rescoring scores in each hit list are normalized, the mean normalized score is computed for each ligand, and the ligands are re-ranked according to their mean normalized scores. A ligand's normalized score $\hat{z}_i$ is computed as

$$\hat{z}_i = \frac{z_i - \bar{z}}{\sigma} .$$

Here, z_i is the ligand's raw score, $\bar{z}$ is the mean of the scores for all of the ligands in the hit list, and σ is the standard deviation of the scores. The mean score $\bar{z}$ is calculated as

$$\bar{z} = \frac{1}{n}\sum_{i=1}^{n} z_i ,$$

where z_i is the raw score of each ligand in the list and n is the number of ligands. The standard deviation σ is computed as

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (z_i - \bar{z})^2} .$$

After normalizing the scores in each list, a ligand's mean normalized score $\bar{\hat{z}}_i$ is computed as

$$\bar{\hat{z}}_i = \frac{1}{n}\sum_{i=1}^{n} \hat{z}_i ,$$

where n is the number of lists in which the ligand is ranked and $\hat{z}_i$ is the ligand's normalized score in each list. The ligands are re-ranked according to their mean normalized scores to create a consensus.

Spearman's rank correlation coefficient ρ, which measures the correlation between a pair of rank lists, was used to compare the docking or rescoring hit list of a sequence to the hit lists of each of its neighbors. Spearman's ρ is computed as

$$\rho = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}.$$

Here, d_i is the difference between a ligand's ranks in two lists that are being compared and n is the number of ligands that the lists have in common. The sign of ρ indicates the direction of the correlation, and the magnitude of the correlation is estimated by ρ^2. The statistical significance of ρ was determined using the Student's t-test.

An algorithm was developed to automate these computations (Consensus.py, Appendix B). The algorithm's inputs consist of docking or rescoring results for a set of homologous sequences; an MSF format multiple sequence alignment containing the sequences; and a BLAST E-value threshold, or list of thresholds, for selection of homologous neighbor sequences for consensus scoring. The algorithm works as follows:

1. Sequences corresponding to the docking or rescoring results are extracted from the multiple sequence alignment.
2. BLAST 2 Sequences is used to align all sequence pairs, and an E-value is computed for each alignment.
3. At each E-value threshold for consensus scoring that was provided as input, homologous neighbor sequences are selected for each sequence.

4. For each sequence, the docking or rescoring results of the sequence and its neighbors are combined using both methods, rank-by-rank and rank-by-number, to create a consensus.
5. To estimate the strength of the consensus, Spearman's rank correlation coefficient is computed for each sequence and each of its neighbors.

A sample of the algorithm's output is included (Appendix C).

Results

From the dataset of 283 MLE subgroup sequences, 17 sequences having experimentally determined functions or highly significant BLAST E-values ($\leq 1e^{-175}$) to experimentally characterized sequences were identified (Brown et al. 2006). The substrates of these sequences include dipeptides such as Ala-Glu, Ala-Ala, and Lys-His (dipeptide epimerase family); N-succinylated amino acids such as N-succinyl-L-Arg (NAAAR/OSBS family); and *cis-cis*-muconate (MLE I family). For each sequence, the rank of the native substrate was determined before and after consensus scoring. The mean change in substrate rank for all 17 sequences was computed at each E-value threshold. Figures 4-1 and 4-2 show the mean change in native substrate rank at specific BLAST E-value thresholds. Figure 4-3 shows the relationship between the BLAST E-value threshold and the mean number of homologous neighbor sequences, which increases exponentially with the log of the E-value threshold. Figure 4-4 shows the consensus of docking results for one specific sequence, MLE I from *Pseudomonas putida* (GI: 1633161), using the rank-by-number method.

Discussion

The mean change in native substrate rank was negative at all E-value thresholds for docking and rescoring results using both consensus methods, rank-by-rank and rank-by-number (Figures 4-1 and 4-2). This indicates that both consensus methods tend to increase overall error in the data that was analyzed. Although the mean rank of the native substrate was lower after consensus scoring, there were specific cases in which consensus scoring caused the native substrate rank to improve. For example, applying the rank-by-number method to the docking results of MLE I from *Pseudomonas putida* (GI: 1633161) caused the rank of the native substrate, *cis-cis*-muconate, to improve (Figure 4-4). The improvement occurred at E-value thresholds from 1e^{-160} to 1e^{-190}, which corresponds to between 12 and 14 homologous neighbor sequences. At E-value thresholds > 1e^{-160}, the number of neighbors increases rapidly and the rank of *cis-cis*-muconate decreases.

On average, the rank-by-number method performed better than the rank-by-rank method (Figures 1 and 2). This result is consistent with the idealized computer experiment of Wang and Wang, which simulates consensus scoring with multiple scoring functions (2006). In the simulation, rank-by-number performed slightly ahead of rank-by-rank, and both methods outperformed a third strategy, rank-by-vote, in which each compound is ranked by the number of scoring functions that place it above some threshold in the ligand database. Wang and Wang postulate that rank-by-vote performs poorly because of the strategy's semiquantitative nature; if *n* scoring functions are combined to create a consensus, the compounds will be divided into *n* + 1 bins. This classification system is coarse-grained and results in loss of quantitative information. Similarly, rank-by-rank may not perform as well as rank-by-number because some quantitative information is lost when raw scores are converted into ranks and compounds are ranked according to mean ranks rather than mean normalized scores.

The best results were usually observed at more stringent BLAST E-value thresholds from $1e^{-190}$ to $1e^{-130}$, which correspond to between 6.5 and 13.3 mean neighbors per sequence. Native substrate rank decreased at less stringent E-value thresholds $> 1e^{-130}$. The mean change in native substrate rank is inversely related to the log of the BLAST E-value threshold, but this relationship appears to be complex and nonlinear. There may be a number of factors that influence the consensus result.

1. *Number of isofunctional sequences.* As the number of isofunctional sequences in the consensus increases, the rank of the native substrate is expected to improve. This is based on a simple statistical reason: the mean value of repeated samplings tends to be closer to the true value. Therefore, as the number of isofunctional sequences increases, random error in the consensus should decrease. Interestingly, the rate of improvement is predicted to slow significantly after three or four observations (Wang and Wang 2006). In other words, there is diminishing marginal benefit to each additional docking or rescoring result that is included in the consensus; this implies that the benefit of increasing numbers of isofunctional sequences may be limited.
2. *Isofunctional versus non-isofunctional sequences.* At less stringent thresholds for neighbor selection, non-isofunctional sequences may be included in the consensus. Non-isofunctional sequences have different preferences for substrate and may decrease the rank of the native substrate when they are added to the consensus.

3. *Random versus systematic errors*. Systematic errors such as scoring function bias, flips of asparagine or glutamine residues in crystal structures, inaccurate positioning of active site residues in models, and assignment of erroneous charges or protonation states are common sources of error in docking results. Consensus scoring decreases random error that falls within a normal distribution; however, it fails to address systematic errors. When docking results with significant systematic errors are included in the consensus, the rank of the native substrate may decrease instead of increasing.

The first two factors—the limited benefit of including large numbers of isofunctional sequences in the consensus, and the cost of including non-isofunctional sequences—may explain the decrease in performance at BLAST E-value thresholds < $1e^{-130}$. The third factor, the introduction of systematic errors into the consensus, may explain the poor overall performance that was observed. For many of the sequences in our dataset, the native substrate ranked below the top 20% of the database, indicating that systematic errors were abundant in the docking and rescoring results. Thus, in this case, any benefit from reducing random error appears to have been overwhelmed by the cost of introducing systematic errors into the consensus (Figures 1 and 2). It may be important to begin with a relatively clean set of docking or rescoring results, free from most systematic errors, before applying consensus methods.

In the future, I would like to repeat this experiment using a cleaner dataset, where the native substrate ranks near the top of the database for most sequences. This may allow the benefit of reducing random error, which was observed in individual cases, to become apparent for more proteins. I would also like to apply consensus methods to proteins that are adjacent in metabolic pathways, which may have different folds but have evolved to bind chemically similar substrates. Further study is needed to determine the importance of random error versus systematic error in docking and rescoring results.

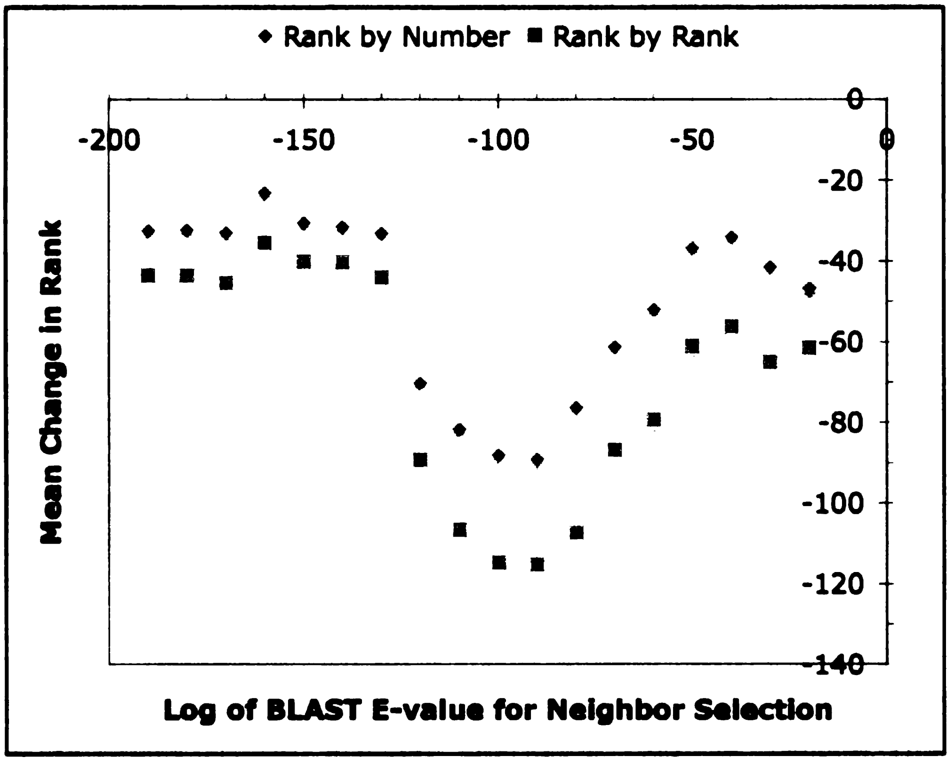


Figure 4-1. Consensus of docking results. The mean change in native substrate rank is plotted against the logarithm of the BLAST E-value threshold for neighbor selection.

