Published in final edited form as:
Biochemistry. 2011 November 22; 50(46): 9950–9962. doi:10.1021/bi201312u.

The Enzyme Function Initiative†

John A. Gerlt[1,2,*], **Karen N. Allen**[3], **Steven C. Almo**[4], **Richard N. Armstrong**[5], **Patricia C. Babbitt**[6], **John E. Cronan**[2,7], **Debra Dunaway-Mariano**[8], **Heidi J. Imker**[2], **Matthew P. Jacobson**[9], **Wladek Minor**[10], **C. Dale Poulter**[11], **Frank M. Raushel**[12], **Andrej Sali**[6], **Brian K. Shoichet**[9], and **Jonathan V. Sweedler**[2,13]

[1]Departments of Biochemistry and Chemistry, University of Illinois at Urbana-Champaign, Urbana, IL 61801

[2]Institute for Genomic Biology, University of Illinois at Urbana-Champaign, Urbana, IL 61801

[3]Department of Chemistry, Boston University, Boston, MA

[4]Department of Biochemistry, Albert Einstein College of Medicine, Bronx, NY 10461

[5]Department of Biochemistry, Vanderbilt University Medical Center, Nashville, TN 37232

[6]Departments of Bioengineering and Therapeutic Sciences and of Pharmaceutical Chemistry, California Institute of Quantitative Biosciences, University of California, San Francisco, CA 94143

[7]Departments of Microbiology and Biochemistry, University of Illinois at Urbana-Champaign, Urbana, IL 61801

[8]Department of Chemistry and Chemical Biology, Albuquerque, NM 87131

[9]Department of Pharmaceutical Chemistry, University of California, San Francisco, CA 94143

[10]Department of Department of Molecular Physiology and Biological Physics, University of Virginia, Charlottesville, VA 22908

[11]Department of Chemistry, University of Utah, Salt Lake City, UT 84112

[12]Department of Chemistry, Texas A&M University, College Station, TX 77843

[13]Department of Chemistry, University of Illinois at Urbana-Champaign, Urbana, IL 61801

Abstract

The Enzyme Function Initiative (EFI) was recently established to address the challenge of assigning reliable functions to enzymes discovered in bacterial genome projects; in this Current Topic we review the structure and operations of the EFI. The EFI includes the Superfamily/Genome, Protein, Structure, Computation, and Data/Dissemination Cores that provide the infrastructure for reliably predicting the *in vitro* functions of unknown enzymes. The initial targets for functional assignment are selected from five functionally diverse superfamilies (amidohydrolase, enolase, glutathione transferase, haloalkanoic acid dehalogenase, and isoprenoid synthase), with five superfamily-specific Bridging Projects experimentally testing the predicted *in vitro* enzymatic activities. The EFI also includes the Microbiology Core that evaluates the *in vivo* context of *in vitro* enzymatic functions and confirms the functional predictions of the EFI. The deliverables of the EFI to the scientific community include: **1)** development of a large-scale,

†This research was supported by NIH 5U54GM093342-02. Molecular graphics images were produced using the UCSF Chimera package from the Resource for Biocomputing, Visualization, and Informatics at the University of California, San Francisco (supported by NIH P41 RR-01081).

*To whom correspondence should be addressed: J.A.G.: Institute for Genomic Biology, University of Illinois, 1206 West Gregory Drive, Urbana, IL 61801. Phone: (217) 244-7414. Fax: (217) 333-0508. j-gerlt@uiuc.edu.

multidisciplinary sequence/structure-based strategy for functional assignment of unknown enzymes discovered in genome projects (target selection, protein production, structure determination, computation, experimental enzymology, microbiology, and structure-based annotation); **2)** dissemination of the strategy to the community *via* publications, collaborations, workshops, and symposia; **3)** computational and bioinformatic tools for using the strategy; **4)** provision of experimental protocols and/or reagents for enzyme production and characterization; and **5)** dissemination of data via the EFI's website, enzymefunction.org. The realization of multidisciplinary strategies for functional assignment will begin to define the full metabolic diversity that exists in nature and will impact basic biochemical and evolutionary understanding, as well as a wide range of applications of central importance to industrial, medicinal and pharmaceutical efforts.

As genome sequencing has become routine, the number of protein sequences in the databases has expanded exponentially. In early October 2011, the UniProtKB/TrEMBL database contained 16,886,838 entries. This abundance of protein sequences is a boon for biology and biomedical science, because understanding the genomic capabilities of an organism will allow its metabolism and physiology to be defined and targets for chemotherapeutic or antibiotic intervention to be identified. Furthermore, understanding the functions of proteins that are enzymes and their associated metabolic pathways should enable advances in medicine, chemistry, synthetic biology, and industry.

However, achievement of this potential is confounded by the problem that reliable *in vitro* functions have been assigned to only a small (and diminishing) fraction of the proteins in the TrEMBL database (1). Every sequenced genome encodes a large number of "hypothetical" proteins that share sufficiently low sequence similarity with those previously identified that even hints of their molecular functions cannot be deduced. An even more acute problem is that the functional annotations for many proteins in GenBank are either misleading or incorrect, as the result of incorrect computational assignment based on annotations for the closest sequence homologues. As additional incorrect annotations are made, these are propagated throughout the databases, expanding the problem. A recent critical analysis performed by one of us (P.C.B.) for members of 37 characterized protein families concluded that 40% of the sequences deposited as recently as 2005 were misannotated (1). As long as the deposited annotations remain uncorrected, this problem is certain to become more prevalent and increasingly problematic.

Therefore, determining reliable functions for unknown proteins (biochemically uncharacterized proteins with uncertain functions) discovered in genome projects is a major challenge in contemporary biology. Although the impetus for assigning these functions is clear, effective methods for doing so are not. Strategies for functional assignment of unknown proteins have utilized clues provided by many approaches, including **1)** sequence similarity by comparison to orthologous or paralogous proteins; **2)** colocalization of genes providing operon/metabolic context for prokaryotic proteins; **3)** transcriptional analysis through chip-based and RNAseq technologies; **4)** identification of upstream DNA motifs that might coregulate transcription; **5)** functions of multidomain proteins to identify coupled activities in a pathway; **6)** protein-protein interaction studies; and **7)** phenotypes of gene deletion/knockout mutants. For enzymes, sequence similarity and/or genome/operon context often can provide coarse functional clues, e.g., the enzyme is a kinase, aldolase, or dehydrogenase, but they are rarely sufficient to provide information about the substrate specificity and, therefore, the actual reaction that is catalyzed.

How might the identity of the substrate and, therefore, the molecular function for an unknown enzyme be deduced in a *high throughput* fashion to meet the challenges presented by the increasing number of genome projects? Given the number of unknown sequences,

biochemical experimentation alone is clearly not a feasible strategy. Rather, computational approaches are necessary to guide experimental verification and, also, to annotate proteins that cannot be experimentally characterized. Indeed, computational tools can play critical roles in functional assignment:

1. Bioinformatic analyses can cluster sequences into probable isofunctional groups, thereby assigning tentative functions to be investigated by structure determination, structural modeling and docking, and biochemical experimentation.
2. Homology modeling methods can expand the use of structural models to guide function assignment to proteins without experimentally determined structures.
3. Computational docking methods can leverage structure to guide functional assignment by suggesting substrates and ligands for biochemical experimentation.

In fact, all three computational strategies play critical roles in our efforts to develop a high throughput, multidisciplinary sequence/structure-based strategy for functional assignment, as described in this Current Topic.

Enzyme Function Initiative (EFI): Overview

With these considerations in mind, we proposed formation of the Enzyme Function Initiative (EFI) in which computation-based prediction of substrate specificity is the centerpiece of a multidisciplinary strategy for functional assignment of unknown enzymes (2). The strategy includes bioinformatics, experimental structural biology, structural modeling and docking, and experimental enzymology to assign *in vitro* substrate specificities and enzymatic functions as well as microbiology (phenotypic analyses, genetics, and transcriptomics); it also includes metabolomics to validate (or disprove) the predicted and experimentally confirmed *in vitro* enzymatic function as the authentic *in vivo* function (Figure 1).

The EFI started in May 2010 with the support of a Large Scale Collaborative Project (U54GM093342) from the National Institute of General Medical Sciences (NIGMS). The EFI is a five-year cooperative agreement among NIGMS, the host institution (University of Illinois, Urbana-Champaign), and the subcontracting institutions (refer to the author list for details). A cooperative agreement is a support mechanism in which NIGMS provides substantial scientific and programmatic involvement, i.e., program staff assist, guide, coordinate, and/or participate in project activities. The EFI is reviewed by NIGMS on a continuing basis, with formal reviews after 18 and 36 months. This *modus operandi* differs from investigator-initiated research grants (R01) and program project grants (P01) where the scientific direction and progress usually are not subject to active oversight by NIGMS staff during the project period. Peter Preusch, chief of the Biophysics Branch in the NIGMS Division of Cell Biology and Biophysics, is the Scientific Officer and a member of the EFI's internal Steering Committee. Warren C. Jones, chief of the Biochemistry and Biorelated Chemistry Branch in the NIGMS Division of Pharmacology, Physiology, and Biological Chemistry, is the Program Officer who oversees the budgetary and administrative aspects of the EFI within NIGMS. An external Scientific Advisory Committee meets annually with the EFI to assess progress and provide guidance for programmatic direction; the members include Helen Berman, Rutgers University and Director of the Protein Data Bank (PDB); Benjamin Cravatt, The Scripps Research Institute; Barry Honig, Columbia University Medical Center; Eaton Lattman, Hauptman-Woodward Medical Research Institute, University at Buffalo; and Rowena Matthews, University of Michigan.

