Chemistry & Biology, Vol. 11, June, 2004, DOI 10.1016/j.chembiol.2004.06.001

Back to Basics: Assigning Biochemical Function in the Post-Genomic Era

In the March and May issues of *Chemistry & Biology*, Croteau and coworkers [16, 17] reported the molecular cloning, expression, and biochemical characterization of two new cytochrome P450 hydroxylases of the taxol biosynthesis pathway. The results reported not only mark a major advance toward the long-term goal of engineering the taxoid pathway to increase biological production yields of this important anticancer agent, but also highlight the challenges and opportunities in assigning specific biochemical function to newly discovered biosynthetic genes.

For many decades, the dominant paradigm for the discovery of new biochemical reactions was grounded in the search for enzymes that would catalyze transformations of defined substrates to identifiable products. By contrast, both classical and molecular genetics begin with the assignment of observable phenotypes to specific alleles or genes. The fields of biochemistry and genetics are united by the classic one gene-one enzyme hypothesis [1]. Numerous experimental approaches have been used to correlate specific genes with the enzymes that they encode. The groundbreaking studies of Beadle and Tatum involved a top-down elucidation of biochemical pathways, based on the identification and ordering of individual metabolites that accumulated in blocked mutants of *Neurospora crassa* [1, 2]. Since the widespread introduction of recombinant DNA methodology, enzymologists have utilized a predominantly bottom-up strategy that uses amino acid sequence information or specific antibodies obtained from purified proteins to clone the corresponding structural genes, thus opening the door to the rational manipulation of both protein sequence and structure and, in so doing, fueling the current explosive growth in structural biology.

The genomic revolution now promises to alter fundamentally the experimental paradigms of enzymology. With completed genome sequences now measuring in the hundreds, the vast majority of new protein sequences are deduced from DNA sequence data rather than determined by direct amino acid sequencing of isolated proteins. Analysis of this torrent of new sequence information has required the development of powerful new bioinformatics algorithms designed to recognize protein-coding sequences and to assign likely function based on conserved protein sequence or structural motifs. At the same time, the rapidly emerging field of structural proteomics, with its ambitious goal of the determination of the three-dimensional structures of the majority of the proteins in key microbial, plant, and mammalian proteomes, promises to unleash an additional flood of protein structural data. In spite of, or perhaps because of, these breathtaking advances, a substantial fraction of newly discovered protein sequences have no assigned biochemical function. In fact, it is today possible to determine the complete amino acid sequence and full three-dimesional structure of a new protein without knowing either the natural substrate for the enzyme nor even the biochemical reaction that it catalyzes. Although there are well-established library screening techniques for identifying the preferred macromolecular protein or nucleic acid substrates of "hypothetical proteins" annotated simply as proteases, protein kinases or phosphatases, and nucleases, there is as yet no general method for screening for the corresponding small-molecule substrates of proteins belonging to widely occurring, well-recognized enzyme classes such as kinases, hydrolases, dioxygenases, dehydrogenases, or cytochromes P450. Indeed, the more novel the biochemical reaction that is catalyzed by a newly discovered gene product, the less useful are homology-based cloning or bioinformatics approaches likely to be in identifying the individual substrates or even the nature of the chemical transformation.

An increasing number of research teams have taken up the challenge of elucidating biosynthetic pathways by starting with the discovery of pathway genes and determining the individual biochemical reactions that they encode. Several examples illustrate both the obstacles and the rewards of such studies. In a few instances, the likely biochemical substrate can be reasonably inferred from analysis of the sequence of the encoded protein. Such new biochemical reactions can be directly elucidated by incubating the putative substrate with the recombinant enzyme and determining the structure of the newly formed reaction product. Using such an approach, we recently showed that an apparent terpene cyclase gene uncovered in the genomic sequencing of *Streptomyces coelicolor* encodes a protein that converts the universal sesquiterpene precursor farnesyl diphosphate to the cyclic alcohol germacradienol, a possible precursor of the widely occurring earthy odorant geosmin [3, 4]. Similarly, Matsuda has identified a new plant triterpene and its cognate synthase by incubation of oxidosqualene with a recombinant cyclase from *Arabidopsis thaliana* that showed strong homology to cycloartenol synthase, a technique that has been dubbed "genome mining" [5]. In most other cases, however, the actual substrate of a new gene product is a good deal more difficult to identify. Although molecular genetic studies by Winkler had established that two *E. coli* gene products, PdxA and PdxJ, were responsible for formation of the characteristic pyridoxine ring of vitamin B_6, and Spenser had independently established that 1-deoxyxylulose and 4-hydroxythreonine were the key biosynthetic precursors, it nevertheless took several years of extensive screening to establish that the true substrates of recombinant PdxA and PdxJ were the corresponding monophosphate esters, 4-hydroxythreonine-4-phosphate and 1-deoxyxylulose-5-phosphate,