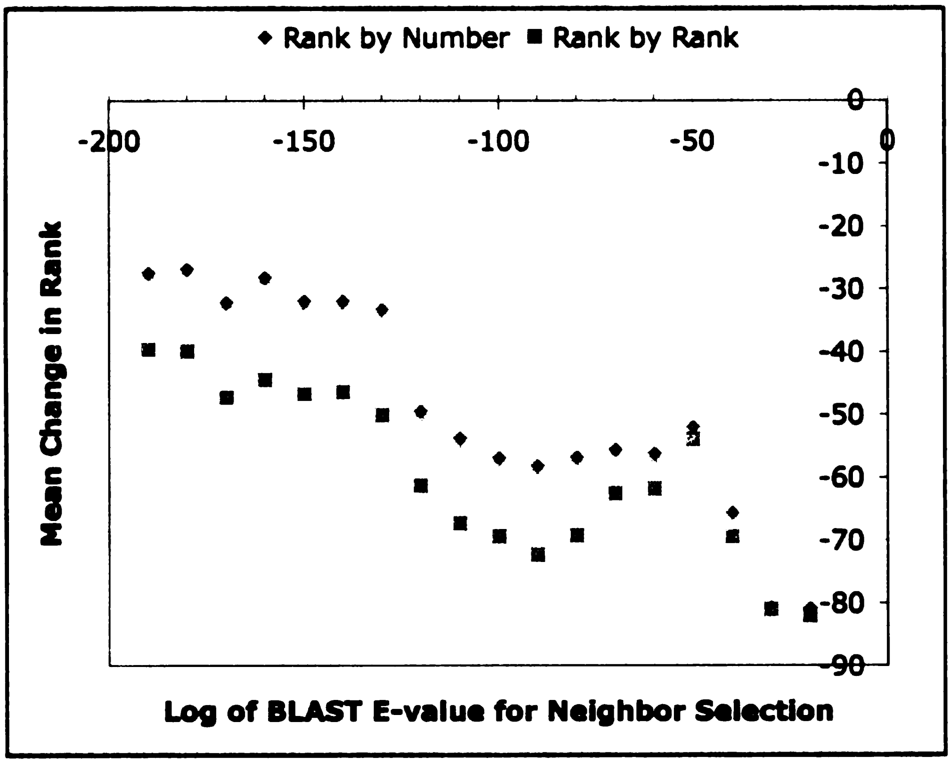


Figure 4-2. Consensus of rescoring results. The mean change in native substrate rank is plotted against the logarithm of the BLAST E-value threshold for neighbor selection.

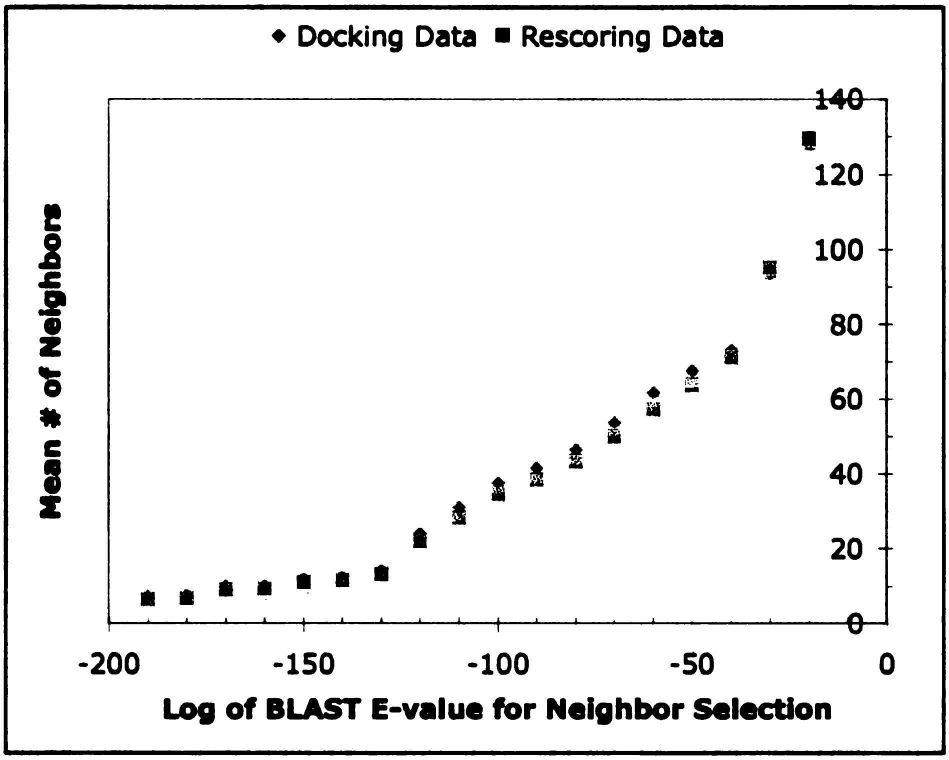


Figure 4-3. Mean number of neighbors. The mean number of homologous neighbor sequences is plotted as a function of the threshold for neighbor selection.

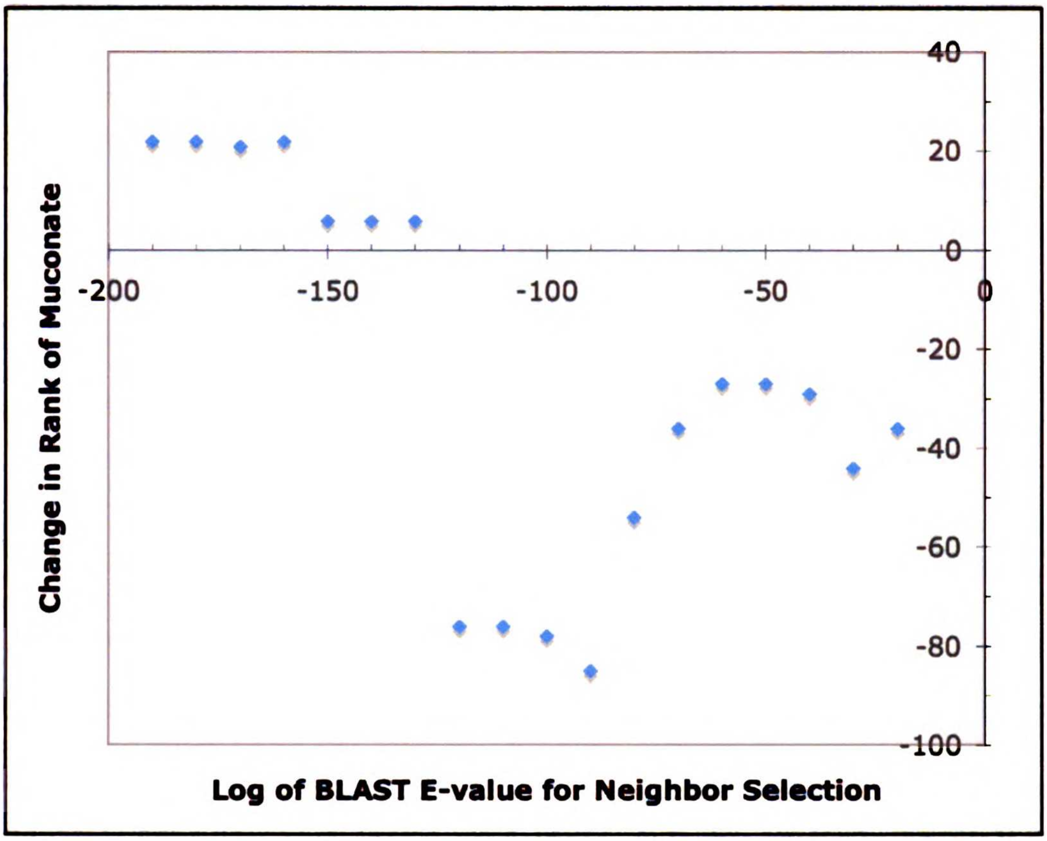


Figure 4-4. Consensus of docking results for *Pseudomonas putida* MLE I (GI: 1633161). The change in rank of *cis-cis*-muconate is plotted against the threshold for neighbor selection. The rank-by-number method was used to compute the consensus result.

References

Introduction

1. Lee, D., Redfern, O. & Orengo, C. (2007). Predicting protein function from sequence and structure. *Nat Rev Mol Cell Biol* **8**, 995-1005.

Chapter 1

1. Ycas, M. (1974). On earlier states of the biochemical system. *J Theor Biol* **44,** 145-60.

2. Jensen, R. A. (1976). Enzyme recruitment in evolution of new function. *Annu Rev Microbiol* **30**, 409-25.

3. O'Brien, P. J. & Herschlag, D. (1999). Catalytic promiscuity and the evolution of new enzymatic activities. *Chem Biol* **6**, R91-R105.

4. Gerlt, J. A. & Babbitt, P. C. (2001). Divergent evolution of enzymatic function: mechanistically diverse superfamilies and functionally distinct suprafamilies. *Annu Rev Biochem* **70**, 209-46.

5. Matsumura, I. & Ellington, A. D. (2001). In vitro evolution of beta-glucuronidase into a beta-galactosidase proceeds through non-specific intermediates. *J Mol Biol* **305**, 331-9.

6. Schmidt, D. M. Z., Mundorff, E. C., Dojka, M., Bermudez, E., Ness, J. E., Govindarajan, S., Babbitt, P. C., Minshull, J. & Gerlt, J. A. (2003). Evolutionary potential of $(\beta/\alpha)_8$-barrels: functional promiscuity produced by single substitutions in the enolase superfamily. *Biochemistry* **42**, 8387-8393.

7. Copley, S. D. (2003). Enzymes with extra talents: moonlighting functions and catalytic promiscuity. *Curr Opin Chem Biol* **7**, 265-72.

8. Hughes, A. L. (1994). The evolution of functionally novel proteins after gene duplication. *Proc Biol Sci* **256**, 119-24.

9. Schultes, E. A. & Bartel, D. P. (2000). One sequence, two ribozymes: implications for the emergence of new ribozyme folds. *Science* **289**, 448-52.

10. Aharoni, A., Gaidukov, L., Khersonsky, O., Mc, Q. G. S., Roodveldt, C. & Tawfik, D. S. (2005). The 'evolvability' of promiscuous protein functions. *Nat Genet* **37**, 73-6. Epub 2004 Nov 28.

11. James, L. C. & Tawfik, D. S. (2003). Conformational diversity and protein evolution--a 60-year-old hypothesis revisited. *Trends Biochem Sci* **28**, 361-8.

12. Gerlt, J. A., Babbitt, P. C. & Rayment, I. (2005). Divergent evolution in the enolase superfamily: the interplay of mechanism and specificity. *Arch Biochem Biophys* **433**, 59-70.

13. Lebioda, L. & Stec, B. (1988). Crystal structure of enolase indicates that enolase and pyruvate kinase evolved from a common ancestor. *Nature* **333**, 683-6.

14. Neidhart, D. J., Howell, P. L., Petsko, G. A., Powers, V. M., Li, R. S., Kenyon, G. L. & Gerlt, J. A. (1991). Mechanism of the reaction catalyzed by mandelate racemase. 2. Crystal structure of mandelate racemase at 2.5-A resolution: identification of the active site and possible catalytic residues. *Biochemistry* **30**, 9264-73.

15. Landro, J. A., Gerlt, J. A., Kozarich, J. W., Koo, C. W., Shah, V. J., Kenyon, G. L., Neidhart, D. J., Fujita, S. & Petsko, G. A. (1994). The role of lysine 166 in the mechanism of mandelate racemase from Pseudomonas putida: mechanistic and

crystallographic evidence for stereospecific alkylation by (R)-alpha-phenylglycidate. *Biochemistry* **33**, 635-43.