The EFI's strategy for functional assignment can be summarized by the "funnel" depicted in Figure 2. With the available resources, the initial computational prediction of substrate specificity can be performed in a relatively high throughput (tens of enzymes per month);

the subsequent experimental enzymology that tests the computational predictions can be performed with modest throughput (several enzymes per month); and *in vivo* studies of the *in vitro* assigned functions are labor and time intensive and, therefore, low throughput (one or two per month), limiting the number of *in vivo* functions that can be evaluated. However, without reliable computational prediction, experimental evaluation would be a random walk through substrate space, preventing efficient functional assignment. Furthermore, without *in vivo* "testing", the *in vitro* assigned functions may be uninformative about the *in vivo* function (*vide infra*) or enzymes with promiscuous *in vitro* substrate specificities could have uncertain physiological importance.

The protein "targets" selected to develop the strategy for functional assignment are members of functionally diverse enzyme superfamilies (conserved partial reactions or chemical capability but divergent overall function) so that assignment of function is not trivial, i.e., homology inferred from simple sequence comparisons alone does not allow assignment of function (3, 4). For example, the members of the functionally diverse enolase superfamily catalyze different reactions that always are initiated by Mg^{2+}-assisted enolization of carboxylate anions and include β-elimination (dehydration, deamination, and cycloisomeriation) and 1,1-proton transfer (racemization and epimerization) reactions (5, 6). In another example, members of the functionally diverse amidohydrolase superfamily catalyze metal-assisted hydrolysis of C-O, C-N, and P-O bonds in diverse substrates (7).

Briefly, our approach ("pipeline" in Figure 3) is to **1)** use sequence relationships to identify putative isofunctional families within functionally diverse superfamilies from which targets are selected to develop, test, and improve the strategy; **2)** for bacterial enzymes, analyze the genome/operon contexts within the families to identify other enzymes that are part of the same metabolic pathway to provide additional functional clues; **3)** when possible, purify and structurally characterize the targets and, when appropriate, other enzymes in the metabolic pathway; **4)** if structures cannot be determined experimentally, use homology modeling to obtain reliable models; **5)** perform *in silico* ligand docking to generate rank-ordered lists of predicted substrates; **6)** experimentally screen predicting substrates for activity, as well as synthesize and screen novel compounds suggested by docking, to determine *in vitro* function; **7)** determine structures of liganded complexes so that the predicted and experimental binding "poses" of the substrate (or analog/product) can be compared to both evaluate as well as improve the computational procedures for homology modeling and/or ligand docking; **8)** when possible, elucidate the *in vivo* function by a combination of focused genetics (knockouts and overexpression), transcriptomics, and metabolomics; and **9)** when possible to do so with high confidence, transfer annotations from the proteins for which the EFI has established reliable functions to other unknowns (1, 8). Elements of this strategy had been demonstrated by some of the authors (J.A.G., S.C.A, P.C.B., M.P.J., F.M.R., A.S., and B.K.S.) for the functionally diverse amidohydrolase and enolase superfamilies (*vide infra*); with the support of the EFI those efforts are being expanded to include dedicated protein production and structure determination for targets from additional functionally diverse superfamilies as well as microbiology and metabolomics.

The EFI's efforts are not organized according to Specific Aims that are integral to traditional research grants, e.g., NIH R01 and P01 funding mechanisms. Instead, the EFI focuses on ***deliverables*** that will benefit the biomedical community. These deliverables include:

1. Development of a multidisciplinary sequence/structure-based strategy for predicting the functions of unknown enzymes discovered in genome sequencing projects.

2. Dissemination of the strategy to the community by publications, web-based interfaces, workshops, symposia, and collaboration of external investigators with the bioinformatics and computational components of the EFI.
3. Development of computational and bioinformatic tools for utilizing the strategy.
4. The genes encoding all targets are made available to the community *via* the PSI-MR (http://psimr.asu.edu/). To the extent possible, compounds used for experimental studies of enzymatic activity will be disseminated; if these are not available in sufficient quantities to allow distribution, the procedures for their synthesis will be made available. Protocols for protein expression and functional assays also will be available via PepcDB (pepcdb.sbkb.org) and the EFI's website (enzymefunction.org), respectively.
5. Dissemination of both computational predictions and experimental data *via* the EFI's website.

In the following sections, we describe the organization of the EFI as well as its internal collaborative interactions and operations.

Enzyme Function Initiative (EFI): Scientific Cores

Central to the EFI's strategy is exploitation of developments in bioinformatics, structural genomics, homology modeling, and *in silico* screening for high throughput prediction of the substrate specificities of unknown enzymes. The EFI is composed of six Scientific Cores (Superfamily/Genome, Protein, Structure, Computation, Microbiology, and Data/Dissemination) and five Bridging Projects that focus on a different functionally diverse superfamily selected as model systems for development of the strategy [amidohydrolase (AH), enolase (EN), glutathione transferase (GST), haloalkanoic acid dehalogenase (HAD), and isoprenoid synthase (IS)].

The Scientific Cores constitute the intellectual and technological "heart" of the EFI. Each is responsible for one of the multidisciplinary approaches that is essential for the successful development and dissemination of the multidisciplinary sequence/structure-based strategy for facilitating functional assignment.

The "pipeline" that describes the flow of information and materials among the Cores and Bridging Projects is shown in Figure 3. Their individual and collaborative roles are summarized in the following paragraphs.

Superfamily/Genome Core

As the sequence databases expand, the increasing number of members of individual protein families and functionally diverse superfamilies makes traditional approaches for viewing sequence relationships, i.e., trees and dendrograms, difficult. Sequence similarity networks developed in part by one of the authors (P.C.B.) provides a powerful approach to identify and classify members of large groups of homologous proteins (9). The Superfamily/Genome Core provides regular updates of the membership of the EFI's superfamilies that are then subjected to additional bioinformatic analyses. Automated scripts and new structure/sequence motif methods are used to identify members of each superfamily, with expert curators overseeing the grouping of the sequences into isofunctional families. The sequences are maintained in the Structure-Function Linkage Database2 (SFLD; http://sfld.rbvi.ucsf.edu) (10) that also provides sequence similarity networks and other tools

[2]The SFLD was developed by the NIH NCR Resource for Biocomputing, Visualization, and Informatics (supported by NIH P41 RR-01081) as well as R01GM60595 and NSF DBI 0234768, and NSF DBI 0640476 (to P.C.B).

that allow facile organization of the members of functionally diverse superfamilies into putative isofunctional families ("clusters") (Figure 4) using the open source software Cytoscape (11). These resources are used, in collaboration with the Computation Core and Bridging Projects, to identify and prioritize targets for functional assignment as well as assist the other Cores and Bridging Projects in their studies.

Protein Core

The Protein Core is responsible for high throughput cloning, protein expression, and protein purification to provide samples for structure determination by X-ray crystallography by the Structure Core and enzymatic assays and library screening by the Bridging Projects. As the EFI enters its second year, the infrastructure is in place for large-scale protein production and distribution to both the Structure Core and Bridging Projects (as many as 600 proteins per year). In collaboration with the Structure Core, the Protein Core screens proteins for ligands using thermal denaturation-based approaches, i.e., ThermoFluor (12, 13). Ligand screening both provides functional clues and, more importantly can support cocrystallization experiments with ligands that yield structures in conformations relevant to enzymatic catalysis. These "catalytically competent" structures are the most valuable, as they provide productively "dockable" templates for *in silico* screening by the Computation Core.

Structure Core

Considerable economies have been realized in protein production and structure determination, in part due to the efforts of the Protein Structure Initiative (PSI). Based on these advances we anticipate that the Structure Core will be able to determine as many as 50 "new" structures and 50 liganded structures per year.

The availability of high-resolution structures enables the Computation Core to use *in silico* analyses that provide predictions of substrate specificity and, also, to construct models of homologous sets of proteins to predict how function diverges as sequence diverges. The X-ray structures are critical for evaluating the structural bases for specificity and thereby accessing the accuracy of the computational predictions against experimentally liganded structures. Concurrently, the X-ray structures also provide an important check on the ability of computational algorithms to correctly predict the structure of the liganded active site.

