GUEST COMMENTARY

The Difficult Road from Sequence to Function

Robert H. White*

Department of Biochemistry, Virginia Polytechnic Institute and State University, Blacksburg, Virginia 24061-0308

The recent report by Patridge and Ferry (20) on the identification of a function for WrbA in the archaea emphasizes the continued importance of experimental measurements leading the way to the identification of the function of a protein encoded by a hypothetical gene. WrbA was first discovered in 1993 when Yang et al. (25) found that it copurified with the tryptophan repressor protein and thus assigned it the name tryptophan (W) repressor-binding protein (WrbA). The enzyme showed strong sequence similarity to the flavodoxin family of proteins, having the α/β core of the flavodoxin fold but with an additional α/β unit unique to the WrbA family (9). These authors concluded that members of the WrbA family of proteins were unlikely to function as DNA-binding proteins. Biochemical experiments with the purified protein showed that it contained bound flavin mononucleotide and was multimeric in solution (10). Despite this early work, the physiological role of the protein remained unclear. Thus, we had an example of a protein for which the gene sequence was known and some of the biochemical properties of which had been determined, but its function was not discernible. The breakthrough finally came from experimental work with fungi (1, 13) and plants (17), where several biochemically characterized NAD(P)H:quinone oxidoreductases with low sequence identity to the WrbA family were identified. This allowed Patridge and Ferry to make the connection between WrbAs in the archaea and this family of enzymes, a connection discounted by these earlier authors because of the low (32%) sequence identity to the archaeal enzymes.

This situation is in contrast to the large numbers of open reading frames that have only been identified *in silico* and for which we have at present no or limited biochemical data as to their function. Suggestions as to the possible functions of these gene products come only from bioinformatic studies. From genomic and protein sequence data, we have “piles of information but only flakes of knowledge” about these proteins (8). All of this clearly reveals an ever-growing problem in this postgenomic world: our inability to predict the functions of the thousands of genes that have been and are continually being identified. Not only are the predictions sometimes not correct, but many times even with a relatively firm annotation of sequence similarity to proteins of known function, surprises are found when their enzymatic reactions are studied experimentally in detail. The fundamental problem is that we cannot predict protein function from sequence and/or structure with

any degree of certainty (16). The number of folding patterns for proteins is surprisingly limited, with ~80% of the proteins utilizing 1 of the 400 structural folds identified to date (6). Specific functions evolve by duplication, recombination, and divergence of this core repertoire (5). Understanding the molecular paths that lead to the evolution of one function from another in a given superfamily is one of the next challenges of structural biology, impacting not only our understanding of how proteins evolve but also the task of correctly annotating the genes identified by whole-genome sequencing (15).

Our inability to relate a specific reaction to a specific folding pattern is the major reason for the increased interest in obtaining crystal structures of enzymes containing bound substrates, cofactors, products, and transition analogs. In addition to WrbA, there remain many established “enzymes,” such as the old yellow enzyme (24) and rhodanese (3), that have been studied for years yet whose true functions remain obscure.

Two obstacles currently stand out in the annotation of specific roles of proteins. The first is the increasing numbers of protein families that can be defined from gene sequence data but cannot be related to proteins of known function. The second is the occurrence of nonorthologous gene replacements of enzymes in known biosynthetic pathways. Problems with genome annotation (21) have also been documented. Genomics and bioinformatics provide a good system for filing data and ultimately establishing the limits of the possible metabolism that an organism can do but do not establish what the reactions or pathways are. The genomic and bioinformatic disciplines supplied the informational framework that initiated the discovery of new biosynthetic pathways to ribose-phosphate (11), aromatic amino acids (22), and cysteine (23), but only through experimental work were the new pathways actually defined.

Even when we can predict the type of reaction catalyzed by a specific gene product, we are often unable to establish the substrates used by the enzyme. One area where this is a real problem is in predicting the enzymology of carbohydrate-active enzymes (7). Bioinformatics can predict genes related to known glycosyltransferases, but prediction of the structures of the substrates is not possible. We still need biochemists to prepare the substrates to test the specificity of the recombinant glycosyltransferases (7). Where are these scientists going to come from as we see a decrease in the numbers of students doing experimental biochemistry? The reasons for this decrease are many, from a reluctance to use chemicals to an increased belief that computers can solve all the problems. This is why laboratory courses in chemistry and biochemistry are continuing to fall by the wayside and are being replaced by computer-generated experiments. Proteomics is more complex than genomics by

* Mailing address: Department of Biochemistry (0308), Virginia Polytechnic Institute and State University, Blacksburg, VA 24061. Phone: (540) 231-6605. Fax: (540) 231-9070. E-mail: rhwhite@vt.edu.

several orders of magnitude (4), and enzymology (enzymomics?) is an order of magnitude still more complex and time-consuming.

