Beyond Cataloging the Library of Babel

Marc Ostermeier[1,*]
[1] Department of Chemical and Biomolecular Engineering, Johns Hopkins University, 3400 N. Charles Street, Baltimore, MD 21218, USA
*Correspondence: oster@jhu.edu
DOI 10.1016/j.chembiol.2007.03.002

In this issue of *Chemistry & Biology*, Landwehr et al. [1] characterize a diverse set of chimeric cytochrome P450s acting on a diverse set of substrates. Statistical analysis of the activity profiles allows the substrates to be grouped such that one substrate can be used as a surrogate for others.

"The Library of Babel," [2] a short story by Argentine author Jorge Luis Borges, describes an infinite universe in the form of a library containing all possible 410-page books derived from a set of 25 characters, seemingly arranged in random order. The librarians that inhabit this universe seek to comprehend the library. Since the library is infinite and the number of possible books is finite, the library must contain books with all possible useful information including biographies of any person, predictions of the future, critiques of obscure economic theories, poetic descriptions of clouds, translations of any book in all languages and thousands of slight imperfections of any book (e.g., a treatise on the sky that gets everything correct except stating that the sky is red). Since the crushing majority of books in the library contain gibberish and the books appear to be randomly distributed, the library is utterly useless and the librarians are despondent.

Borges' library has been used as an analogy for protein sequence space (and its relationship with functional space) [3]. The analogy helps one grasp the vastness of sequence space and the complexity of proteins. Complex systems appear random in the absence of understanding. The relationship between protein sequence and function, on the surface, does appear to have an element of randomness in it. One mutation can completely eliminate a protein's function—sometimes for reasons that may be obscure. Like the nonsensical books in the library, a random protein sequence is almost always devoid of meaningful function, though if we look at enough random sequences a few functional ones can be found [4]. However, protein space and Borges' library have a basic difference. Whereas the contents of one book in the library give no information about the next book on the shelf, related protein sequences tend to have related function. The protein engineer's task is thus fundamentally easier than that of Borges' librarians. The relationship between sequence space and functional space is a logical one, even if we fall significantly short of fully understanding it.

Protein engineers explore this space both computationally and experimentally. In particular, directed evolution—our implementation of nature's algorithm for modifying and creating protein function—has been used successfully for navigating sequence space to arrive at proteins with properties we desire. However, directed evolution has been underutilized as a tool for gaining insight into the relationship between sequence and function, a problem for which the combinatorial nature of directed evolution is ideally suited. In addition, carefully designed and analyzed directed evolution experiments are the best route for answering the question "How does one best employ directed evolution?" For example, directed evolution has been used to show that thermostability and evolvability (the capacity to evolve) are positively correlated, suggesting that thermostable variants are better starting points for protein engineering than their less stable homologs [5].

In this issue of *Chemistry & Biology*, Landwehr et al. [1] follow up on work in an earlier paper [6] and begin to catalog efficiently the section of the protein library labeled "cytochrome P450s." The P450 superfamily catalyzes hydroxylation and demethylation reactions on a vast array of substrates, and the constituent enzymes play key roles in the break down of xenobiotics and in secondary metabolite biosynthesis. The architecture of cytochrome P450s usually includes a NADPH-P450 reductase domain and a heme-binding P450 domain. Whereas the reductase and heme-binding domains in many P450s occur as separate polypeptides, the bacterial P450s in this study incorporate both domains into a contiguous sequence. Importantly, the authors' characterization is not primarily of natural cytochrome P450s, but rather of a library of hybrid proteins designed to sparsely cover the sequence space around three bacterial P450s. The library was diverse by design and its members differ on average from the parent sequences by about 72 out of 463 amino acids. Such a library can address to what extent a high diversity of sequences in libraries results in a high diversity of functions [7].

In order that the library be as rich in functional members as possible, the authors created the 3,000 member library of new, properly-folded heme domains of P450s using SCHEMA recombination, a simple but elegant structure-based method for guiding the sites for recombination of the genes [8]. Most members (>73%) of the resulting library were catalytically active peroxygenases. In research described in this issue [1], the authors selected a set of 14 of the active, chimeric heme domains, reconstituted them with all three parent reductase domains and determined the monooxygenase and peroxygenase activities on a diverse set of eleven substrates.