Computation Core

The Computation Core develops, applies, and disseminates computational tools that leverage structural information to infer enzymatic function. As discussed in the following paragraphs, the two primary classes of tools are homology modeling for enzymes without experimental structures and *in silico* metabolite docking.

Comparative protein structure modeling (homology modeling) is leveraging the results from experimental structural biology so that useful models of large a numbers of proteins can be obtained (14–17). In early October 2011, the ModBase database (18), developed by one of the authors (A.S.), contained 21,092,755 comparative models for domains in 3,505,676 unique sequences (http://salilab.org/modbase/). Thus, the large number of experimental and predicted structures of unknown proteins enables the use of *in silico* docking to tackle the challenge of high throughput functional prediction.

Virtual screening (*in silico* docking), using computational algorithms to evaluate complementarity between a protein receptor and a virtual library of small molecules, is a widely used strategy in both academia and the pharmaceutical industry to identify lead compounds for drug discovery (19–22). The lead compounds so identified need not have any structural similarity to the natural ligands; an effective inhibitor provides a scaffold on

which substituents are placed to optimize steric and polar interactions with the receptor site. Computational docking can screen extremely large virtual ligand libraries, ranking hits using an energy scoring function to identify those that are predicted to best fit the receptor site.

Until recently (*vide infra*) docking had not been used to screen virtual metabolite libraries for substrates of enzymes. Identification of substrates is a much more difficult problem than identification of inhibitors because drugs only need to "fill" the receptor site so that they can act as competitive inhibitors and do not need to structurally resemble the natural ligand. In contrast, substrates require a precise orientation via specificity-determining residues so that the reactive portion is positioned productively adjacent to the catalytic residues.

Experimental screening of physical ligand libraries is time consuming and inefficient; negative results from experimental screening rarely provide useful information for discovery of the correct ligand. However, virtual screening offers the potential to be a high throughput predictive method that can focus experimental assignment of function to specific substrate candidates, thereby facilitating the discovery of either known or novel substrates. Unlike physical library screening, virtual screening is not limited to known metabolites, commercially available compounds, and/or those that can be readily synthesized. Virtual libraries can include novel substrates as well as structural variants of known metabolites that, based on genome/operon context or physiology, are candidate substrates. If novel compounds are prominent in the energy-ordered list ("hit" list) of predicted substrates, focused synthetic efforts by the Bridging Projects can be justified to test the predictions.

The Computational Core applies these tools to unknown members of the five Bridging Project superfamilies to guide the selection and/or synthesis of specific metabolites or focused libraries for use in enzymatic assays as well as ligand binding screens performed by the Protein Core. The computational methods are subjected to continuous development and refinement as the results of *in silico* docking are compared with the results of enzymatic assays by the Bridging Projects and liganded structures by the Structure Core. As feasible, the Computation Core will collaborate with the community to apply these computational tools to enzymes outside the five superfamilies (*vide infra*).

Microbiology Core

The Microbiology Core examines *in vitro* assigned functions using *in vivo* approaches, including **1)** construction of knockout (null) and overexpression mutants of targets in genetically tractable bacteria; **2)** phenotypic evaluation of wild type and mutant strains in chemically defined media; **3)** transcriptomic analyses of wild type and mutant strains under conditions in which a phenotype is identified; and **4)** mass spectrometric identification of metabolites in wild type and mutant strains grown in chemically defined media to detect and quantitate the abundance of the substrates and products as well as related intermediates in metabolic pathways.

To facilitate *in vivo* studies, most targets for functional assignment are selected from bacterial genomes and, in many cases, from organisms that are genetically tractable so that knockout mutations can be constructed for phenotypic analyses. The extension to *in vivo* function provides a check on the predicted and experimentally confirmed *in vitro* function and, more importantly, allows the metabolic and physiological contexts of novel reactions to be defined. For example, *in vivo* studies may reveal a single substrate for an enzyme that is functionally promiscuous in *in vitro* studies. Alternatively, *in vivo* experiments may reveal that the identity of the *in vitro* enzymatic function does not apply in the context of the organism (*vide infra*) and provide essential information to improve the computational predictions and inform the *in vitro* characterization.

Data/Dissemination Core

The Data/Dissemination Core is responsible for developing and maintaining **1)** the EFI's public website (www.enzymefunction.org) that serves as a resource for information on development of the multidisciplinary strategy and as a "user friendly" portal to the EFI's selected targets, ensuing experimental data, and the computational and experimental tools; **2)** a public database of experimental data (EFI-DB; http://kiemlicz.med.virginia.edu/efi/) that allows interrogation of data gathered on each target, e.g., cloning, purification, and structure determination as well as the results of enzymatic assays and phenotypic/transcriptomic/metabolomic analyses as the latter become publicly available (as determined by NIH policy for data sharing); **3)** an internal database (LabDB) for semi-automated recording and analysis of experimental data to be transferred into EFI-DB; and **4)** the SFLD database that provides highly curated sequence information, links to external databases containing sequence, genomic context, structural, and computationally-derived information for the functionally diverse superfamilies under study by the EFI as well as an expanding number of other superfamilies, e.g., enoyl-CoA hydratase, vicinal oxygen chelate, RuBisCO, nucleophilic-6-bladed beta propeller (N6P), and those of the thioredoxin fold class.

Enzyme Function Initiative (EFI): Bridging Projects

The targets for developing the EFI's multidisciplinary strategy for functional assignment are selected from functionally diverse superfamilies that are the experimental foci of the Bridging Projects. The five Bridging Projects are focused on functionally diverse superfamilies (AH, EN, GST, HAD, and IS) that span four of the six reaction classes defined by the Enzyme Nomenclature Classification System (E.C.) (23) and four fold classes (Figure 5). The selected superfamilies range in size from several thousand to tens of thousand members and differ in domain organization and architecture, substrate chemotypes and structures, metal requirements, and catalytic strategies. They represent a broad sampling of the enzyme universe and, together with associated operon-encoded proteins, provide appropriate targets to develop and test the general utility of the EFI's strategy and inspire further generalization of its methods.

AH Bridging Project

The members of the AH superfamily (~25,000 members) catalyze diverse reactions that involve stabilization of an anionic intermediate by a conserved metal center (one to three Zn^{2+}, Mn^{2+}, Fe^{2+}, or Ni^{2+} metal ions). Most reactions involve hydrolysis of phosphate esters, esters, and amides, although divergent members catalyze 1,2-proton transfer and decarboxylation reactions (7). The polypeptides fold as a single domain that has the ubiquitous $(\beta/\alpha)_8$-barrel (TIM-barrel) fold; thus, both substrate specificity and chemical mechanism are determined by the same domain. The AH superfamily was selected for inclusion in the EFI because **1)** substrate specificity often is defined by flexible loops consisting of residues that determine substrate specificity; **2)** the superfamily is estimated to catalyze a large number of reactions (≥ 100); and **3)** organisms often contain paralogues with different substrate specificities.

EN Bridging Project

The members of the EN superfamily (> 6,000 nonredundant[1] members) catalyze diverse reactions involving a Mg^{2+}-stabilized enolate anion intermediate obtained by abstraction of the α-proton of a carboxylate substrate, including β-elimination (cycloisomerization, dehydration, or deamination) and 1,1-proton transfer (racemization or epimerization)

[1]Nonredundant sequences are those obtained by excluding those that share >98% sequenced identity over 95% of the length of the functional domain.

reactions (5, 6, 24). The polypeptides fold as two domains, with loops in the N-terminal (α+β) capping domain providing determinants for substrate specificity and the C-terminal $(\beta/\alpha)_7\beta$-barrel (TIM-barrel) domain providing the residues that deliver the chemistry. The EN superfamily was selected for inclusion in the EFI because it arguably is the best characterized functionally diverse superfamily and, therefore, provides a "gold standard" set of enzymes/reactions that can be used to test new computational methods by both retrospective and prospective analyses.

GST Bridging Project

The members of the GST superfamily (> 13,000 nonredundant members in the cytosolic GST superfamily) catalyze a diverse range of redox reactions as well as conjugation reactions in xenobiotic metabolism (25–27). The canonical GST superfamily members are composed of an N-terminal domain that has a thioredoxin-like fold and a C-terminal domain that has a unique α-helical fold; the active sites are located at the domain interface. An alternate fold where the thioredoxin-like domain is interrupted by the α-helical domain is also found in eukaryotes and prokaryotes (28, 29). This fold represents the so-called kappa GSTs, another superfamily in the thioredoxin fold class that catalyzes the GST reaction (30, 31). The canonical superfamily harbors members that have robust disulfide bond oxidoreductase activity (32, 33); these enzymes likely utilize proteins as substrates, thereby extending the challenge of functional prediction to protein-protein interactions. The GST superfamily was selected for inclusion in the EFI because a large number of its diverse members have not been characterized with respect to the boundaries between sequence, structure, and function.