16. Wedekind, J. E., Poyner, R. R., Reed, G. H. & Rayment, I. (1994). Chelation of serine 39 to Mg2+ latches a gate at the active site of enolase: structure of the bis(Mg2+) complex of yeast enolase and the intermediate analog phosphonoacetohydroxamate at 2.1-A resolution. *Biochemistry* **33**, 9333-42.

17. Thompson, T. B., Garrett, J. B., Taylor, E. A., Meganathan, R., Gerlt, J. A. & Rayment, I. (2000). Evolution of enzymatic activity in the enolase superfamily: structure of o-succinylbenzoate synthase from Escherichia coli in complex with Mg2+ and o-succinylbenzoate. *Biochemistry* **39**, 10662-76.

18. Gulick, A. M., Hubbard, B. K., Gerlt, J. A. & Rayment, I. (2000). Evolution of enzymatic activities in the enolase superfamily: crystallographic and mutagenesis studies of the reaction catalyzed by D-glucarate dehydratase from Escherichia coli. *Biochemistry* **39**, 4590-602.

19. Babbitt, P. C., Hasson, M. S., Wedekind, J. E., Palmer, D. R., Barrett, W. C., Reed, G. H., Rayment, I., Ringe, D., Kenyon, G. L. & Gerlt, J. A. (1996). The enolase superfamily: a general strategy for enzyme-catalyzed abstraction of the alpha-protons of carboxylic acids. *Biochemistry* **35**, 16489-501.

20. Damien Devos, A. V. (2000). Practical limits of function prediction. *Proteins: Structure, Function, and Genetics* **41**, 98-107.

21. Wilson, C. A., Kreychman, J. & Gerstein, M. (2000). Assessing annotation transfer for genomics: quantifying the relations between protein sequence,

structure and function through traditional and probabilistic scores. *J of Mol Biol* **297**, 233-249.

22. Todd, A. E., Orengo, C. A. & Thornton, J. M. (2001). Evolution of function in protein superfamilies, from a structural perspective. *J Mol Biol* **307**, 1113-43.

23. Rost, B. (2002). Enzyme function less conserved than anticipated. *J Mol Biol* **318**, 595-608.

24. Tian, W. & Skolnick, J. (2003). How well is enzyme function conserved as a function of pairwise sequence identity? *J Mol Biol* **333**, 863-82.

25. Palmer, D. R., Garrett, J. B., Sharma, V., Meganathan, R., Babbitt, P. C. & Gerlt, J. A. (1999). Unexpected divergence of enzyme function and sequence: "N-acylamino acid racemase" is o-succinylbenzoate synthase. *Biochemistry* **38**, 4252-8.

26. Taylor Ringia, E. A., Garrett, J. B., Thoden, J. B., Holden, H. M., Rayment, I. & Gerlt, J. A. (2004). Evolution of enzymatic activity in the enolase superfamily: functional studies of the promiscuous o-succinylbenzoate synthase from Amycolatopsis. *Biochemistry* **43**, 224-9.

27. Meganathan, R. (2001). Biosynthesis of menaquinone (vitamin K2) and ubiquinone (coenzyme Q): a perspective on enzymatic mechanisms. *Vitam Horm* **61**, 173-218.

28. Collins, M. D. & Jones, D. (1981). Distribution of isoprenoid quinone structural types in bacteria and their taxonomic implication. *Microbiol Rev* **45**, 316-54.

29. Teichmann, S. A., Rison, S. C., Thornton, J. M., Riley, M., Gough, J. & Chothia, C. (2001). The evolution and structural anatomy of the small molecule metabolic pathways in Escherichia coli. *J Mol Biol* **311**, 693-708.

30. Holt, J. G., Ed. (1984). Bergey's Manual of Systematic Bacteriology. 1st edit. Vol. 1. Baltimore: Williams & Wilkins.

31. Pegg, S. C., Brown, S., Ojha, S., Huang, C. C., Ferrin, T. E. & Babbitt, P. C. (2005). Representing structure-function relationships in mechanistically diverse enzyme superfamilies. *Pac Symp Biocomput*, 358-69.

32. Pegg, S. C., Brown, S. D., Ojha, S., Seffernick, J., Meng, E. C., Morris, J. H., Chang, P. J., Huang, C. C., Ferrin, T. E. & Babbitt, P. C. (2006). Leveraging Enzyme Structure-Function Relationships for Functional Inference and Experimental Design: The Structure-Function Linkage Database. *Biochemistry* **45**, 2545-2555.

33. Doolittle, W. F. (1999). Phylogenetic classification and the universal tree. *Science* **284**, 2124-9.

34. Teichmann, S. A. & Mitchison, G. (1999). Is there a phylogenetic signal in prokaryote proteins? *J Mol Evol* **49**, 98-107.

35. Gribaldo, S. & Philippe, H. (2002). Ancient phylogenetic relationships. *Theor Popul Biol* **61**, 391-408.

36. Delsuc, F., Brinkmann, H. & Philippe, H. (2005). Phylogenomics and the reconstruction of the tree of life. *Nat Rev Genet* **6**, 361-75.

37. Garrity, G. M., Bell, J. A. & Lilburn, T. G. (2004). Taxonomic Outline of the Prokaryotes. *Bergey's Manual of Systematic Bacteriology*.

38. Wolf, Y. I., Rogozin, I. B., Grishin, N. V., Tatusov, R. L. & Koonin, E. V. (2001). Genome trees constructed using five different approaches suggest new major bacterial clades. *BMC Evol Biol* **1**, 8.

39. Brown, J. R., Douady, C. J., Italia, M. J., Marshall, W. E. & Stanhope, M. J. (2001). Universal trees based on large combined protein sequence data sets. *Nat Genet* **28**, 281-5.

40. Battistuzzi, F. U., Feijao, A. & Hedges, S. B. (2004). A genomic timescale of prokaryote evolution: insights into the origin of methanogenesis, phototrophy, and the colonization of land. *BMC Evol Biol* **4**, 44.

41. Yang, S., Doolittle, R. F. & Bourne, P. E. (2005). Phylogeny determined by protein domain content. *Proc Natl Acad Sci U S A* **102**, 373-8. Epub 2005 Jan 3.

42. Deeds, E. J., Hennessey, H. & Shakhnovich, E. I. (2005). Prokaryotic phylogenies inferred from protein structural domains. *Genome Res* **15**, 393-402.

43. Gophna, U., Doolittle, W. F. & Charlebois, R. L. (2005). Weighted genome trees: refinements and applications. *J Bacteriol* **187**, 1305-16.

44. Klenchin, V. A., Taylor Ringia, E. A., Gerlt, J. A. & Rayment, I. (2003). Evolution of enzymatic activity in the enolase superfamily: structural and mutagenic studies of the mechanism of the reaction catalyzed by o-succinylbenzoate synthase from Escherichia coli. *Biochemistry* **42**, 14427-33.

45. Thoden, J. B., Taylor Ringia, E. A., Garrett, J. B., Gerlt, J. A., Holden, H. M. & Rayment, I. (2004). Evolution of enzymatic activity in the enolase superfamily: structural studies of the promiscuous o-succinylbenzoate synthase from Amycolatopsis. *Biochemistry* **43**, 5716-27.

46. Taylor, E. A., Palmer, D. R. & Gerlt, J. A. (2001). The lesser "burden borne" by o-succinylbenzoate synthase: an "easy" reaction involving a carboxylate carbon acid. *J Am Chem Soc* **123**, 5824-5.

47. Duquerroy, S., Camus, C. & Janin, J. (1995). X-ray structure and catalytic mechanism of lobster enolase. *Biochemistry* **34**, 12513-23.

48. Kuhnel, K. & Luisi, B. F. (2001). Crystal structure of the Escherichia coli RNA degradosome component enolase. *J Mol Biol* **313**, 583-92.

49. Chai, G., Brewer, J. M., Lovelace, L. L., Aoki, T., Minor, W. & Lebioda, L. (2004). Expression, purification and the 1.8 angstroms resolution crystal structure of human neuron specific enolase. *J Mol Biol* **341**, 1015-21.

50. da Silva Giotto, M. T., Hannaert, V., Vertommen, D., de A. S. Navarro, M. V., Rider, M. H., Michels, P. A., Garratt, R. C. & Rigden, D. J. (2003). The crystal structure of Trypanosoma brucei enolase: visualisation of the inhibitory metal binding site III and potential as target for selective, irreversible inhibition. *J Mol Biol* **331**, 653-65.

51. Hosaka, T., Meguro, T., Yamato, I. & Shirakihara, Y. (2003). Crystal structure of Enterococcus hirae enolase at 2.8 A resolution. *J Biochem (Tokyo)* **133**, 817-23.

52. Stubbe, J. (2000). Ribonucleotide reductases: the link between an RNA and a DNA world? *Curr Opin Struct Biol.* **10**, 731-6.

53. Kolberg, M., Strand, K. R., Graff, P. & Andersson, K. K. (2004). Structure, function, and mechanism of ribonucleotide reductases. *Biochim Biophys Acta* **1699**, 1-34.

54. Sakai, A., Xiang, D. F., Xu, C., Song, L., Yew, W. S., Raushel, F. M. & Gerlt, J. A. (2006). Evolution of Enzymatic Activities in the Enolase Superfamily: N-Succinylamino Acid Racemase and a New Pathway for the Irreversible Conversion of D- to L-Amino Acids. *submitted.*

55. Helin, S., Kahn, P. C., Guha, B. L., Mallows, D. G. & Goldman, A. (1995). The refined X-ray structure of muconate lactonizing enzyme from Pseudomonas putida PRS2000 at 1.85 A resolution. *J Mol Biol* **254**, 918-41.

56. Minshull, J., Ness, J. E., Gustafsson, C. & Govindarajan, S. (2005). Predicting enzyme function from protein sequence. *Curr Opin Chem Biol* **9**, 202-9.

57. Glasner, M. E., Gerlt, J. A. & Babbitt, P. C. (in press). Mechanisms of Protein Evolution and Their Application to Protein Engineering. In *Advances in Enzymology and Related Areas of Molecular Biology* (Toone, E., ed.). Wiley & Sons.

58. Ochman, H., Lawrence, J. G. & Groisman, E. A. (2000). Lateral gene transfer and the nature of bacterial innovation. *Nature* **405**, 299-304.

59. Koonin, E. V., Makarova, K. S. & Aravind, L. (2001). Horizontal gene transfer in prokaryotes: quantification and classification. *Annu Rev Microbiol* **55**, 709-42.

60. Garcia-Vallve, S., Romeu, A. & Palau, J. (2000). Horizontal gene transfer in bacterial and archaeal complete genomes. *Genome Res* **10**, 1719-25.

61. Doolittle, W. F., Boucher, Y., Nesbo, C. L., Douady, C. J., Andersson, J. O. & Roger, A. J. (2003). How big is the iceberg of which organellar genes in nuclear genomes are but the tip? *Philos Trans R Soc Lond B Biol Sci* **358**, 39-57; discussion 57-8.