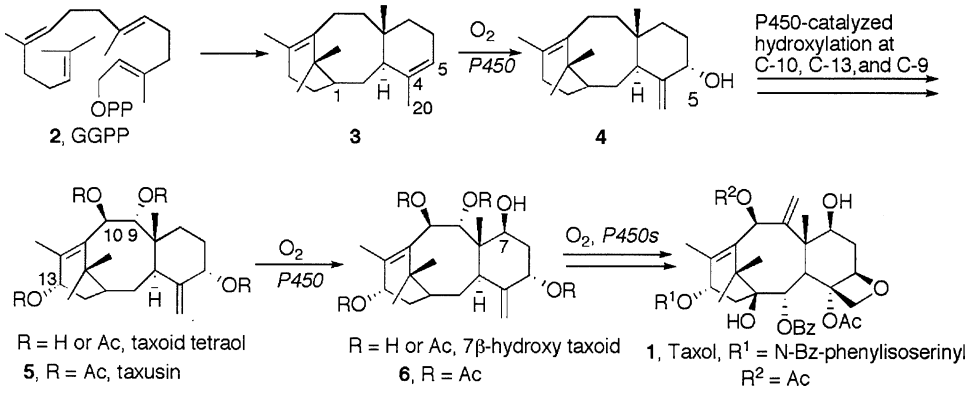

Figure 1. Taxol Biosynthetic Pathway from Geranylgeranyl Diphosphate (2, GGPP) and Taxadiene (3), Involving a Series of Cytochrome P450-Catalyzed Hydroxylations

respectively, thereby allowing the elucidation of the actual biosynthetic reactions [6, 7]. Intriguingly, pyridoxine biosynthesis in the bacterium *Bacillus subtilis* [8, 9], as well as in yeast and other eukaryotic microorganisms [10] and in higher plants, involves a completely distinct and as yet undefined set of biochemical reactions encoded by two closely linked and highly conserved genes of still incompletely understood biochemical function. Clustering of microbial biosynthetic genes can often be a powerful aid to the discovery of the complete set of biosynthetic enzymes. For example, Begley and his collaborators have carried out a remarkable set of studies on the thiamine biosynthetic gene cluster of *E. coli* that has resulted in the discovery of a wealth of novel biochemical reactions [11]. Similarly, a good deal is now known about the gene cluster that encodes all of the enzymes required for the biosynthesis of the broad spectrum antibiotic erythromycin [12, 13]. On the other hand, only a minor fraction of the dozens of known polyketide biosynthetic gene clusters have been directly characterized at the biochemical level. Even for well-established classes of enzymes within these clusters, definition of specific biochemical reactions can present a significant challenge. For example, PicK is a cytochrome P450 monooxygenase that is responsible for introduction of the C-12 hydroxyl group of the 14-membered ring antibiotic picromycin. Although recognition of the P450 reaction type was relatively straightforward, identification of the actual biosynthetic substrate proved to be a considerable experimental challenge. Eventually, it was established that while PicK has an absolute requirement for a desosamine sugar at the C-3 hydroxyl of its macrolide substrate, the enzyme not only catalyzes the hydroxylation of 14-membered ring narbomycin to picromycin, it also efficiently oxidizes the closely related 12-membered macrolide, 10-deoxymethymycin, at both C-10 and C-12 [14, 15].