HAD Bridging Project

The members of the HAD superfamily (> 32,000 nonredundant members) catalyze a diverse range of reactions that involve the Mg^{2+}-dependent formation of a covalent intermediate with an active site Asp. The reactions include dehalogenation, phosphoryl transfer, and hydrolysis of phosphate esters, phosphate anhydrides, and phosphonates (34–37). The polypeptides share a Rossman-like fold, and most contain a cap module that regulates access of substrates to the active site while providing substrate specificity determinants. Phosphatases are prevalent in the HAD superfamily and often have promiscuous substrate specificities and unknown biological functions. The HAD superfamily was selected for inclusion in the EFI because of the challenges it offers for the development of **1)** computational methods for substrate prediction; and **2)** microbiological- and metabolomic-based strategies for *in vivo* function assignment.

IS Bridging Project

The members of the IS Type 1 superfamily (> 7,600 nonredundant members) catalyze often complex C-C bonding forming reactions initiated by Mg^{2+}-assisted dissociation of a pyrophosphate moiety from an allylic diphosphate substrate followed by reactions/rearrangements in which the conformations of electrophilic carbocation intermediates relative to nucleophilic double bonds determine the structure of the product (38, 39). The polypeptides share an α-helical bundle fold, with the shape of the active sites controlling the conformation of the bound substrate and, therefore, the identity of the product. Unlike the other EFI superfamilies, the range of substrates is almost exclusively limited to only one homoallylic and four allylic diphosphate substrates, so the functional assignment challenge is primarily product prediction. The IS superfamily also was selected for inclusion in the EFI because only a small number of sequences has been functionally characterized, so priorities for target selection are difficult to define.

Integration of the Cores and Bridging Projects: The Integrated Strategy

Success of the EFI's integrated sequence/structure-based strategy to facilitate functional assignment will be judged by its ability to facilitate the discovery of new functions for enzymes of enormous diversity. Interactions between the components of the pipeline for target selection and downstream evaluation of functional predictions are critical for optimizing this strategy (Figure 3).

EFI Target Selection

The Superfamily/Genome, Protein, Structure, Computation, and Microbiology Cores together with the Bridging Projects collaborate on the selection of targets. The Superfamily/Genome Core collects member sequences for each of the superfamilies and defines sequence and structure boundaries expected to be useful for identification of isofunctional families, based on multiple bioinformatic analyses, including similarity networks (e.g., Figure 4). Using this visualization as well as information about genome/operon context, divergent families are identified and then evaluated by the Computation Core to assess feasibility for ligand docking as well as whether the various families provide challenges for docking that allow the enhancement of the computational algorithms. The Bridging Projects contribute their accumulated experimental experience to reveal insights into possible functions and substrates, based on conservation of active site functional groups and divergence of specificity-determining residues. The Protein and Structure Cores provide input about feasibility of protein expression, purification, crystallization, and structure determination, based on accumulated experiences for members of each of the superfamilies, e.g., position of affinity tags for protein purification, exploration of fermentation conditions to optimize metal loading, genome availability, and gene synthesis. Finally, the Microbiology Core provides information about genetic tractability. Although some targets are selected to explore divergent sequence and, therefore, function space, many targets are chosen to address specific scientific questions such as exploring the boundaries between substrate specificities as sequence diverges. The latter targets provide the ability to test and develop the computational algorithms on homologous proteins as the sequence similarity decreases.

EFI Target Initiation

Selected targets are communicated to the Protein Core for inclusion in the "pipeline" (Figure 3) for gene cloning, protein expression and purification, ThermoFluor screening, and experimental structure determination by the Structure Core. Protein samples also are provided to the Bridging Projects for focused library screening and enzymatic assays to test the substrate specificity predictions from the Computation Core.

Structural Characterization of EFI Targets

When the Computation Core concludes that an existing liganded structure shares sufficient similarity with the target, a homology model is generated to provide a template for docking, thereby providing a faster and higher throughput approach for computational predictions of substrate specificity. In such cases, the type of reaction catalyzed by the template, e.g., acid sugar dehydration in the EN superfamily, may be the type of reaction catalyzed by the target, with the template and target differing in substrate specificity.

Parallel protein production by the Protein Core and structure determination by the Structure Core occurs when possible so that **1)** the predicted substrate specificity can be experimentally tested by the Bridging Projects, and **2)** the accuracy of the "pose" of the liganded active site predicted by the Computation Core can be assessed, thereby validating the results of the docking predictions.

Generation of Functional Predictions for EFI Targets

Irrespective of the method by which the target structure is obtained, the Computation Core uses its methodologies, including flexible receptor (40, 41) and high-energy intermediate (HEI) docking (42, 43), to assemble energy-ordered hit lists of predicted substrates. In flexible receptor docking, the rotameric conformations of the side chains of the active site residues are varied to identify the lowest energy complex; in HEI docking, the structures of reactive intermediates, e.g., tetrahedral intermediates hydrolyses of esters and amides, are used for docking. These hit lists are computationally filtered to prioritize the identities of focused substrate libraries for the Bridging Projects.

Testing of Functional Predictions for EFI Targets

The Bridging Projects in combination with the Microbiology Core evaluate the functional predictions (substrate hit lists) provided by the Computation Core and procure (by purchase, in-house synthesis, and/or custom synthesis) predicted substrates for use in focused libraries for enzymatic assays. In the AH, EN, and HAD superfamilies, the reactions are either unimolecular or use water as a co-substrate, so identification of the substrate is equivalent to function prediction. In the GST superfamily, glutathione (or spermidinylglutathione) is always a substrate, so the *in silico* docking predicts the co-substrate and, therefore, the function. In the IS superfamily, the identities of both the predicted substrate(s) (from a set of five allylic pyrophosphates) and predicted product are tested.

Functional Assignment and Rescue of EFI Targets

Criteria for deciding the flow of targets through the experimental (enzymological and microbiological) components to functional assignment include (Figure 3):

1. If the kinetic constants for the *in vitro* function are consistent with those expected for a typical metabolic enzyme, e.g., $k_{cat}/K_M \geq 10^4\ M^{-1}\ sec^{-1}$ (44), and the target is from a tractable organism, it is referred to the Microbiology Core for genetic, phenotypic, transcriptomic, and/or metabolomic "confirmation". If the predicted reaction is catalyzed, but the value of k_{cat}/K_M is less than expected, the substrate/product/analog is provided to the Structure Core for cocrystallization, and the resulting liganded structure is provided to the Computation Core for additional *in silico* ligand docking.
2. If the predicted reaction is not catalyzed but another reaction is identified with "unfocused" library screening by the Bridging Project, the substrate/product/analog for that reaction is provided to the Structure Core for cocrystallization, and the liganded structure is provided to the Computation Core for assessment of prediction failure. Such situations are instructive, in fact essential, for development of the strategy because the structure-based explanation for an incorrect predicted function suggests how the algorithms for docking and/or homology modeling can be improved.
3. If no reaction is identified, the target is placed "on hold" for salvage as the integrated strategy is improved.

Successful Examples of the Integrated Strategy

The feasibility of using *in silico* ligand docking to facilitate functional assignment was demonstrated in a smaller program focused on the AH and EN superfamilies (J.A.G, S.C.A., P.C.B., M.P.J, F.M.R., A.S., and B.K.S.). Those efforts resulted in several successful focused predictions of substrate specificity in the functionally diverse AH (43, 45–48) and EN (41, 49, 50) superfamilies. Recently, this methodology was used to generate high

throughput substrate specificity predictions for the entire dipeptide epimerase family in the EN superfamily[3].

Noteworthy among these examples is the prediction of the function of an unknown member of the AH superfamily encoded by the *Thermotoga maritima* genome (Tm0936) as S-adenosylhomocysteine deaminase, a novel enzymatic reaction. This prediction was accomplished by docking a library of high-energy intermediates to the three dimensional structures determined by PSI-2 centers (1P1M and 1J6P) (43) (Figure 6, panel A). In another example, the N-succinyl Arg racemase function was predicted for a member of the *cis,cis*-muconate lactonizing enzyme (MLE) subgroup of the EN superfamily encoded by *Bacillus cereus* ATCC 14579 (BC0371). This prediction was accomplished by flexible receptor docking of a virtual library of dipeptides and N-succinyl amino acids to a homology model generated using the structure of the L-Ala-D/L-Glu epimerase from *B. subtilis* (1TKK) as the template (41) (Figure 6, panel B). Finally, new specificities for many of the >700 members of the dipeptide epimerase family in the EN superfamily were predicted by *in silico* docking to homology models based on the 1TKK template and experimentally verified by enzymology; in addition, several of the liganded structures were determined by X-ray crystallography, allowing validation of the liganded active site models. Based on these results, virtually all of the predicted dipeptide epimerases in the enolase superfamily can be annotated; these annotations will be made available in the SFLD.