62. Brown, J. R. (2003). Ancient horizontal gene transfer. *Nat Rev Genet* **4**, 121-32.

63. Philippe, H. & Douady, C. J. (2003). Horizontal gene transfer and phylogenetics. *Curr Opin Microbiol* **6**, 498-505.

64. Kennedy, S. P., Ng, W. V., Salzberg, S. L., Hood, L. & DasSarma, S. (2001). Understanding the Adaptation of Halobacterium Species NRC-1 to Its Extreme Environment through Computational Analysis of Its Genome Sequence. *Genome Res.* **11**, 1641-1650.

65. Overbeek, R., Disz, T. & Stevens, R. (2004). The SEED: a peer-to-peer environment for genome annotation. *Communications of the ACM* **47**, 46-5`.

66. Meganathan, R., Bentley, R. & Taber, H. (1981). Identification of Bacillus subtilis men mutants which lack O-succinylbenzoyl-coenzyme A synthetase and dihydroxynaphthoate synthase. *J Bacteriol* **145**, 328-32.

67. Taber, H. W., Dellers, E. A. & Lombardo, L. R. (1981). Menaquinone biosynthesis in Bacillus subtilis: isolation of men mutants and evidence for clustering of men genes. *J Bacteriol* **145**, 321-7.

68. Driscoll, J. R. & Taber, H. W. (1992). Sequence organization and regulation of the Bacillus subtilis menBE operon. *J Bacteriol* **174**, 5063-71.

69. Rowland, B., Hill, K., Miller, P., Driscoll, J. & Taber, H. (1995). Structural organization of a Bacillus subtilis operon encoding menaquinone biosynthetic enzymes. *Gene* **167**, 105-9.

70. Koike-Takeshita, A., Koyama, T. & Ogura, K. (1997). Identification of a novel gene cluster participating in menaquinone (vitamin K2) biosynthesis. Cloning and sequence determination of the 2-heptaprenyl-1,4-naphthoquinone

methyltransferase gene of Bacillus stearothermophilus. *J Biol Chem* **272**, 12380-3.

71. Johnson, T. W., Shen, G., Zybailov, B., Kolling, D., Reategui, R., Beauparlant, S., Vassiliev, I. R., Bryant, D. A., Jones, A. D., Golbeck, J. H. & Chitnis, P. R. (2000). Recruitment of a foreign quinone into the A(1) site of photosystem I. I. Genetic and physiological characterization of phylloquinone biosynthetic pathway mutants in Synechocystis sp. pcc 6803. *J Biol Chem* **275**, 8523-30.

72. Pegg, S. C. & Babbitt, P. C. (1999). Shotgun: getting more from sequence similarity searches. *Bioinformatics* **15**, 729-40.

73. Gish, W. (1994-1997). BLASTP 2.0a19MP-WashU. In *unpublished.* unpublished.

74. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**, 3389-402.

75. Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucl. Acids Res.* **32**, 1792-1797.

76. Jewett, A. I., Huang, C. C. & Ferrin, T. E. (2003). MINRMS: an efficient algorithm for determining protein structure similarity using root-mean-squared-distance. *Bioinformatics* **19**, 625-34.

77. Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C. & Ferrin, T. E. (2004). UCSF Chimera--a visualization system for exploratory research and analysis. *J Comput Chem* **25**, 1605-12.

78. Ronquist, F. & Huelsenbeck, J. P. (2003). MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* **19**, 1572-4.

79. Altekar, G., Dwarkadas, S., Huelsenbeck, J. P. & Ronquist, F. (2004). Parallel Metropolis coupled Markov chain Monte Carlo for Bayesian phylogenetic inference. *Bioinformatics* **20**, 407-15. Epub 2004 Jan 22.

80. Whelan, S. & Goldman, N. (2001). A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Mol Biol Evol* **18**, 691-9.

81. Felsenstein, J. (2004). PHYLIP (Phylogeny Inference Package) version 3.6. *Distributed by the author. Department of Genome Sciences, University of Washington, Seattle.*

82. Jones, D. T., Taylor, W. R. & Thornton, J. M. (1992). The rapid generation of mutation data matrices from protein sequences. *Comput Appl Biosci* **8**, 275-82.

83. Shindyalov, I. N. & Bourne, P. E. (1998). Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Eng* **11**, 739-47.

84. Myers, E. W. & Miller, W. (1988). Optimal alignments in linear space. *Comput Appl Biosci* **4**, 11-7.

85. Venter, J. C., Remington, K., Heidelberg, J. F., Halpern, A. L., Rusch, D., Eisen, J. A., Wu, D., Paulsen, I., Nelson, K. E., Nelson, W., Fouts, D. E., Levy, S., Knap, A. H., Lomas, M. W., Nealson, K., White, O., Peterson, J., Hoffman, J., Parsons, R., Baden-Tillson, H., Pfannkoch, C., Rogers, Y. H. & Smith, H. O. (2004). Environmental genome shotgun sequencing of the Sargasso Sea. *Science* **304**, 66-74.

Chapter 2

1. Gu, C., Zhou, W., Puthenveedu, M. A., Xu, M., Jan, Y. N., and Jan, L. Y. (2006). The microtubule plus-end tracking protein EB1 is required for Kv1 voltage-gated K+ channel axonal targeting. *Neuron* **52**, 803-16.

2. Gulbis, J. M., Mann, S., MacKinnon, R. (1999). Structure of a voltage-dependent K+ channel beta subunit. *Cell* **97**, 943-52.

3. Hille, B. (1992). Ionic Channels of Excitable Membranes (Sunderland, MA: Sinauer Associates, Inc.).

4. Jez, J. M., Bennett, M. J., Schlegel, B. P., Lewis, M., and Penning, T. M. (1997). Comparative anatomy of the aldo-keto reductase superfamily. *Biochem J* **326**, 625-36.

5. Tempel, B. L., Papazian, D. M., Schwarz, T. L., Jan, Y. N., and Jan, L. Y. (1987). Sequence of a probable potassium channel component encoded at Shaker locus of Drosophila. *Science* **237**, 770-5.

6. Weng, J., Cao, Y., Moss, N., and Zhou, M. (2006). Modulation of voltage-dependent Shaker family potassium channels by an aldo-keto reductase. *J Biol Chem* **281**, 15194-200.

Chapter 4

1. Babbitt, P. C., Hasson, M. S., Wedekind, J. E., Palmer, D. R., Barrett, W. C., Reed, G. H., Rayment, I., Ringe, D., Kenyon, G. L., Gerlt, J. A. (1996). The enolase superfamily: a general strategy for enzyme-catalyzed abstraction of the alpha-protons of carboxylic acids. *Biochem* **35**, 16489-501.

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

3. Feher, M. (2006). Consensus scoring for protein-ligand interactions. *Drug Discov Today* **11**, 421-8.

4. Wang, R., Wang, S. (2001). How does consensus scoring work for virtual library screening? An idealized computer experiment. *J Chem Inf Comput Sci* **41**, 1422-1426.

Appendix A

Draft of Metacaspase Paper

Plasmodium berghei metacaspase 1 is involved in controlling parasite numbers in the mosquito

Alida Coppi[1], Shahid M. Khan[2#], Gunnar R. Mair[2], N. Fayazmanesh[3], M. Jacobson[3], Jannik Fonager[2], M. Bogyo[4], M. Lee[5], Chris J. Janse[2], Andrew P. Waters[2], Photini Sinnis[1#], M. Sajid[5#]

[1]Department of Medical & Molecular Parasitology, 341 E. 25th Street, New York University School of Medicine NY, NY 10010. [2]Leiden University Medical Centre, Department of Parasitology, Albinusdreef 2, 2333 ZA Leiden, Netherlands. [3]UCSF MC 2240 Genentech Hall Room N472C 600 16th Street, San Francisco, CA. [4]Department of Pathology, Stanford University Medical School, Edwards Building Rm R-343, 300 Pasteur Dr. Stanford, CA 94305-5324. [5]Sandler Center for Basic Research in Parasitic Diseases, QB3, 1700 4th Street, Rm 509, University of California San Francisco, SF, CA, 94158-2550.

#corresponding authors

Methods

Parasite culture

The gametocyte producing reference clone, cl15cy1 (HP) of the ANKA strain of P. berghei was used and maintained as previously described. In addition, the non-gametocyte producer clone (HPE) of the ANKA strain was also used.

Characterization of PbMC1

The sequence for full-length cDNA of PbMC1 was deposited at GenBank, accession number AJ555625. Chromosomal location of the *P. berghei* Metacaspase 1 gene, hereby termed *PbMC1*, was determined by standard hybridization using a gene specific probe to pulse-field gel electrophoresis separated chromosomes {Ponzi, 1990 #229}. Transcription of the *PbMC1* was analysed by standard Northern blotting of RNA isolated form synchronized asexual blood stages from HP and HPE parasites and gametocytes {Paton, 1993 #230}.

Phylogeny, sequence comparison and domain structure analysis

Annotated metacaspase genes were downloaded from www.plasmodb.org, www.apidb.org and http://www.ncbi.nlm.nih.gov/Genbank. These were translated using the standard codon table and aligned in ClustalX at www.sacs.ucsf.edu, using default settings for gap opening and extension. For this set, Phylogenetic analysis was performed using the Neighbour Joining method with 1000 bootstrap replicates, as implemented the PHYLIP v.3.66 package {Felsenstein, 1997 #231} (Version 3.2)). Tree drawing was conducted by Mega 3.1 {Kumar, 2004 #232}. The metacaspase domain structures were determined by using the Simple Modular Architecture Research Tool (SMART at www.smart.embl-heidelberg.de)

Resynthesis of PbMC1 (rPbMC1)

The 77% AT rich coding region for PbMC1 was resynthesised by Geneart, Germany (sequence available on request). The codon usage of the GC balanced resynthesised *P. berghei* metacaspase 1 (rPbMC1) was biased for optimal heterologous expression in *Pichia pastoris.*

Mosquito stages

3 to 5 day-old *Anopheles stephensi* mosquitoes were fed on Swiss Webster mice infected with *P. berghei* ANKA strain wild type or metacaspase 1 knockout parasites. On days 10 through 22 post-infective blood meal, mosquitoes were anesthetized on ice, rinsed in 70% ethanol, washed in Dulbecco's Modified Eagle Medium (DMEM), and midguts, hemolymph and salivary glands were harvested for determination of sporozoite numbers in these different compartments. Midguts and salivary glands were homogenized to release sporozoites, centrifuged to remove mosquito debris and sporozoites were counted in a hemocytometer. Hemolymph sporozoites were counted directly without processing. For visualization of GFP-transgenic oocysts infected midguts were removed post-infective blood meal, mounted in PBS and photographed using a Nikon E600 Fluorescence Microscope and a DXM1200 digital camera.