Normalized activities were analyzed by a clustering statistical algorithm [9] that segregated the enzymes into four groups based on similar substrate specificities. In an interesting twist that sought relationships among substrates in chemical space, correlations between substrates were used to identify groups of substrates with similar chimera profiles. This approach addresses whether a chimera's activity on one substrate predicts activity on another with the outcome being the identification of substrates that can be surrogates for other substrates in the same group. This classification has the potential to facilitate discovery of useful catalysts by high-throughput screening using these surrogates. Not too surprisingly, a correspondence was found between the protein clusters and the substrate clusters. The analysis of the similarities and differences in physical, structural, and chemical properties within and between clusters—either protein or substrate—should be a powerful tool for gaining insight into P450 function and for predicting the enzymatic activity of untested chimeras on tested substrates and tested chimeras on untested substrates. The authors use their analysis to suggest that swapping of the reductase domains always resulted in functional enzymes because key interdomain interactions were conserved.

The results illustrate that diversity in sequence leads to diversity in function. The less expected, but more important finding is that such diversity in function need not come at the expense of absolute activity. For all eleven substrates, the top-performing chimeric enzyme's activity exceeded that of the three parental enzymes. However, the work most importantly speaks to the power of directed evolution approaches as a tool for not just exploring, but for analyzing sequence space and its relationship with functional space. Statistical analysis of such combinatorial protein engineering experiments, among other approaches, will allow us to move beyond simply cataloging protein space, but understanding it as well.

REFERENCES

1. Landwehr, M., Carbone, M., Otey, C.R., Li, Y., and Arnold, F.H. (2007). Chem. Biol. *14*, 1–2.
2. Borges, J.L. (1962). Labyrinths: Selected Stories and Other Writings, D.A. Yates and J.E. Irby, eds. (New York: New Directions), pp. 51–58.
3. Arnold, F.H. (2000). Adv. Protein Chem. *55*, ix–xi.
4. Keefe, A.D., and Szostak, J.W. (2001). Nature *410*, 715–718.
5. Bloom, J.D., Labthavikul, S.T., Otey, C.R., and Arnold, F.H. (2006). Proc. Natl. Acad. Sci. USA *103*, 5869–5874.
6. Otey, C.R., Landwehr, M., Endelman, J.B., Hiraga, K., Bloom, J.D., and Arnold, F.H. (2006). PLoS Biol. *4*, e112.
7. Ostermeier, M. (2003). Trends Biotechnol. *21*, 244–247.
8. Voigt, C.A., Martinez, C., Wang, Z.G., Mayo, S.L., and Arnold, F.H. (2002). Nat. Struct. Biol. *9*, 553–558.
9. Larsson, A.K., Emren, L.O., Bardsley, W.G., and Mannervik, B. (2004). Protein Eng. Des. Sel. *17*, 49–55.

Antifungal Tradecraft by Cholesterol Oxidase

Natasha M. Nesbitt[1] **and Nicole S. Sampson**[1,*]
[1] Department of Chemistry, Stony Brook University, Stony Brook, NY 11794, USA
*Correspondence: nicole.sampson@stonybrook.edu
DOI 10.1016/j.chembiol.2007.03.003

In this issue of *Chemistry & Biology*, Aparicio and coworkers report that secreted bacterial cholesterol oxidase is required for the biosynthesis of the antifungal polyene pimaricin by *Streptomyces natalensis* [1]. Their discovery expands the inventory of tasks this biotechnologically important enzyme performs.

Bacterial cholesterol oxidases have been utilized by humans for more than 30 years to analyze the bane of Western eating habits, serum cholesterol [2–4]. For their own use, bacteria, actinomycetes in particular, secrete cholesterol oxidase to metabolize 3β-hydroxy sterols to 3-ketosteroids, which serve as their energy source [5, 6]. Cholesterol oxidase, more properly 3β-hydroxysterol oxidase, is the first enzyme in that metabolic pathway that yields propionate and acetate as the ultimate products. Since their first isolation, cholesterol oxidases have been rediscovered in screens for insecticides [7], and have been widely used in the search for lipid rafts [8]. In the presence of a sufficient quantity of either cholesterol or ergosterol mixed with saturated lipids, liquid-phase membranes are ordered, that is, they can form lipid rafts. Oxidation of the sterol to cholest-4-en-3-one or ergosta-4,7,22-trien-3-one (Figure 1) results in dissolution of these lipid domains. These liquid-disordered membranes render the lipid bilayer more permeable and susceptible to lysis, which is the mechanism of insecticidal activity [9].

Now, Aparicio and coworkers report a new function for the *Streptomyces*