Challenges for Development of the Integrated Strategy

Despite these examples of success, *in silico* ligand docking is not always successful in correctly predicting substrate specificities. One reason for failure is that experimentally determined structures are not necessarily in “dockable” conformations, e.g., substrates often induce conformational changes, and the conformational sampling methods are not capable of finding the bound conformation. One way to circumvent this problem is to screen unknown enzymes for ligand/substrate fragment binding *via* thermal stabilization using libraries of small molecule substrate fragments and/or potential mimics of intermediates, e.g., hydroxamates for enolate anions in the case of the EN superfamily. Such scanning can be monitored in a high throughput manner using the ThermoFluor assay that can measure binding of a hydrophobic dye as a function of temperature in a 96-well format (12, 13); the infrastructure for these analyses has been implemented by the Protein Core.

Another reason for incorrect predictions is that while the actual substrate may be present in the docking hit list, it may not score highly due to inaccurate scoring functions. Improved prediction specificity may be possible by the addition of orthogonal information. For example, when the target participates in a metabolic pathway and its gene is encoded by an operon that encodes other enzymes in the pathway, common characteristics among the ligands of enzymes that catalyze successive reactions in a pathway may be revealed by *in silico* docking results for *all* enzymes in the pathway, thereby providing functional clues that restrict the identities of the substrates for each of the enzymes. This approach has been illustrated by a retrospective analysis of the glycolysis pathway in *Escherichia coli* published by one of us (M.P.J.) (51).

As described earlier, the role of the Microbiology Core is to provide *in vivo* evaluation of *in vitro* assigned functions in the context of physiology as well as assign the metabolic roles of functions in novel metabolic pathways. For example, the metabolic role of the N-succinyl Arg racemase reaction that was computationally predicted and experimentally verified

[3]T. Lukk, A. Sakai, C. Kalyanaraman, S. Brown, H. J. Imker, L. Song, A. A. Fedorov, E. V. Fedorov, R. Toro, B. Hillerich, R. Seidel, Y. Patskovsky, M. V. Vetting, S. K. Nair, P. C. Babbitt, S. C. Almo, J. A Gerlt, and M. P. Jacobson, manuscript submitted.

remains unknown (41). Although the encoding gene is not located in an operon, the metabolic function may be the conversion of D- to L-amino acids *via* N-succinylated intermediates (52); a knockout of the encoding gene in the *B. cereus* ATCC 14579 genome may provide a phenotype, and these experiments are underway. Determination of phenotypes for knockouts under a wide range of growth conditions may be necessary to discover the *in vivo* function, although testing for utilization of D-amino acids as nitrogen source may be sufficient. In either case, the Microbiology Core has implemented a high throughput platform for phenotypic analyses of metabolic activity by using a BioLog PM instrument that allows as many as 4800 growth conditions to be simultaneously examined in a 96-well plate format.

The Microbiology Core already has discovered an example of an *in vitro* assigned function that is "incorrect" in the context of the encoding organism's metabolism. In the RuBisCO superfamily, one of our laboratories (H.J.I. and J.A.G.) characterized a novel 1,3-proton transfer to a RuBisCO-like protein (RLP) from *Rhodospirillum rubrum* in which 5-methylthio-D-ribulose 5-phosphate is converted to a 3:1 mixture of a 1-methylthio-ribulose/xylulose 5-phosphate in two successive 1,2-proton transfer reactions (Scheme I). The identities of the reaction products were established using ^{1}H, ^{13}C, and ^{32}P NMR spectroscopy and mass spectrometry (53). However, the Microbiology Core has obtained evidence that the first 1,2-proton transfer reaction to generate the "3-ulose" intermediate is the physiological reaction in *R. rubrum*.[4] The "4-ulose" product obtained *in vitro* by the second 1,2-proton transfer reaction is the thermodynamically most stable isomer of the substrate and apparently accumulates if the "3-ulose" species is not utilized as substrate by the next enzyme in the pathway.

EFI Interactions with the Community

As noted in the section entitled "Enzyme Function Initiative (EFI): Overview", the EFI's deliverables include not only development of the multidisciplinary strategy for functional assignment using targets selected from the EFI's five functionally diverse superfamilies but also dissemination of the strategy to the community. At this early stage of the EFI, we are focused on developing high throughput, yet still high quality, tools for the strategy, including bioinformatics analyses by the Superfamily/Genome Core as well as modeling and docking tools by the Computation Core. Our resources for the "wet" experimental aspects of the integrated strategy are more limited and currently restricted to the five Bridging Projects. However, the EFI has sufficient resources for establishing collaborations of the Superfamily/Genome and Computation Cores with the scientific community to facilitate assignments of function in other functionally diverse superfamilies.

In most cases, we expect that these collaborations will involve initial interactions with the Superfamily/Genome Core so that sequence similarity networks can be constructed for a given superfamily, thereby providing an overview of the extent of functional diversity (isofunctional families) within the superfamily. This information also will facilitate selecting the most viable targets for addressing specific functional assignment problems and may be sufficient for subsequent investigations of functional assignments in the collaborator's laboratory. However, depending on the availability of experimental structures or homology models, we expect that the sequence similarity networks also will encourage selection of targets for *in silico* ligand docking and substrate prediction by the Computation Core; these predictions would then be tested in the collaborator's laboratory.

[4]T. J. Erb, K. Choi, B. S. Evans, J. Singh, B. M. Wood, J. V. Sweedler, J. E. Cronan, R. F. Tabita, and J. A. Gerlt, manuscript submitted.

The EFI encourages the community to propose collaborations that involve functionally diverse superfamilies that currently are not within its focus. This can be accomplished by completing a form on the "Collaborations" page of the EFI's website (enzymefunction.org/collaborations/overview) that includes a short proposal (two or three paragraphs) for the expected nature and scope of the collaboration as well as several leading references about the superfamily. For example, these collaborations may involve investigations of the functions of specific divergent members of functionally diverse superfamilies, with the interactions with the Superfamily/Genome Core providing sequence/family context and with the Computation Core providing predictions of substrate specificities and, therefore, enzymatic functions. We also encourage interested members of the community to register on the "Home" page of the EFI website for e-mail alerts that will provide news and updates.

The EFI will organize workshops as well as symposia at scientific meetings that will involve participation not only by the EFI PIs but also members of the scientific community who are actively involved or interested in the challenges presented by functional assignment of unknown enzymes. We expect that the EFI will catalyze interactions in the community to facilitate the development of the strategy and necessary methodologies for comprehensive enzyme annotation.

Summary

Assigning reliable functions to unknown enzymes discovered in genome projects is a complex yet critical challenge that will only increase in magnitude as the databases continue to expand. While many strategies ultimately may be required, the EFI is taking the lead by developing and disseminating a systematic and robust approach to meet this challenge. We anticipate that enhancing the ability of bioinformatics to identify informative sequence relationships, of homology modeling to allow accurate high throughput structure prediction, and of *in silico* ligand docking to provide accurate and testable hit lists of potential targets will be valuable contributions by the EFI. Coupling these advances with the enzymology community's extensive experimental knowledge will contribute to a "new era" of enzymology in which genomic sequence information facilitates a wider range of intellectual, physiological, biomedical, and commercial applications.

References

1. Schnoes AM, Brown SD, Dodevski I, Babbitt PC. Annotation error in public databases: misannotation of molecular function in enzyme superfamilies. PLoS Comput Biol. 2009; 5:e1000605. [PubMed: 20011109]
2. Gerlt JA. A Protein Structure (or Function ?) Initiative. Structure. 2007; 15:1353–1356. [PubMed: 17997960]
3. Babbitt PC, Gerlt JA. Understanding enzyme superfamilies. Chemistry as the fundamental determinant in the evolution of new catalytic activities. J Biol Chem. 1997; 272:30591–30594. [PubMed: 9388188]
4. Gerlt JA, Babbitt PC. DIVERGENT EVOLUTION OF ENZYMATIC FUNCTION: Mechanistically Diverse Superfamilies and Functionally Distinct Suprafamilies. Annu Rev Biochem. 2001; 70:209–246. [PubMed: 11395407]
5. Gerlt JA, Babbitt PC, Rayment I. Divergent evolution in the enolase superfamily: the interplay of mechanism and specificity. Arch Biochem Biophys. 2005; 433:59–70. [PubMed: 15581566]
6. Gerlt JA, Babbitt PC, Jacobson MP, Almo SC. Divergent Evolution in the Enolase Superfamily: Strategies for Assigning Function. J Biol Chem. 2011; 286 in press.
7. Seibert CM, Raushel FM. Structural and catalytic diversity within the amidohydrolase superfamily. Biochemistry. 2005; 44:6383–6391. [PubMed: 15850372]