Invasion and development assays

Hepa 1-6 cells (CRL -1830; ATCC, Rockville, MD) were grown in DMEM supplemented with 10% fetal calf serum (FCS) and 1 mM glutamine and seeded in Permanox eight-chambered Lab-Tek wells (2.5 x 10^5/well) and allowed to grow overnight. On the day of the experiment, 5 x 10^5 *P. berghei* wild type or metacaspase knockout sporozoites were added per well. After 1 hr at 37°C, cells were washed, fixed, and sporozoites were stained with a double staining assay that distinguishes intracellular from extracellular sporozoites. To quantify EEF development, cells with sporozoites were grown for an additional 2 days after which they were fixed with methanol, stained with mAb 2E6, directed against *P. berghei* Hsp70 followed by goat anti-mouse Ig conjugated to FITC. In all assays, at least 50 fields per well were counted and each point was performed in triplicate.

Quantification of liver stage burden

4 to 5 week old female Swiss Webster mice were injected intravenously with 10^4 sporozoites. 40 hr later, livers were harvested, total RNA was isolated, and liver parasite burden was quantified by reverse transcription followed by real-time PCR as outlined previously with some modifications. PCR was performed using primers that recognize *P. berghei*-specific sequences within the 18S rRNA and the temperature profile of the real-time PCR was 95°C for 15 min, followed by 40 cycles of 95°C for 30 s, 58°C for 30 s and 72°C for 30 s. Ten-fold dilutions of a plasmid construct containing the *P. berghei* 18S rRNA gene were used to create a standard curve. All *in vivo* data were analyzed using the Student *t* test for unpaired samples.

Generation and immuno-selection of αPbMC1 antibodies

The PbMC1 peptide NFYDSSMNILKLID was synthesised and linked to KLH by Covance, PA, and used to raise antisera in New Zealand White Rabbits using standard methodologies. The fourth bleed was used in all subsequent experiments. The IgG fraction was enriched by loading 8 ml of PD10 (Pharmacia) buffer exchanged serum onto an equilibrated 1ml HiTrap protein-G column (Pharmacia Biotech, NJ) and washed extensively. Buffer exchange and washes were carried out with 25 mM phosphate buffer, pH 7.2. The IgG fraction was eluted with 1.2 ml of 50 mM glycine-HCl, pH 2.7, the eluate was immediately returned to pH 7.2 using 1.0 M phosphate buffer. Preimmune serum control was prepared in tandem.

Immuno-selection of the monovalent antibody preparation was carried out using standard methodologies. Briefly, 500 μg of polyhistidine tagged His-recPbMC1 expressed in *E. coli* (see below) was subjected to SDS-PAGE. Following electrophoresis, proteins were electrotransferred to a PVDF membrane and visualized with 0.1% (v/v) Ponceau-Red (Sigma Co.). The excised His-recPbMC1 was destained with 2mM NaOH and washed in distilled water. The His-recPbMC1 bound PVDF was blocked with 5% (w/v) BSA in PBS containing 0.05% (w/v) Tween 20 (PBS/T) at 4°C for 15hrs, washed extensively with PBS/T prior to incubation with the 500 μl protein-G purified antibody preparation for 15 h at 4°C. Following excessive washing with PBS/T, the bound immunoselected antibodies were eluted in 500 μl of 100 mM glycine-HCl (pH 2.7), and the preparation immediately adjusted to pH 7.2 using 1.0 M phosphate buffer. The monovalant immunoselected antibody preparation (αPbMC1) was used in all subsequent Western blot and IFA studies.

Western blot and IFA

Proteins were denatured and reduced prior to being resolved by SDS-PAGE using pre-cast 4-12% NuPAGE gradient gels (Invitrogen). The gels were either stained with Coomassie Blue or electroblotted onto nitrocellulose (Schleicher and Schuell, NH) or PVDF membranes (Millipore, MA). After transfer, the membranes were blocked with 5% (w/v) bovine serum albumin in PBS/T and incubated for 1 h with 1:400 αPbMC1. After washing with PBS/T, membranes were incubated for a further 1 h with goat anti-rabbit antibodies coupled to horseradish peroxidase (1:4000 in PBS/T; Life Technologies, MD). The membranes were washed extensively with PBS/T prior to incubation in ECL Western Blotting System (Amersham Biosciences). Immunopositive bands were visualized by exposure to X-Omat film or by phosphorimaging using a Typhoon Trio (Amersham Biosciences).

Wild type *P. berghei* salivary gland sporozoites were lysed in nonreducing sample buffer and 5 x 10^4 sporozoite equivalents/lane were loaded and separated by SDS-PAGE, transferred to PVDF membrane and incubated with α-PbMC1 (1:100) followed by anti–rabbit Ig conjugated to horseradish peroxidase (HRP; 1:100,000). Bound antibodies were visualized using the enhanced chemiluminescence detection system.

Expression of rPbMC1

For expression work, a hexahistidine tag was incorporated at the N-terminus of the full length rPBMC1 immediately upstream of the start methionine. Full length rPbMC1 was used as no signal sequence was predicted.

Bacterial expression

BL21(DE3)pLysS strain of *E. coli* containing episomal pET28a/His-rPbMC1 plasmid was grown overnight at 37°C in L-broth containing 50 μg ml^{-1} tetracycline. Routinely, 50 ml of the overnight culture was used to seed 950 ml of *L*-broth containing 50 μg ml^{-1} of tetracycline. The cultures were grown at 37°C until an $O.D._{600nm}$ of 0.6 - 0.8 was achieved. Induction was initiated by the addition of isopropyl-β-*D*-thiogalactopyranoside (Boehringer-Mannheim, IN) to a final concentration of 1 mM and continued for 4 h. The bacteria were then harvested by centrifugation at 4 000 *x g* for 5 min (Sorvall, RC-5B). The pellet obtained was resuspended in 40 ml PBS (1 mM KH_2PO_4, 10 mM Na_2HPO_4, 137 mM NaCl, 2.7 mM KCl, (pH 7.4)) containing 100 μg.ml^{-1} lysozyme (Boehringer-Mannheim, IN) and 5 μg.ml^{-1} DNAase-I (Boehringer-Mannheim, IN) and left on ice for 30 min. The solution was then sonicated with three brief pulses (Fischer, sonic dismembranator model 30; setting at 60), clarified at 4 000 *x g* for 5 min and the supernatant centrifuged at 11 000 *x g* for 60 min. The pellet obtained was resuspended in 20 ml 8.0 M urea, 75 mM Tris-HCl, 10 mM imidazole, 500 mM NaCl (pH 7.2) and left overnight on a flat bed shaker at room temperature prior to centrifugation at 11 000 *x g* for 60 min. The supernatant was mixed with 1.5 ml of Ni^{2+}-charged metal chelating sepharose fast flow (Amersham Biotech). The sepharose beads were extensively washed (3x in the same buffer) prior to elution with 3 ml of 500 mM imidazole in buffer.

Cell free expression

10 μg of the plasmid pET28a/His-rPbMC1 was linearised down stream of the rPbMC1 coding region using SmaI. The linearised plasmid was used to generate single strand transcript using the Ribomax Kit (Promega) and standard methodologies. His-rPbMC1 transcript was added to the nuclease treated T7 TNT rabbit reticulocyte lysate (50 μl total) in the presence of either cold methionine or ^{35}S-methionine. The lysate was incubated at 30°C for 90 min prior to IMAC purification using Ni-NTA (Qiagen) and standard methodologies. Washes were carried out using 50 mM Tris-HCl, 10 mM imidazole, 500 mM NaCl (pH 7.2) and elution in the same buffer with 500 mM imidazole.

Molecular modeling and S1 subsite prediction of PbMC1, AtMC1, LmMC, and TbMC

Alignment of PbMC1 with known metacaspases and caspases was used to identify the core enzymatic domain of PbMC1. Using the protein structure prediction program PRIME (Schrödinger, Inc.), one hundred sequences with structures in the Protein Data Bank were identified as having significant homology to the predicted secondary structure of PbMC1. Human caspase 3 (PDB ID: 1CP3) was selected as the template for modeling PbMC1 due to its high PRIME threading rank (number 3) and because it is also a clan CD enzyme. A BLAST search of the NCBI non-redundant sequence database identified sequences with expectation values better than 1e^{-20} to 1CP3, PbMC1, AtMC1, LmMC, and TbMC. The sequences were filtered at 90% identity to remove redundant sequences, and the multiple sequence alignment program MUSCLE was used to construct a profile-profile alignment of the caspases to the metacaspases.

Homology models of PbMC1, AtMC1, LmMC, and TbMC were generated using the Protein Local Optimization Program (Dr. Matthew P. Jacobson, UCSF). The p17 and p12 heavy and light chains of caspase 3 were modeled as one polypeptide.

Labelling cell free translated His-rPbMC1 product with biotinylated P1 peptidyl-acyloxymethylketones (AOMKs)

5 μl of purified in vitro translated ^{35}S-His-rPbMC1 was made up to 20 μl with 50 mM Tris-HCl pH 7.2, 100 mM NaCl, 0.1% (w/v) NP40, 2 mM DTT, 20 μM biotinylated P1 variable peptidyl-AOMK for 90 min at room temp (P1 residues used were Arg, Lys, Asn, Asp, Phe and Leu). Following incubation the mix was diluted to 1 ml with 100 mM Tris-HCl pH 7.6, 500 mM NaCl, 0.05% (w/v) Triton X100 prior to the addition of 25 μl bed volume of neutravidin-agarose (NA). Following mixing the NA was harvested by centrifugation at 1000 g for 30 sec and washed fives times with 1.5 ml 100 mM Tris-HCl pH 7.8, 500 mM NaCl, 0.05% (w/v) Triton X100. After the final wash the NA, 40 μl 2X reduced protein sample loading buffer was added prior to boiling for 10 min. Following centrifugation the supernatant was resolved by 10-20% SDS-PAGE and the gel visualised by either autoradiography or exposure to phosphorimaging using a Typhoon phosphorimaging system (GE).

Results

There are three metacaspases in Plasmodium

Phylogenic analysis using nearest neighbour/maximum parsimony revealed that metacaspases of parasitic protozoa fall in to three distinct assemblages (Figure 1). Type

I represent the largest group and include PbMC1 and the yeast ScMCA1. Within the Type I MC groups there are cohorts that are exclusively trypanasomatid or solely plasmodial. A third group within the Type I MCs includes the metacaspases from yeast, Acanthamoeba and Toxoplasma. Type II MCs are only found in Plasmodium species, whereas, Type III are restricted to Apicomplexa. Full length PbMC1 cDNA sequence was deposited in Genbank, accession number AJ555625.

Type I, II and III MCs can be defined by respective sequence similarity around the predicted position of the active site catalytic dyad, as deduced by ClustalW pileup analysis with caspase 3 (Figure 2A). ScMCA1 has been shown to be an active enzyme, and it is likely that all Type I MCs would be predicted to be enzymatically active as they all posses the critical active site His and Cys at or preceding the likely position as surmised by pileup analysis. Modelling studies (see below; Figure 6) confirm that a Cys residue preceding the predicted position can indeed spatially overlay the active Cys of caspase 3. Type II MCs are exclusively found in Plasmodium species and may represent the most recent MC due to the high degree of similarity in the homology domain shown. Type II are not predicted to be thiol-proteases as no catalytic Cys is predicted nor is a general base in the vicinity of this homology domain. Interestingly, all the Type II MCs have a Tyr and Thr replacing the catalytic His and Cys, possibly suggesting that these substituted residues are important for biological function. No Tyr as catalytic residue has been reported to date. Type III MCs are restricted to Apicomplexa and may well be active. In addition to a His as the base in Type III MCs, Lys residues are also predicted. A Lys residue can act as a general base so be directly involved in the proteolytic mechanism. TgMC3, PfMC3 and PvMC3 would be predicted to be enzymatically active as they possess a catalytic Cys at or preceding the predicted location. The remainder of the Type III all have a Ser at this predicted site of the catalytic nucleophile. A Ser as a nucleophile is well documented, and there are examples where Ser/His and Ser/Lys are involved in catalysis. Therefore it is plausible that all Type III MCs are active enzymes.

