A Multidisciplinary Strategy for Function Assignment: The Enzyme Function Initiative from a Bioinformatics Perspective

Daniel E. Almonacid, Shoshana D. Brown and Patricia C. Babbitt

Departments of Bioengineering and Therapeutic Sciences, and Pharmaceutical Chemistry, and California Institute for Quantitative Biosciences, University of California San Francisco, San Francisco, California, USA

The Enzyme Function Initiative (EFI) is a multidisciplinary collaboration between 14 groups aiming at developing a large-scale sequence/structure-based strategy for assigning the in vitro functions and in vivo contexts of unknown enzymes from five functionally diverse superfamilies [1]. The EFI is composed of six scientific cores (superfamily/genome, protein, structure, computation, microbiology, and data/dissemination) that work together in five orthogonal bridging projects each focusing on a different functionally diverse superfamily of enzymes (amidohydrolase, enolase, glutathione transferase, haloalkanoic acid dehalogenase, and isoprenoid synthase (IS)).

The pipeline for functional assignment adopted by the EFI starts and ends with bioinformatics scientific cores (superfamily/genome and data/dissemination cores). The superfamily/genome core collects sequences/structures and functions for all members of the EFI superfamilies. Automated scripts implementing sequences/structure methods are used to identify new members of each superfamily, and then sequence similarity networks are used to visualize the relationships among members of these large groups of homologous proteins [2] (Figure 1). All the information is maintained in our Structure–Function Linkage Database (SFLD) [3], a freely available resource that can be found online at: http://sfld.rbvi.ucsf.edu. Using the information in the SFLD, new targets for developing the strategy for functional assignment are indentified and prioritized. Once targets are selected, the pipeline continues with the protein and structure cores focusing on protein production, ligand binding and structure determination. Then the computation core focuses on homology modeling and ligand docking, and finally the experimentalists from the specific bridging projects and from the microbiology core synthesize libraries, measure activities and perform in vivo testing. The pipeline then cycles back to the superfamily/genome and data/dissemination cores where the structures and functions determined are curated into isofunctional families within the context of the superfamilies, and where annotation transfer to other functionally uncharacterized enzymes is performed. Public access is provided to the information through our SFLD as well as two other public websites: the EFI’s website (http://www.enzymefunction.org), and a website containing experimental data (http://kiemlicz.med.virginia.edu/efi/).

Here we will discuss the work performed at the bioinformatics scientific cores of the EFI in relation to the IS bridging project. By the time the EFI started in May 2010, we identified 4407 non-redundant enzyme sequences having high sequence similarity and matching a hidden Markov model based on an alignment of α-helical bundle fold enzymes that had been characterized to catalyze C-C bond forming reactions. All these reactions are initiated by Mg2+-assisted dissociation of a pyrophosphate moiety from an allylic diphosphate substrate with concomitant generation of a carboxylation intermediate, which can then follow a number of different mechanistic routes. What it is interesting about this superfamly of enzymes is that they produce tens of thousands of different natural products, with biological functions including fragrances and flavors, pheromones, defensive agents, antitumor drugs, signal transduction components, membrane constituents, precursor of steroid hormones and bile acids, photoreceptive agents, cofactor side-chains and natural polymers, many of which have industrial and biomedical applications [4,5]. Yet the substrates are limited to only one homoallylic and four allylic diphosphate substrates, so the challenge for functional assignment is determination of product specificity and not substrate specificity like in the rest of the superfamilies targeted by the EFI. The 4407 sequences clustered into 5 distinctive functional subgroups (Figure 1): polyprenyl synthetase like, trichodiene synthase like, squalene/phytoene synthase like, and two subgroups of terpene synthases, one containing a single-domain structure (terpene synthase like 2), and one containing a two-domain structure where the C-terminal domain belongs to the IS superfamily (terpene synthase like 1 C-term) and the N-terminal domain is from a different superfamily. At that point most knowledge about the superfamily resided in the terpene synthase like 1 C-term subgroup so target selection for year one of the EFI focused mainly in the less studied but most populous polyprenyl synthetase like subgroup.
Of 205 selected polyprenyl synthetases, 85 were purified, 78 functionally characterized and 21 structurally characterized. This new knowledge, together with already published structures and functions, has allowed us to manually curate 12 isofunctional families within the polyprenyl synthetase subgroup in the SFLD. It is evident at this time, however, that many families present cases of pseudo-convergent evolution, i.e. the function of the family has been invented multiple times along the evolution of the superfamily and therefore sequences in these families will be further split into two or more groups. We are currently working on developing annotation transfer rules, so that the experimental information for the 12 families can be extrapolated to unknown enzymes from the superfamily, as well as to disseminate these predictions through our SFLD. Work in the computational core is also underway for large-scale homology modeling followed by induced fit in silico docking for prediction of product specificity.

Now that we are half-way through year two of the initiative, the IS superfamily consists of 9925 non-redundant enzyme sequences, of which 91% are assigned to one of five subgroups (colored nodes in Figure 1), but only 3% are assigned to one of 40 isofunctional families. To continue to develop our strategy for function assignment we are currently selecting new targets, mainly sequences that may catalyze cyclopropanations reactions (squalene/phytoene synthase like subgroup) and cyclization reactions from the terpene synthase like 2 subgroup. We will also continue working on better defining the functional boundaries among members of different subgroups and families and we are particularly interested in targeting sequences that lie at the boundary of different subgroups in the hope that they may express dual activity.

By following this multidisciplinary approach the EFI seeks to give deeper insight into the biochemical, metabolic and evolutionary aspects of extant superfamilies of enzymes, so that assigning function to unknown enzymes discovered in the genome projects becomes an easier task, despite the always increasing size of sequence databases. At this early stage, the EFI has already developed quality high-throughput bioinformatics analyses and we have sufficient resources for establishing collaborations with the scientific community to aid function assignment in other functionally diverse enzyme superfamilies. Ultimately, we expect that both the basic and applied aspects of our strategy contribute to a “new era” of enzymology where genomic sequence information assist intellectual, industrial, medicinal and pharmaceutical efforts of the community at large [1].

References