8. Brown SD, Gerlt JA, Seffernick JL, Babbitt PC. A gold standard set of mechanistically diverse enzyme superfamilies. Genome Biol. 2006; 7:R8. [PubMed: 16507141]
9. Atkinson HJ, Morris JH, Ferrin TE, Babbitt PC. Using sequence similarity networks for visualization of relationships across diverse protein superfamilies. PLoS One. 2009; 4:e4345. [PubMed: 19190775]
10. Pegg SC, Brown SD, Ojha S, Seffernick J, Meng EC, Morris JH, Chang PJ, Huang CC, Ferrin TE, Babbitt PC. Leveraging enzyme structure-function relationships for functional inference and experimental design: the structure-function linkage database. Biochemistry. 2006; 45:2545–2355. [PubMed: 16489747]
11. Cline MS, Smoot M, Cerami E, Kuchinsky A, Landys N, Workman C, Christmas R, Avila-Campilo I, Creech M, Gross B, Hanspers K, Isserlin R, Kelley R, Killcoyne S, Lotia S, Maere S, Morris J, Ono K, Pavlovic V, Pico AR, Vailaya A, Wang PL, Adler A, Conklin BR, Hood L, Kuiper M, Sander C, Schmulevich I, Schwikowski B, Warner GJ, Ideker T, Bader GD. Integration of biological networks and gene expression data using Cytoscape. Nat Protoc. 2007; 2:2366–2382. [PubMed: 17947979]
12. Cummings MD, Farnum MA, Nelen MI. Universal screening methods and applications of ThermoFluor. J Biomol Screen. 2006; 11:854–863. [PubMed: 16943390]
13. Ericsson UB, Hallberg BM, Detitta GT, Dekker N, Nordlund P. Thermofluor-based high throughput stability optimization of proteins for structural studies. Anal Biochem. 2006; 357:289–298. [PubMed: 16962548]
14. Baker D, Sali A. Protein structure prediction and structural genomics. Science. 2001; 294:93–96. [PubMed: 11588250]
15. Wallner B, Elofsson A. All are not equal: a benchmark of different homology modeling programs. Protein Sci. 2005; 14:1315–1327. [PubMed: 15840834]
16. Eswar N, Webb B, Marti-Renom MA, Madhusudhan MS, Eramian D, Shen MY, Pieper U, Sali A. Comparative protein structure modeling using Modeller. Curr Protoc Bioinformatics. 2006; Chapter 5(Unit 5–6)
17. Eswar N, Eramian D, Webb B, Shen MY, Sali A. Protein structure modeling with MODELLER. Methods Mol Biol. 2008; 426:145–159. [PubMed: 18542861]
18. Pieper U, Webb BM, Barkan DT, Schneidman-Duhovny D, Schlessinger A, Braberg H, Yang Z, Meng EC, Pettersen EF, Huang CC, Datta RS, Sampathkumar P, Madhusudhan MS, Sjolander K, Ferrin TE, Burley SK, Sali A. ModBase, a database of annotated comparative protein structure models, and associated resources. Nucleic Acids Res. 2010; 39:D465–474. [PubMed: 21097780]
19. Kolb P, Ferreira RS, Irwin JJ, Shoichet BK. Docking and chemoinformatic screens for new ligands and targets. Curr Opin Biotechnol. 2009; 20:429–436. [PubMed: 19733475]
20. Leach AR, Shoichet BK, Peishoff CE. Prediction of protein-ligand interactions. Docking and scoring: successes and gaps. J Med Chem. 2006; 49:5851–5855. [PubMed: 17004700]
21. Shoichet BK. Virtual screening of chemical libraries. Nature. 2004; 432:862–865. [PubMed: 15602552]
22. Shoichet BK, McGovern SL, Wei B, Irwin JJ. Lead discovery using molecular docking. Curr Opin Chem Biol. 2002; 6:439–446. [PubMed: 12133718]
23. Webb, EC. Enzyme Nomenclature 1992: Recommendations of the Nomenclature Committee of the International Union of Bicohemistry and Molecular Biology on the Nomenclature and Classification of Enzymes. Academic Press; San Diego, CA: 1992.
24. Babbitt PC, Hasson MS, Wedekind JE, Palmer DR, Barrett WC, Reed GH, Rayment I, Ringe D, Kenyon GL, Gerlt JA. The enolase superfamily: a general strategy for enzyme-catalyzed abstraction of the alpha-protons of carboxylic acids. Biochemistry. 1996; 35:16489–16501. [PubMed: 8987982]
25. Armstrong RN. Structure, catalytic mechanism, and evolution of the glutathione transferases. Chem Res Toxicol. 1997; 10:2–18. [PubMed: 9074797]
26. Armstrong RN. Mechanistic imperatives for the evolution of glutathione transferases. Curr Opin Chem Biol. 1998; 2:618–623. [PubMed: 9818188]
27. Armstrong, RN. Glutathione Transferases. In: Guengerich, FP., editor. Comprehensive Toxicology. 2. Elsevier Science; Oxford: 2010. p. 295-321.

28. Ladner JE, Parsons JF, Rife CL, Gilliland GL, Armstrong RN. Parallel evolutionary pathways for glutathione transferases: structure and mechanism of the mitochondrial class kappa enzyme rGSTK1-1. Biochemistry. 2004; 43:352–361. [PubMed: 14717589]
29. Thompson LC, Ladner JE, Codreanu SG, Harp J, Gilliland GL, Armstrong RN. 2-Hydroxychromene-2-carboxylic acid isomerase: a kappa class glutathione transferase from Pseudomonas putida. Biochemistry. 2007; 46:6710–6722. [PubMed: 17508726]
30. Atkinson HJ, Babbitt PC. An atlas of the thioredoxin fold class reveals the complexity of function-enabling adaptations. PLoS Comput Biol. 2009; 5:e1000541. [PubMed: 19851441]
31. Atkinson HJ, Babbitt PC. Glutathione transferases are structural and functional outliers in the thioredoxin fold. Biochemistry. 2009; 48:11108–11116. [PubMed: 19842715]
32. Wadington MC, Ladner JE, Stourman NV, Harp JM, Armstrong RN. Analysis of the structure and function of YfcG from Escherichia coli reveals an efficient and unique disulfide bond reductase. Biochemistry. 2009; 48:6559–6561. [PubMed: 19537707]
33. Stourman NV, Branch MC, Schaab MR, Harp JM, Ladner JE, Armstrong RN. Structure and function of YghU, a nu-class glutathione transferase related to YfcG from Escherichia coli. Biochemistry. 50:1274–1281. [PubMed: 21222452]
34. Lahiri SD, Zhang G, Dai J, Dunaway-Mariano D, Allen KN. Analysis of the substrate specificity loop of the HAD superfamily cap domain. Biochemistry. 2004; 43:2812–2820. [PubMed: 15005616]
35. Allen KN, Dunaway-Mariano D. Phosphoryl group transfer: evolution of a catalytic scaffold. Trends Biochem Sci. 2004; 29:495–503. [PubMed: 15337123]
36. Burroughs AM, Allen KN, Dunaway-Mariano D, Aravind L. Evolutionary genomics of the HAD superfamily: understanding the structural adaptations and catalytic diversity in a superfamily of phosphoesterases and allied enzymes. J Mol Biol. 2006; 361:1003–1034. [PubMed: 16889794]
37. Lu Z, Dunaway-Mariano D, Allen KN. The catalytic scaffold of the haloalkanoic acid dehalogenase enzyme superfamily acts as a mold for the trigonal bipyramidal transition state. Proc Natl Acad Sci U S A. 2008; 105:5687–5692. [PubMed: 18398008]
38. Thulasiram HV, Poulter CD. Farnesyl diphosphate synthase: the art of compromise between substrate selectivity and stereoselectivity. J Am Chem Soc. 2006; 128:15819–15823. [PubMed: 17147392]
39. Thulasiram HV, Erickson HK, Poulter CD. Chimeras of two isoprenoid synthases catalyze all four coupling reactions in isoprenoid biosynthesis. Science. 2007; 316:73–76. [PubMed: 17412950]
40. Sherman W, Day T, Jacobson MP, Friesner RA, Farid R. Novel procedure for modeling ligand/receptor induced fit effects. J Med Chem. 2006; 49:534–553. [PubMed: 16420040]
41. Song L, Kalyanaraman C, Fedorov AA, Fedorov EV, Glasner ME, Brown S, Imker HJ, Babbitt PC, Almo SC, Jacobson MP, Gerlt JA. Prediction and assignment of function for a divergent N-succinyl amino acid racemase. Nat Chem Biol. 2007; 3:486–491. [PubMed: 17603539]
42. Hermann JC, Ghanem E, Li Y, Raushel FM, Irwin JJ, Shoichet BK. Predicting substrates by docking high-energy intermediates to enzyme structures. J Am Chem Soc. 2006; 128:15882–15891. [PubMed: 17147401]
43. Hermann JC, Marti-Arbona R, Fedorov AA, Fedorov E, Almo SC, Shoichet BK, Raushel FM. Structure-based activity prediction for an enzyme of unknown function. Nature. 2007; 448:775–779. [PubMed: 17603473]
44. Bar-Even A, Noor E, Savir Y, Liebermeister W, Davidi D, Tawfik DS, Milo R. The moderately efficient enzyme: evolutionary and physicochemical trends shaping enzyme parameters. Biochemistry. 2011; 50:4402–4410. [PubMed: 21506553]
45. Xiang DF, Kolb P, Fedorov AA, Meier MM, Fedorov LV, Nguyen TT, Sterner R, Almo SC, Shoichet BK, Raushel FM. Functional annotation and three-dimensional structure of Dr0930 from Deinococcus radiodurans, a close relative of phosphotriesterase in the amidohydrolase superfamily. Biochemistry. 2009; 48:2237–2247. [PubMed: 19159332]
46. Cummings JA, Nguyen TT, Fedorov AA, Kolb P, Xu C, Fedorov EV, Shoichet BK, Barondeau DP, Almo SC, Raushel FM. Structure, mechanism, and substrate profile for Sco3058: the closest bacterial homologue to human renal dipeptidase. Biochemistry. 2010; 49:611–622. [PubMed: 20000809]