The domain structure of PfMC1, PfMC2 and PfMC3 as type examples of Type I, II and III MCs was carried out by SMART analysis and all are shown to have the clan CD family C14 catalytic domain. PfMC1, like PbMC1 (data not presented) has a predicted N-terminal calcium dependent C2 domain. The C2 domain has been shown to be involved in signal transduction or membrane trafficking through phospholipids binding. C2 domains can associate with membrane lipid moieties such as phosphatidylserine and phosphatidylcholine. Whether the C2 domain of PfMC1 is involved in signal transduction and associates with PCD marker, phosphatidylserine, remains to be shown. PfMC2 contains a histone deacetylase interaction (HDAC) domain which is thought to be involved in DNA stability and integrity. PfMC3 is the largest of the three types at well over 250 kDa and has no additional predicted SMART domains.

PbMC1 is highly transcribed in transmission stages

PbMC1, like PfMC1, is a single copy gene as determined by the completed genome sequencing of P. berghei and P. falciparum, respectively. Consistent with PfMC1 data (Figure 2C) from PlasmoDB, Northern blot analysis reveals that PbMC1 mRNA was also tightly regulated with transcription levels high during the gametocyte and sporozoite stages. No expression is evident during the asexual stages of the lifecycle.

-PbMC1 (is female restricted) is a full length polypeptide and localises to the cytoplasm
Western blot analysis revealed that PbMC1 was expressed as an unprocessed protein of approximately 70 kDa during the gametocyte and ookinete stages (Figure 3B). IFA using αPbMC1 antibodies localise PbMC1 to the cytoplasm of female gametocytes (Figure 3C). Parasites that were transfected with a construct comprised of the upstream region of PbMC1 fused to the coding region of GFP were selected (Figure 4A). GFP expression was exclusively restricted to female gametocytes. No GFP fluorescence was observed in either male gametocytes nor asexual stages (Figure 4B).

Phenotype of metacaspase 1 null parasites in the mosquito host

The demonstration of metacaspase 1 expression in sporozoites suggested that this protein may function during the mosquito stages of the parasite's life cycle. To test this we allowed mosquitoes to feed on mice infected with either wild type or ΔPbMC1 parasites and followed the progression of the cycle in the mosquito. Interestingly, oocyst numbers per midgut had a wider range in ΔPbMC1 parasites compared to wild type. Although there is generally a large range in oocyst numbers among different mosquitoes, ΔPbMC1 parasites consistently gave very high oocyst counts in a proportion of the infected mosquitoes. These high oocyst counts however did not result in higher sporozoite numbers. In fact, there were slightly lower numbers of oocyst sporozoites in ΔPbMC1 parasites compared to controls. Interestingly, numbers of hemolymph sporozoites and salivary gland sporozoites were more significantly decreased in ΔPbMC1 parasites.

We then went on to test the infectivity of ΔPbMC1 sporozoites in the mammalian host. In vitro, using a hepatoma cell line we found that invasion efficiency was similar to wild type sporozoites. In addition, these sporozoites could develop normally to exoerythrocytic stages. In vivo, using an RT-PCR assay to quantify liver stage burden we found that the infectivity of ΔPbMC1 sporozoites was similar to wild type.

PbMC1 null parasites (ΔPbMC1) have increased oocyst number and reduced sporozoite numbers

Selection of ΔPbMC1 parasites was confirmed by western blot analysis, where no PbMC1 protein expression was observed in PbMC1 null gametocyte and ookinetes stages (Figure 4E). ΔPbMC1 parasite generated significantly greater numbers of midgut oocysts compared to wild type, the maxima of which was at day 12 (Figure 5B). Although there were greater numbers of oocysts in the ΔPbMC1, there were fewer sporozoites emerging from these oocysts compared to wild type (Figure 5C). Sporozoites numbers in the hemolyph and salivary glands were also reduced in ΔPbMC1 parasites compared to wild type (Figure 5E). PbMC1 did not have a significant role in sporozoite invasion of hepatocytes (Figure 5F).

A PbMC1 model reveals a hydrophobic S1 subsite

A structural model of PbMC1 using caspase 3 as template was generated (Figure 6A). The model showed significant similarity around the active site, where the active site His and Cys served as anchor residues. The main chain backbone around the S1 subsites shows considerable superimposition. The S1 subsites of caspase 3, the model of PbMC1 and model of A. thaliana AtMC1 were compared (Figure 6B). The AtMC1 has been shown to hydrolyse substrates with Arg at P1 and so AtMC1 was used to increase confidence in the modelling. As expected the AtMC1 predicted S1 subsite was predicted to be negatively charged. The PbMC1 S1 subsite suggested a very different preference at the S1 position, where hydrophobic residues may be preferred. An inhibitor subtractive gametocyte preparation incubated with a number of authentic AMC substrates resulted in some residual activity. In some cases, as in Pro and Phe at P1, this activity was inhibited by alkylation and so demonstrated that a Cys residue was involved catalysis.

Figure Legends

Figure 1

Metacaspases from parasitic protozoa are distributed into three main groups. Type I MCs comprise the largest and most widely distributed and includes the yeast enzyme, ScMCA1. Type II MCs are exclusively plasmodial whereas type III are restricted to the apicomplexa. Orthologues of protozoan parasite metacaspases were obtained from GenBank, ApiDB, PlasmoDB and ToxoDB accession number or gene identifier numbers are shown. Key: AcMC1 (Acanthamoeba castellanii), AAL87229; ChMC3 (Cryptosporidium hominis), XP_665082; CpMC3 (Cryptosporidium parvum), XP_625430; LdMC1 (Leishmania donovani), ABD19717; LdMC2 (Leishmania donovani), ABD19718; LmMC2 (Leishmania major strain Friedlin), XP_843263; PbMC1 (Plasmodium berghei), CAD88480; PbMC1a (Plasmodium berghei strain ANKA), XP_680261; PbMC2 (Plasmodium berghei), CAD88481; PbMC3 (Plasmodium berghei), XP_670090; PcMC1 (Plasmodium chabaudi chabaudi), XP_743169; PcMC2

(Plasmodium chabaudi chabaudi), XP_744101; PcMC3 (Plasmodium chabaudi chabaudi), XP_737372; PfMC1 (Plasmodium falciparum), NP_705432; PfMC2 (Plasmodium falciparum 3D7), NP_702252; PfMC3 (Plasmodium falciparum), NP_702048; PvMC1 (Plasmodium vivax), Pv114725; PvMC2 (Plasmodium vivax), Pv118575; PvMC3 (Plasmodium vivax), Pv085640; PyMC1 (Plasmodium yoelii), XP_725264; PyMC2 (Plasmodium yoelii), XP_726149; PyMC3 (Plasmodium yoelii), XP_725052; TbMC2a (Trypanosoma brucei), CAD24803; TbMC3a (Trypanosoma brucei), CAD24804; TbMC4a (Trypanosoma brucei), CAD24805; TbMC4b (Trypanosoma brucei TREU927), XP_828404; TbMC2 (Trypanosoma brucei TREU927), XP_826272; TbMC3 (Trypanosoma brucei TREU927), XP_826271; TbMC4 (Trypanosoma brucei TREU927), XP_822497; TbMC5 (Trypanosoma brucei TREU927), XP_827616; TbMC5a (Trypanosoma brucei), CAD24806; TcMC3 (Trypanosoma cruzi), AAY84580; TcMC3a (Trypanosoma cruzi), AAY84583; TcMC3b (Trypanosoma cruzi strain CL Brener), XP_805953; TcMC3c, (Trypanosoma cruzi), AAY84581; TcMC3d (Trypanosoma cruzi), AAY84582; TcMC3e (Trypanosoma cruzi strain CL Brener), XP_804238; TcMC3f (Trypanosoma cruzi strain CL Brener), XP_818074; TcMC3g (Trypanosoma cruzi strain CL Brener), XP_810937; TcMC5, (Trypanosoma cruzi strain CL Brener), XP_816130; TcMC5a, (Trypanosoma cruzi strain CL Brener), XP_817238; TgMC1 (Toxoplasma gondii), 20m03956; TgMC3 (Toxoplasma gondii), CB752529; TpMC3 (Theileria parva strain Muguga), XP_764113; TaMC3 (Theileria annulata strain Ankara), XP_953141. The numbers denote bootstrap values.

Figure 2

(A) Comparison of conserved amino acid residues spanning the catalytic residues of apicomplexan parasite MCs reveals that there are three distinct MC types. Like the PbMC1 in this study (shown by an asterix) the yeast enzyme, SmMCA1 is a type I MC. Caspase 3 is used to compare the active site His and Cys, which are indicated by arrows and highlighted by green shading; this reveals that the predicted residue at the position of the catalytic nucleophile of PbMC1 is Pro, however, a Cys residue immediately precedes this. Caspase 3 is included as a distant but related member of the same family of cysteine peptidases. Colours reflect physicochemical similarities, whereas residues that are black share no similarity at that position. Colour key for residues: Green, hydrophobic; light blue, polar but negative; dark blue, acidic; red, basic; purple, polar but positive; yellow, cysteine. (B) Domain structures of PfMC1 (Type I), PfMC2 (Type II) and PfMC3 (Type III) MCs from P. falciparum generated using SMART at www.smart.embl-heidelberg.de. All three MC types have the family C14, clan CD catalytic domain, PfMC1 contains a conserved C2 domain and PfMC2 has a predicted HDAC domain. Scale of polypeptide length is shown. (C) Expression profile of PfMC1, 2 and 3 was compiled using PlasmoDB.

Figure 3

(A) Northern blot analysis of the PbMC1 expression reveals that PbMC1 is transcribed in gametocytes and sporozoites. Top panel is a loading control, the lower panel shows the absence/presence of mRNA for PbMC1. (B) Western blot analysis reveals expression of full length PbMC1 in ookinetes and gametocytes. No processed PbMC1 product is seen. (C) IFA studies using immunoselected anti-PbMC1 antibodies.

Figure 4

Generation of PbMC1 promotor-GFP construct, in vivo read out. Generation of KO construct, selection, and confirmation. (A) Construct of PbMC1 promo-GFP. (B) In vivo Gfp +ve fems. (C) Schematic of construct. (D) Confirm PCR of KO. (E) Western of wild type KO.

Figure 5

Phenotypic analysis of ΔPbMC1 parasites. (A) Gametocytes, female degeneracy. (B) Oocyst number. (C) Midgut oocyst sporozoite number. (D) Hemolymph sporozoite. (E) Salivary gland sporozoites. (F) Sporozoite invasion of hepatocytes.