Natural product biosynthetic genes are rarely clustered in plants, thereby raising the barriers to complete genetic and biochemical characterization of biosynthetic pathways by an order of magnitude and requiring the development of novel approaches to the discovery of biosynthetic genes and the assignment of their biochemical function. In the March and May issues of *Chemistry & Biology*, Rodney Croteau and his collaborators have described impressive studies that resulted in the molecular cloning, eukaryotic expression, and biochemical characterization of two new cytochrome P450 hydroxylases involved in the biosynthesis of the antimitotic anticancer agent Taxol (**1**, generic name paclitaxel) [16, 17]. This complex diterpene metabolite is produced by various species of yew (*Taxus*), and pharmaceutical supplies of the drug currently depend on semisynthetic manipulation of biologically produced, late-stage taxoid metabolites. In a wide-ranging biosynthetic study, Croteau, frequently in close collaboration with Robert M. Williams, has defined many of the biosynthetic enzymes at both the protein and the DNA level and identified many of the key biosynthetic steps in the formation of Taxol. Following cyclization of the universal acyclic diterpene precursor geranylgeranyl diphosphate (**2**) to the parent hydrocarbon taxa-4(5),11(12)-diene (**3**), a series of some eight cytochrome P450-catalyzed oxygenations, along with four CoA-dependent acylations, are required for formation of the end-product Taxol (**1**) (Figure 1). The classical biochemical approach of screening for individual hydroxylase activities in *Taxus* extracts would be complicated not only by the absence of suitable substrates and the low natural titers of the target enzymes, but by expected difficulties in chromatographically resolving the individual enzymes from one another. Instead, Croteau designed a powerful homology-based cloning strategy using a cDNA library obtained from induced *Taxus* cells, taking advantage of universally conserved P450 sequence motifs to design degenerate PCR primers. Sequencing of the resulting individual *E. coli* clones identified several open reading frames encoding P450 proteins. The presumptive P450 cDNA clones were functionally expressed in an engineered *Saccharomyces cerevisiae* stain, WAT11, harboring an inducible NADPH-cytochrome P450 reductase. The presence of the recombinant P450 was verified by the observation of the characteristic reduced P450-CO-difference spectra, and the transformed yeast cells were screened for the ability to hydroxylate a battery of exogenously administered taxoid substrates. In this manner, it was found that one P450 clone converted taxa-4(5),11(12)-diene (**3**) to the previously known taxa-4(20),11(12)-dien-5α-ol (**4**) [16]. The 5α-hydroxylase was subsequently obtained in purified form by baculovirus expression in insect cells (*Spodoptera fugiperda*), and

the recombinant enzyme was characterized biochemically. Although the 5α-hydroxylase displayed a narrow substrate specificity, interestingly it also converted the isomeric taxa-4(20),11(12)-diene to the same 5α-hydroxyl product **4** with a V_{rel}/K_m that was twice that of the presumed natural substrate **3**. Further screening for P450 enzymes catalyzing downstream oxidations was complicated by the absence of suitable natural substrates. Croteau chose instead to incubate microsomes of transformed yeast with a surrogate substrate, taxusin (**5**), a presumed analog of the natural but as yet unidentified taxoid intermediate (Figure 1) [17]. The resulting product, shown to be the corresponding 7β-hydroxyl derivative **6**, was rigorously identified by NMR spectroscopic methods. The 7β-hydoxylase itself was then expressed in insect cells and characterized biochemically. Notably, the 7β-hydoxylase displayed a relatively broad substrate specificity for polyoxygenated and acylated taxoid substrates. Thus, although the *site* of hydroxylation is beyond doubt, the natural substrate for this P450-catalyzed 7β-hydroxylation remains to be determined. Combined with other studies of taxoid-oxidizing P450s, the Croteau group continues to unmask the remaining secrets of this intricate and important biosynthetic pathway.