47. Hall RS, Fedorov AA, Marti-Arbona R, Fedorov EV, Kolb P, Sauder JM, Burley SK, Shoichet BK, Almo SC, Raushel FM. The hunt for 8-oxoguanine deaminase. J Am Chem Soc. 2010; 132:1762–1763. [PubMed: 20088583]

48. Kamat SS, Fan H, Sauder JM, Burley SK, Shoichet BK, Sali A, Raushel FM. Enzymatic deamination of the epigenetic base N-6-methyladenine. J Am Chem Soc. 2011; 133:2080–2083. [PubMed: 21275375]

49. Kalyanaraman C, Imker HJ, Fedorov AA, Fedorov EV, Glasner ME, Babbitt PC, Almo SC, Gerlt JA, Jacobson MP. Discovery of a dipeptide epimerase enzymatic function guided by homology modeling and virtual screening. Structure. 2008; 16:1668–1677. [PubMed: 19000819]

50. Rakus JF, Kalyanaraman C, Fedorov AA, Fedorov EV, Mills-Groninger FP, Toro R, Bonanno J, Bain K, Sauder JM, Burley SK, Almo SC, Jacobson MP, Gerlt JA. Computation-facilitated assignment of the function in the enolase superfamily: a regiochemically distinct galactarate dehydratase from Oceanobacillus iheyensis. Biochemistry. 2009; 48:11546–11558. [PubMed: 19883118]

51. Kalyanaraman C, Jacobson MP. Studying enzyme-substrate specificity in silico: a case study of the Escherichia coli glycolysis pathway. Biochemistry. 2010; 49:4003–4005. [PubMed: 20415432]

52. Sakai A, Xiang DF, Xu C, Song L, Yew WS, Raushel FM, Gerlt JA. 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. Biochemistry. 2006; 45:4455–4462. [PubMed: 16584181]

53. Imker HJ, Singh J, Warlick BP, Tabita FR, Gerlt JA. Mechanistic diversity in the RuBisCO superfamily: a novel isomerization reaction catalyzed by the RuBisCO-like protein from Rhodospirillum rubrum. Biochemistry. 2008; 47:11171–11173. [PubMed: 18826254]

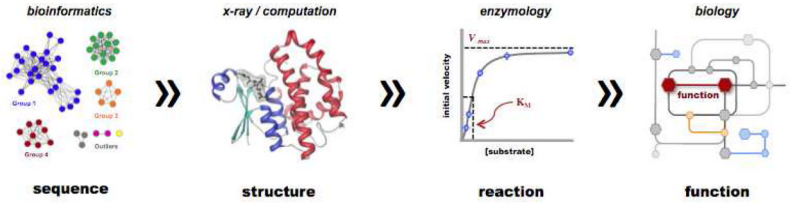


Figure 1.
The goal of the EFI is to develop a multidisciplinary, high throughput strategy for functional assignment of unknown enzymes.

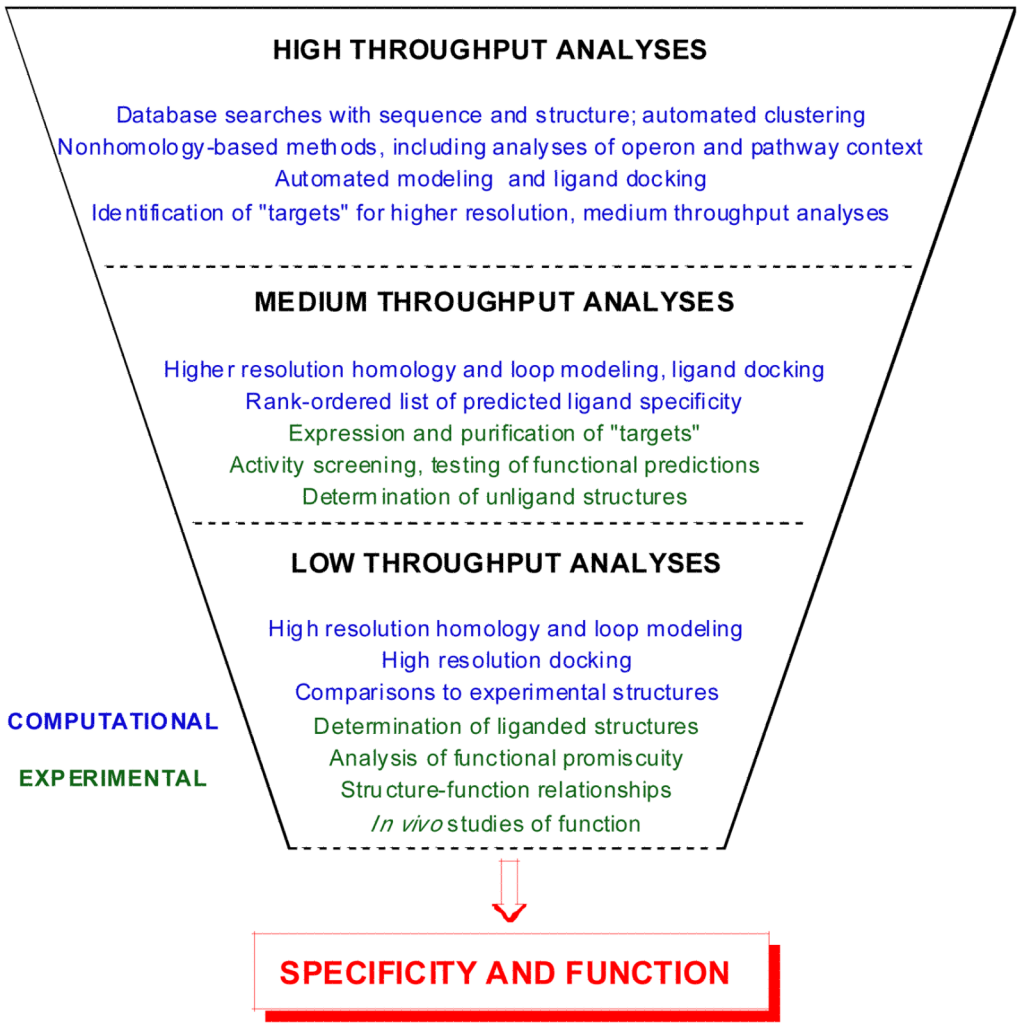


Figure 2.
The “funnel” for functional assignment, showing the roles and relative throughputs of the computational and experimental stages in functional assignment.

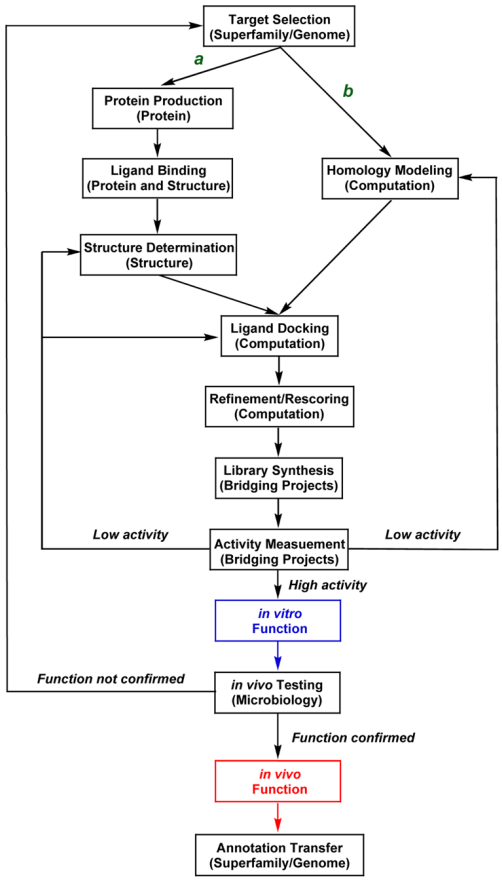


Figure 3.
The pipeline for functional assignment adopted by the EFI.