Figure 6

(A) A structural model of PbMC1 (blue) was generated using caspase 3 (red) as a template. The main chain backbone around the active His and Cys and the S1 subsite are in good agreement. (B) Additional models were generated of Arabidopsis thaliana MC1 and MC6 that can accommodate Arg and Lys, respectively, where no structures are available. This allowed comparison of key predicted amino acid residue at the S1 subsite. The S1 subsite from the structure of caspase 3 and the S1 pocket from models of PbMC1 and AtMC1. (C) An inhibitor subtractive gametocyte preparation was incubated with a number of authentic AMC substrates, in the absence or presence of iodoacetamide.

Substrates with Phe or Pro at P1 were hydrolysed, and this activity was sensitive to thiol alkylation.

Appendix B

Consensus.py

```
#!/usr/local/bin/python

##########################################################################
#                                                                        #
#      ______                                                            #
#     / ____/___  ____  ________  ____  _______  _______  ____  __  __   #
#    / /   / __ \/ __ \/ ___/ _ \/ __ \/ ___/ / / / ___/ / __ \/ / / /   #
#   / /___/ /_/ / / / (__  )  __/ / / (__  ) /_/ (__  ) / /_/ / /_/ /    #
#   \____/\____/_/ /_/____/\___/_/ /_/____/\__,_/____(_) .___/\__, /     #
#                                                     /_/    /____/      #
#                                                                        #
#                             written by                                 #
#                                                                        #
#                          Nima Fayazmanesh                              #
#                        Jacobson Group, UCSF                            #
#                          February 20, 2007                             #
#                                                                        #
# This program takes a set of docking/rescoring files, an msf format     #
# multiple sequence alignment, and a list of evalue cutoffs. It uses     #
# Blast 2 Sequences to select homologous "neighbor" sequences for each   #
# sequence, and combines the files for each sequence and its neighbors   #
# to create a consensus using the rank by rank and the rank by number    #
# methods. To estimate the strength of the consensus, the program        #
# computes Spearman's rank correlation coefficient between the hit list  #
# of each sequence and its neighbors.                                    #
#                                                                        #
# usage: consensus.py <le-i> <le-j> ... <le-n>                           #
#                                                                        #
# Note: To run this program, Blast 2 Sequences must be installed         #
# locally. Blast can be downloaded by FTP from the National Center for   #
# Biotechnology Information at ftp://ftp.ncbi.nih.gov/blast/. The path  #
# to the local installation is specified in the function bl2seq below.   #
##########################################################################

import os, re, copy, sys

# This function searches the current directory for the msf format multiple
# sequence alignment file and returns the file name.

def get_alignment():
  files = os.listdir('.')
  alignment = None
  for file in files:
    m = re.search('.msf$', file)
    if m:
      alignment = file
  if not alignment:
    print "Error: missing .msf file in current directory."
  return alignment

# This function searches the current directory for docking/rescoring files
# and returns a dictionary of file names, indexed by sequence title.

def get_names(type):
  files = os.listdir('.')
  names = {}
  if type == 'dock':
    for file in files:
      m = re.search('.rept$', file)
      if m:
        seq = re.split('.rept', file)[0]
        names[seq] = file
  elif type == 'rescore':
    for file in files:
      m = re.search('-flexible-rescore.txt$', file)
      if m:
        seq = re.split('-flexible-rescore.txt', file)[0]
        names[seq] = file
  return names

# This function searches the current directory for a file containing blast
# evalues for pairs of sequences.

def get_evalues(type):
  files = os.listdir('.')
  evalues = None
  for file in files:
    m = re.match(str(type) + '.blast', file)
    if m:
      f = open(file, 'r')
      evalues = eval(f.read())
      f.close()
  return evalues

# This function takes a dictionary of blast evalues for each sequence pair
# and writes it to a file. Obtaining the evalues for all sequence pairs is
# computationally intensive; by writing to a file, we save time if the
# program needs to be run again under different conditions, for example,
# using a different evalue cutoff for neighbor selection or updated
# docking/rescoring files.

def write_evalues(evalues, type):
  f = open(str(type) + '.blast', 'w')
  f.write(str(evalues))
  f.close()
  return None

# This function takes the sequence titles and the name of the msf format
# multiple sequence alignment. It reads in the alignment, extracts the
# sequences of interest, removes gap characters, and returns a dictionary
# of sequences, indexed by title.

def msf2dict(names, alignment):
  seqs = names.keys()
  dict = {}
  for title in seqs:
    dict[title] = ''
  f = open(alignment, 'r')
  lines = f.readlines()
  for line in lines:
    words = line.split()
    if words:
      title = words[0]
      if title in dict:
        seq = ''
        for word in words[1:]:
          for char in word:
            if char != '.':
              seq = seq + char
        dict[title] = dict[title] + seq
  return dict

# This function takes a dictionary of sequence titles and sequences and
# writes each sequence to a file. These files will be used as input to
# bl2seq to compute the blast evalues for each pair of sequences.

def write_fasta(dict):
  seqs = dict.keys()
  for title in seqs:
    seq = dict[title]
    file = title + '.fa'
    f = open(file, 'w')
    f.write(seq)
    f.close()

# This function finds and deletes sequence files in the current directory.
# After running bl2seq, such files are not needed and may be removed.

def remove_fasta():
  files = os.listdir('.')
  for file in files:
    m = re.search('.fa$', file)
    if m:
      os.remove(file)
  return None

# This function takes the names of two sequences, runs bl2seq, and parses
# the output for the evalue. The correct path to the local installation of
# bl2seq must be specified. For the result to match the Blast 2 Sequences
# web server at NCBI, the option -d is used to specify the theoretical
# database size. The current size is listed in the web server output under
# "Length of database."

def bl2seq(seq1, seq2):
  title1 = seq1 + '.fa'
  title2 = seq2 + '.fa'
  command = '/Applications/blast-2.2.15/bin/bl2seq -i ' + title1 + ' -j ' +
title2 + ' -p blastp -d 1603721534'
  stdout = os.popen(command)
  lines = stdout.readlines()
  for line in lines:
    if re.search('Expect', line):
      words = line.split()
      n = len(words)
      last = words[n-1]
      if last[0] == 'e':
        last = '1' + last
      e = eval(last)
      break
    elif re.search('No hits found', line):
      e = '-'
      break
  return e

# This function takes a dictionary of sequence titles and sequences and
# returns a dictionary of evalues for each sequence pair. All permutations
# are evaluated because the order in which sequences are passed to bl2seq
# matters; the evalue of (seq1, seq2) is not identical to that of
# (seq2, seq1). By evaluating only sequences for which we have
# docking/rescoring results, rather than the full sequence set in the
# multiple alignment, the computational time is minimized.

def run_blast(dict):
  seqs = dict.keys()
  evalues = {}
  for seq1 in seqs:
    for seq2 in seqs:
      pair = (seq1, seq2)
      e = bl2seq(seq1, seq2)
      evalues[pair] = e
  return evalues

# This function reads in a set of docking/rescoring files and returns a
# dictionary containing ligand names, ranks, and scores, indexed by file
# name. If columns of data are missing from any of the rescoring files,
# the names of the problematic files are printed and the program exits.

def read(names, type):
  error = None
  filenames = names.values()
  if type == 'dock':
    data = {}
    for file in filenames:
      i = []
      f = open(file, 'r')
      lines = f.readlines()
      for line in lines:
        m = re.match('\s+\d+.+', line)
        if m:
          words = m.group(0).split()
          rank = eval(words[0])
          ligand = words[1]
          score = eval(words[3])
          j = [ligand, rank, score]
          i.append(j)
      data[file] = i
  elif type == 'rescore':
    data = {}
    for file in filenames:
      try:
        i = []
        f = open(file, 'r')
        lines = f.readlines()
        for line in lines:
          m = re.match('\s+\d+.+', line)
          if m:
            words = m.group(0).split()
            rank = eval(words[0])
            ligand = words[6]
            score = eval(words[4])
            j = [ligand, rank, score]
            i.append(j)
        data[file] = i
      except IndexError:
        error = 1
        print "Error reading file:", file
        continue
  if error:
    print "Exiting..."
    return None
  else:
    return data

# This function takes a dictionary of ligand data, indexed by file name,
# and identifies redundant ligands that appear more than once in a list.
# It changes their names so that they are distinct and returns a
# dictionary of nonredundant ligand data, indexed by file name.
#
# Note: Redundancy arises when stereoisomers in a ligand library have
# identical names. Redundant ligands must be filtered out because they
# cause errors in computing the rank correlation between pairs of lists.
# Redundant ligands also decrease the rank of the best scoring
# stereoisomer, because the ranks and normalized scores are averaged in
# the consensus result.
```

```
def remove_redundancy(data):
  for prot, set in data.items():
    nr = {}
    for lig in set:
      name = lig[0]
      if name in nr:
        nr[name] = nr[name] + 1
        lig[0] = name + '_' + str(nr[name])
      else:
        nr[name] = 1
  return data

# This function takes a list of numbers and returns their mean average.

def mean(scores):
  n = len(scores)
  sum = 0.0
  for score in scores:
    sum = sum + score
  m = sum / n
  return m

# This function takes a list of numbers and their mean average, and
# returns the standard deviation.

def std_dev(scores, m):
  n = len(scores)
  sum = 0.0
  for score in scores:
    d = score - m
    sum = sum + d * d
  s = (sum / (n - 1)) ** 0.5
  return s

# This function takes a dictionary of ligand data, indexed by file name.
# For each sequence, the raw scores are normalized by subtracting the mean
# and dividing by the standard deviation. It returns a dictionary of
# normalized ligand data, indexed by file name.

def normalize(data):
  error = None
  ndata = copy.deepcopy(data)
  titles = ndata.keys()
  for title in titles:
    scores = []
    report = ndata[title]
    for lig in report:
      score = lig[2]
      scores.append(score)
    try:
      m = mean(scores)
    except ZeroDivisionError:
      error = 1
      print "Error reading file:", title
      continue
    s = std_dev(scores, m)
    for lig in report:
      score = lig[2]
      nscore = (score - m) / s
      lig[2] = nscore
  if error:
    print "Exiting..."
    return None
  else:
    return ndata

# This function takes the sequence titles and file names, a dictionary of
# evalues for each sequence pair, and an evalue cutoff for neighbor
# selection. It returns a dictionary of neighboring sequences, with
# evalues less than the cutoff, for each sequence. It also returns a
# dictionary of files to merge, indexed by sequence.

def select(names, evalues, cutoff):
  seqs = names.keys()
  neighbors = {}
  files2merge = {}
  for seq1 in seqs:
    n = []
    f = []
    for seq2 in seqs:
      pair = (seq1, seq2)
      e = evalues[pair]
      if e <= cutoff:
        n.append(seq2)
        file2 = names[seq2]
        f.append(file2)
    neighbors[seq1] = n
    files2merge[seq1] = f
  return neighbors, files2merge

# This function takes a sequence, file pair, a dictionary of normalized
# ligand data indexed by file name, and a list of files to merge. It
# returns a dictionary of ligand ranks, normalized scores, and initial
# rank in the hit list of the sequence of interest, indexed by ligand.

def ligand_dictionary(name, ndata, files):
  lig_dict = {}
  for file in files:
    data = ndata[file]
    for lig in data:
      title = lig[0]
      rank = lig[1]
      score = lig[2]
      init_rank = '-'
      if file == name[1]:
        init_rank = rank
      if title not in lig_dict:
        rlist = [rank]
        slist = [score]
        lig_dict[title] = [rlist, slist, init_rank]
      else:
        rlist = lig_dict[title][0]
        slist = lig_dict[title][1]
        rlist.append(rank)
        slist.append(score)
        if init_rank != '-' and init_rank < lig_dict[title][2]:
          lig_dict[title][2] = init_rank
  return lig_dict

# This function takes a dictionary of ligand ranks, normalized scores, and
# initial rank, indexed by ligand. For each ligand, it computes the mean
# rank, mean normalized score, and number of votes, or hit lists in which
# the ligand is ranked. It returns a list of the ligands sorted by mean
# rank and a list sorted by mean normalized score.

def consensus(d):
  rsort = []
  ssort = []
  for lig, data in d.items():
    rlist = data[0]
    slist = data[1]
    init_rank = data[2]
    mean_rank = mean(rlist)
    mean_score = mean(slist)
    votes = len(rlist)
    rsort.append([mean_rank, lig, init_rank, votes])
    ssort.append([mean_score, lig, init_rank, votes])
  rsort.sort()
  ssort.sort()
  return rsort, ssort

# This function takes a list of rank differences and returns Spearman's
# rank correlation, the square of the correlation, the t value to
# determine whether r is significantly different from zero, and the sample
# size n.

def compute_r(rank_diff):
  n = len(rank_diff)
  sum = 0.0
  for d in rank_diff:
    sum = sum + d * d
  r = 1 - 6 * sum / (n * (n ** 2 - 1))
  r_sq = r * r
  if n >= 10 and r_sq < 1:
    t = r * (n - 2) ** 0.5 / (1 - r_sq) ** 0.5
  else:
    t = '-'
  return r, r_sq, t, n

# This function takes two sequences and data for their common ligands, and
# computes the difference in ranks for each ligand between the two lists.
# It returns the sequence names and correlation values.

def correlation(i_common, j_common):
  Pi, Di = i_common
  Pj, Dj = j_common
  rank_diff = []
  for x in Di:
    Li = x[0]
    Ri = x[1]
    for y in Dj:
      Lj = y[0]
      Rj = y[1]
      if Li == Lj:
        d = Ri - Rj
        rank_diff.append(d)
  r, r_sq, t, n = compute_r(rank_diff)
  return [Pi, Pj, r, r_sq, t, n]

# This function takes two sequences and their ligand data, filters the
# data to include ligands common to the lists of both proteins, and
# re-ranks the ligands in each list. This is necessary for correctly
# computing Spearman's correlation coefficient between the two lists.

def common(i, j):
  Pi, Si = i
  Pj, Sj = j
  Si_common = copy.deepcopy(Si)
  Sj_common = copy.deepcopy(Sj)
  common_ligs = []
  for Li in Si:
    Ni = Li[0]
    for Lj in Sj:
      Nj = Lj[0]
      if Ni == Nj:
        common_ligs.append(Ni)
        break
  for lig in Si:
    name = lig[0]
    if name not in common_ligs:
      Si_common.remove(lig)
  for lig in Sj:
     name = lig[0]
     if name not in common_ligs:
       Sj_common.remove(lig)
  for set in [Si_common, Sj_common]:
    rank = 1
    for lig in set:
      lig[1] = rank
      rank = rank + 1
  i_common = Pi, Si_common
  j_common = Pj, Sj_common
  return i_common, j_common

# This function takes a sequence, file pair, a dictionary of normalized
# ligand data indexed by file name, and a list of files to merge. It
# returns a list of correlation values for the sequence and each of its
# neighbors.

def run_correlation(name, data_nr, files):
  corr = []
  filename = name[1]
  i = filename, data_nr[filename]
  for file in files:
    j = file, data_nr[file]
    i_common, j_common = common(i, j)
    corr.append(correlation(i_common, j_common))
  return corr

# This function writes a list of neighbors, rank correlation values, and
# consensus results for each protein. Two result files are written using
# the rank by rank and rank by number methods.

def write(name, type, cutoff, evalues, neighbors, rsort, ssort, corr):
  for data in [rsort, ssort]:
    file = name[1]
    if data == rsort:
      method = 'rank'
    elif data == ssort:
      method = 'number'
    saveout = sys.stdout
    f = open(file + '_' + str(cutoff) + '.' + method, 'w')
    sys.stdout = f
    print "Sequence:", name[0], "\n"
    print "Threshold for neighbor selection: E =", cutoff, "\n"
    print "List of neighbors and E values of sequence to neighbors:"
    print "            Neighbor               E value "
    print "================================== ========="
    for neighbor in neighbors[name[0]]:
      print neighbor, " " * (35 - len(neighbor)), evalues[(name[0], neighbor)]
    print "\nSpearman rank correlation between hit lists:"
    print "                   Protein #1
Protein #2                           r     r_sq      t       n   "
    print "======================================================
====================================================== ======= ======= =======
======="
    for i in corr:
      prot1 = i[0]
      prot2 = i[1]
      r = str(round(i[2], 3))
      r_sq = str(round(i[3], 3))
      t = i[4]
      if t != '-':
        t = str(round(t, 3))
      n = str(round (i[5], 3))
      print prot1, " " * (53 - len(prot1)), prot2, " " * (53 - len(prot2)), r,
" " * (6 - len(r)), r_sq, " " * (6 - len(r_sq)), t, " " * (6 - len(t)), n
    if data == rsort:
      print "\nConsensus method is rank by rank.\n"
      print "Consensus scoring results:"
      print "Rank           Ligand                  Mean rank     Init rank
Votes   "
      print "====== ============================== ============
============ =========="
    elif data == ssort:
      print "\nConsensus method is rank by number.\n"
      print "Consensus scoring results:"
      print "Rank           Ligand                   Mean score   Initial
rank     Votes   "
      print "====== ============================== =============
============= =========="
    c = 1
    for i in data:
      rank = str(c)
      mean = str(round(i[0], 3))
      lig = i[1]
      init_rank = str(i[2])
      votes = str(i[3])
      print rank, " " * (7 - len(rank)), lig, " " * (30 - len(lig)), mean, " "
* (13 - len(mean)), init_rank, " " * (13 - len(init_rank)), votes
      c = c + 1
    sys.stdout = saveout
    f.close()
  return None

# This function makes a directory for the consensus results created at a
# specific evalue cutoff and moves the result files to the directory.

def cleanup(cutoff):
  name = str(cutoff)
  try:
    os.mkdir(name)
  except OSError:
    pass
  files = os.listdir('.')
  for file in files:
    pattern = str(cutoff) + '.rank$' + '|' + str(cutoff) + '.number$'
    m = re.search(pattern, file)
    if m:
      old = str(file)
      new = str(cutoff) + '/' + str(file)
      os.rename(old,new)

#################
# main function #
#################

def run(cutoffs):
  alignment = get_alignment()
  if not alignment:
    sys.exit(0)
  types = ['dock', 'rescore']
  for type in types:
    names = get_names(type)
    if not names:
      continue
    evalues = get_evalues(type)
    if not evalues:
      dict = msf2dict(names, alignment)
      write_fasta(dict)
      evalues = run_blast(dict)
      write_evalues(evalues, type)
      remove_fasta()
    data = read(names, type)
    if not data:
      break
    data_nr = remove_redundancy(data)
    ndata = normalize(data_nr)
    if not ndata:
      break
    for cutoff in cutoffs:
      neighbors, files2merge = select(names, evalues, cutoff)
      for name in names.items():
        files = files2merge[name[0]]
        lig_dict = ligand_dictionary(name, ndata, files)
        rsort, ssort = consensus(lig_dict)
        corr = run_correlation(name, ndata, files)
        write(name, type, cutoff, evalues, neighbors, rsort, ssort, corr)
      cleanup(cutoff)
  return None

################
# program body #
################

if (len(sys.argv) == 1):
  print "usage: consensus.py <le-i> <le-j> ... <le-n>"
  sys.exit(0)
input = sys.argv[1:]
cutoffs = []
for i in input:
  cutoffs.append(eval(i))
  cutoffs.sort()
run(cutoffs)
```