David E. Cane
Department of Chemistry, Box H
Brown University
Providence, Rhode Island 02912

Selected Reading

1. Beadle, G.W., and Tatum, E.L. (1941). Proc. Natl. Acad. Sci. USA *27*, 499–506.
2. Horowitz, N.H. (1991). Genetics *127*, 631–635.
3. Cane, D.E., and Watt, R.M. (2003). Proc. Natl. Acad. Sci. USA *100*, 1547–1551.
4. Gust, B., Challis, G.L., Fowler, K., Kieser, T., and Chater, K.F. (2003). Proc. Natl. Acad. Sci. USA *100*, 1541–1546.
5. Xu, R., Fazio, G.C., and Matsuda, S.P. (2004). Phytochemistry *65*, 261–291.
6. Cane, D.E., Hsiung, Y., Cornish, J.A., Robinson, J.K., and Spenser, I.D. (1998). J. Am. Chem. Soc. *120*, 1936–1937.
7. Cane, D.E., Du, S., Robinson, J.K., Hsiung, Y.-J., and Spenser, I.D. (1999). J. Am. Chem. Soc. *121*, 7722–7723.
8. Belitsky, B.R. (2004). J. Bacteriol. *186*, 1191–1196.
9. Bauer, J.A., Bennett, E.M., Begley, T.P., and Ealick, S.E. (2004). J. Biol. Chem. *279*, 2704–2711.
10. Wetzel, D.K., Ehrenshaft, M., Denslow, S.A., and Daub, M.E. (2004). FEBS Lett. *564*, 143–146.
11. Xi, J., Ge, Y., Kinsland, C., McLafferty, F.W., and Begley, T.P. (2001). Proc. Natl. Acad. Sci. USA *98*, 8513–8518.
12. Khosla, C., Gokhale, R.S., Jacobsen, J.R., and Cane, D.E. (1999). Annu. Rev. Biochem. *68*, 219–253.
13. Staunton, J., and Weissman, K.J. (2001). Nat. Prod. Rep. *18*, 380–416.
14. Graziani, E.I., Cane, D.E., Betlach, M.C., Kealey, J.T., and McDaniel, R. (1998). Bioorg. Med. Chem. Lett. *8*, 3117–3120.
15. Xue, Y., Wilson, D., Zhao, L., Liu, H.-W., and Sherman, D.H. (1998). Chem. Biol. *5*, 661–667.
16. Jennewein, S., Long, R.M., Williams, R.M., and Croteau, R. (2004). Chem. Biol. *11*, 379–387.
17. Chau, M., Jennewein, S., Walker, K., and Croteau, R. (2004). Chem. Biol. *11*, 663–672.

Chemistry & Biology, Vol. 11, June, 2004, DOI 10.1016/j.chembiol.2004.06.003

Plumbing New Depths in Drug Discovery

Marine natural products are an untapped source for drug development. In this month's *Chemistry & Biology*, Gerwick and coworkers [1] describe the biosynthetic genes for novel neurotoxins, jamaicamides A–C from *Lyngbya majuscula*, bringing closer the goal of producing marine natural products in terrestrial hosts.

Following on the failure of combinatorial chemistry to deliver the anticipated wealth of new drug candidates, interest in natural products as pharmaceutical leads has been significantly rekindled [2]. One of the least characterized but most promising sources for new drug candidates is the ocean: to survive in this environment, it seems that microorganisms must produce a tremendous diversity of toxic metabolites [3]. Not only are these compounds cytotoxic, but many exhibit important medicinal activities including anticancer, antifungal, antimalarial, antiviral, immunosuppressive, and antimicrobial properties. As an illustration of their potential utility, approximately half of all current anticancer discovery efforts are focused on marine organisms [3].

Exploiting the chemical diversity of marine metabolites, however, has been greatly hampered by their extremely low natural abundance. In addition, all but a small fraction of marine microbes are difficult to culture using conventional methods, particularly those that are symbionts of aquatic invertebrates [4]. One promising alternative is to clone the biosynthetic genes and transfer them wholesale from the natural hosts to more productive alternatives. This strategy not only offers a more secure source of these compounds, but also paves the way for the genetic engineering of unnatural variants of these molecules for use in drug discovery [5, 6]. However, research in this area is relatively new, with only three complete gene clusters sequenced to date [7–9].

From a "retrobiosynthetic" analysis, it appears that many marine metabolites are of polyketide or nonribosomal polypeptide origin, and a significant number incorporate features of both [4]. Assembly of such structures in terrestrial microorganisms occurs on gigantic multienzymes called polyketide synthases (PKSs) and nonribosomal polypeptide synthases (NRPSs), respec-