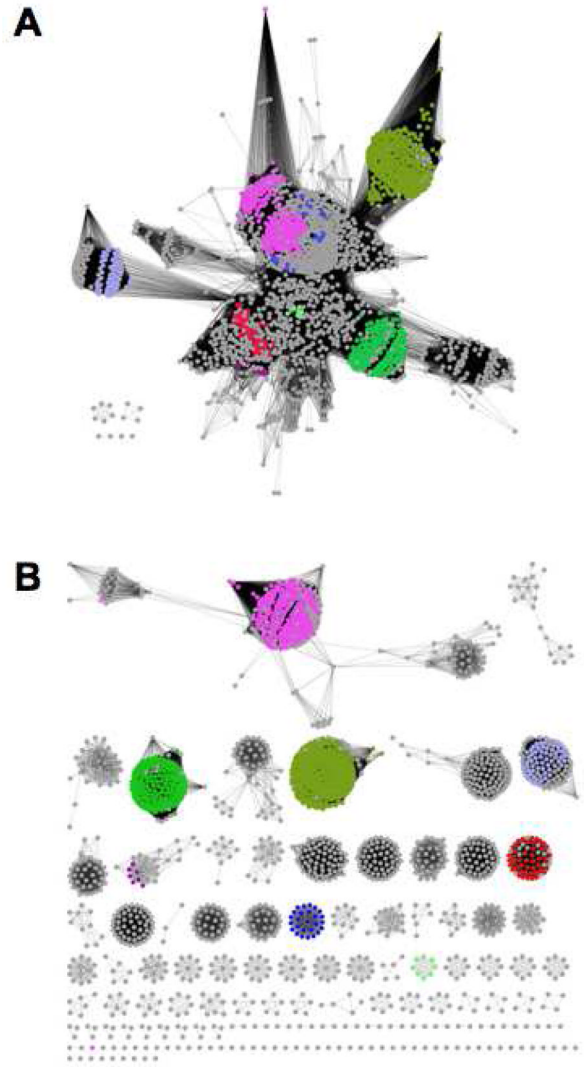


Figure 4.
Representative sequence similarity networks for the mandelate racemase (MR) subgroup of the enolase superfamily. Sequences are shown as nodes (dots); connections with BLASTP E-values more stringent than a specified threshold are shown as edges (lines). Panel A, BLASTP E-value < 10^{-40}. Panel B, BLASTP E-value < 10^{-80}. As the BLASTP E-value threshold is made more stringent, the sequences separate into discrete clusters; at < 10^{-80}, many of the clusters are isofunctional families. Nodes colored grey have unknown functions.

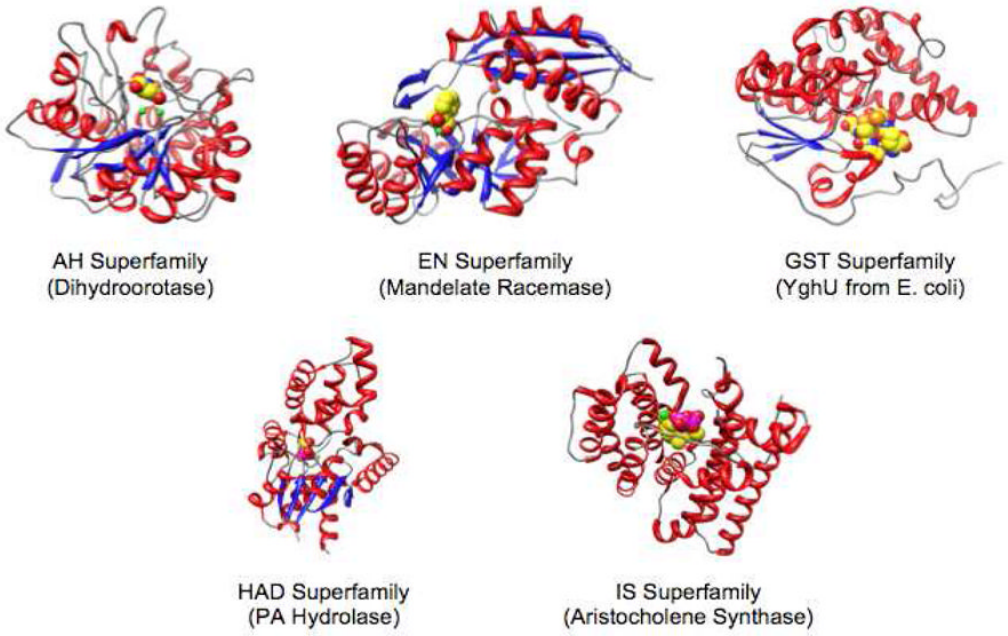


Figure 5.
The architectures/folds for the five functionally diverse superfamilies from which targets are selected for development of the EFI's multidisciplinary functional assignment strategy.

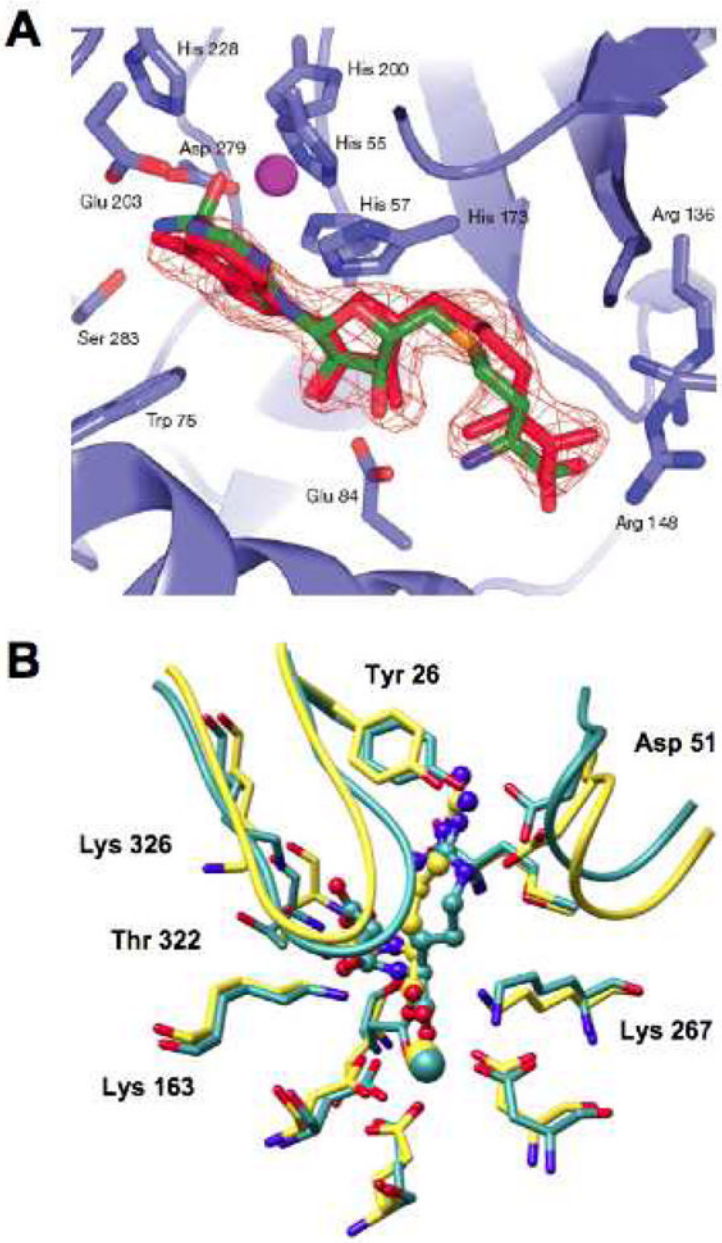
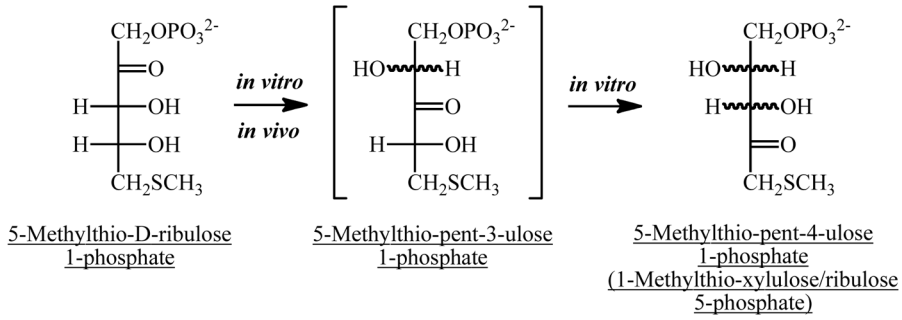


Figure 6.

Panel A, Tm0936 (AH superfamily). Computationally predicted pose of the high-energy intermediate (green) superimposed on experimental structure (red, with electron density contours) (43). Panel B, BC0371 (EN superfamily) in complex with substrate N-succinyl Arg, as predicted by homology modeling and docking (cyan) as well as determined by crystallography (yellow) (41). Both panels are reproduced with permission from the publisher.



5-Methylthio-D-ribulose 1-phosphate

5-Methylthio-pent-3-ulose 1-phosphate

5-Methylthio-pent-4-ulose 1-phosphate (1-Methylthio-xylulose/ribulose 5-phosphate)

Scheme 1.