Appendix C

Sample Output of Consensus.py

```
Sequence: 164_MLEI_MLE_38198158_RHOER_IS

Threshold for neighbor selection: E = 1e-150

List of neighbors and E values of sequence to neighbors:
            Neighbor                E value
================================== =========
164_MLEI_MLE_5915882_RHOOP_CH        1e-172
164_MLEII_MLE_77362681_RHOAN_I       1e-150
164_MLEII_MLE_82548049_NOCC-_I       1e-153
164_MLEI_MLE_38198158_RHOER_IS       0.0

Spearman rank correlation between hit lists:
                   Protein #1
Protein #2                           r     r_sq      t       n
=====================================================
====================================================== ======= ======= =======
=======
164_MLEI_MLE_38198158_RHOER_IS.rept
164_MLEI_MLE_5915882_RHOOP_CH.rept                     0.576   0.331   10.765
236.0
164_MLEI_MLE_38198158_RHOER_IS.rept
164_MLEII_MLE_77362681_RHOAN_I.rept                    0.504   0.254   9.544
269.0
164_MLEI_MLE_38198158_RHOER_IS.rept
164_MLEII_MLE_82548049_NOCC-_I.rept                    0.761   0.579   19.286
273.0
164_MLEI_MLE_38198158_RHOER_IS.rept
164_MLEI_MLE_38198158_RHOER_IS.rept                    1.0     1.0     -
289.0

Consensus method is rank by rank.

Consensus scoring results:
Rank            Ligand                   Mean rank      Init rank     Votes
====== ================================ =============  ============= ==========
1        D-Iduronate                     11.0           4              4
2        3-deoxy-D-Glucarate             14.0           3              4
3        3-phosphoglycerate              19.75          22             4
4        L-Allarate                      20.0           25             4
5        D-Alluronate                    20.5           6              4
6        GLY-ARG                         21.75          27             4
7        L-Tartarate                     24.0           20             4
8        2-phosphoglycerate              25.0           7              4
9        L-Arabarate-L-Lyxarate          28.25          16             4
10       5-keto-4-deoxy-Idarate          29.0           13             4
11       D-6D-Allonate                   30.75          2              4
12       D-6D-Altronate                  32.0           28             4
13       3-phosphoglycerate_2            34.25          33             4
14       2,3-dideoxy-D-Glucarate         34.75          17             4
15       2,3-dideoxy-D-Glucarate_2       35.75          18             4
16       L-Mannonate                     36.0           58             4
17       L-Xylonate                      37.25          1              4
18       D-Mannarate                     38.25          9              4
19       5-keto-4-deoxy-Mannarate        38.75          14             4
20       L-Allonate                      38.75          54             4
21       D-Altrarate-D-Talarate          41.5           76             4
...
```