Adding to our difficulty in predicting protein function is the enormous ability of enzymes to adapt new functions and catalyze many different reactions. Enzymes have long been known to be catalytically promiscuous, and many of them can catalyze more than one chemical reaction (14, 19). They have an enormous potential to evolve, and the smallest of changes in their structures can have large catalytic consequences (18). Larger changes can completely change their substrate specificities (26). In view of these facts, it is amazing that we have seen so few examples of nonorthologous replacement of different enzymes in biosynthetic pathways. Because most metabolic enzymes have evolved to "catalytic perfection" (2), they are unlikely to be replaced by newly evolving enzymes which are far from perfect. These observations indicate that the enzymes in the metabolic pathways present in the last universal common ancestor had already evolved close to perfection. New enzymes may arise only in response to drastically different environmental needs, e.g., in the extremophiles.

The establishment of a specific enzymatic reaction by a given gene product, however, does not prove that this is the reaction catalyzed by the enzyme in the cell. It does not prove the physiological purpose of enzymatic reaction. This requires the generation of a knockout mutant and characterization of the resulting phenotype.

The discipline of biochemistry has been inverted since its inception as a result of genomics. Originally enzymes were isolated based on assay of the specific reactions they catalyzed (function). Thus, when, for example, a specific mutation in histidine biosynthesis was linked to a specific enzymatic activity, we had conclusive proof of the function of the gene encoding this enzyme. Assignments of function based on genomic data must be verified by experimental data generated by assays of the isolated purified enzyme or recombinantly produced enzymes.

Given the lack of biochemical information for most enzymes from microorganisms in general, without even considering the enormous number of uncultured bacteria (12), it is apparent that determining the functions of all annotated genes is an enormous task, requiring the work of biochemists well into the future.

REFERENCES

- Akileswaran, L., B. J. Brock, J. L. Cereghino, and M. H. Gold. 1999. 1,4-Benzoquinone reductase from *Phanerochaete chrysosporium*: cDNA cloning and regulation of expression. *Appl. Environ. Microbiol.* **65**:415–421.
- Albery, W. J., and J. R. Knowles. 1976. Evolution of enzyme function and the development of catalytic efficiency. *Biochemistry* **15**:5631–5640.
- Bordo, D., and P. Bork. 2002. The rhodanese/Cdc25 phosphatase superfamily. Sequence-structure-function relations. *EMBO Rep.* **3**:741–746.
- Brower, V. 2001. Proteomics: biology in the post-genomic era. Companies all over the world rush to lead the way in the new post-genomics race. *EMBO Rep.* **2**:558–560.
- Chothia, C., J. Gough, C. Vogel, and S. A. Teichmann. 2003. Evolution of the protein repertoire. *Science* **300**:1701–1703.
- Coulson, A. F., and J. Moulton. 2002. A unfold, mesofold, and superfold model of protein fold use. *Proteins* **46**:61–71.
- Davies, G. J., and B. Henrissat. 2002. Plant glyco-related genomics. *Biochem. Soc. Trans.* **30**:292–297.
- Eisenberg, D., E. M. Marcotte, I. Xenarios, and T. O. Yeates. 2000. Protein function in the post-genomic era. *Nature* **405**:823–826.
- Grandori, R., and J. Carey. 1994. Six new candidate members of the α/β twisted open-sheet family detected by sequence similarity to flavodoxin. *Protein Sci.* **3**:2185–2193.
- Grandori, R., P. Khalifah, J. A. Boice, R. Fairman, K. Giovanielli, and J. Carey. 1998. Biochemical characterization of WrbA, founding member of a new family of multimeric flavodoxin-like proteins. *J. Biol. Chem.* **273**:20960–20966.
- Grochowski, L. L., H. Xu, and R. H. White. 2005. Ribose-5-phosphate biosynthesis in *Methanocaldococcus jannaschii* occurs in the absence of a pentose-phosphate pathway. *J. Bacteriol.* **187**:7382–7389.
- Handelsman, J. 2004. Metagenomics: application of genomics to uncultured microorganisms. *Microbiol. Mol. Biol. Rev.* **68**:669–685.
- Jensen, K. A., Jr., Z. C. Ryan, A. Vanden Wymelenberg, D. Cullen, and K. E. Hammel. 2002. An NADH:quinone oxidoreductase active during biodegradation by the brown-rot basidiomycete *Gloeophyllum trabeum*. *Appl. Environ. Microbiol.* **68**:2699–2703.
- Jensen, R. A. 1976. Enzyme recruitment in evolution of new function. *Annu. Rev. Microbiol.* **30**:409–425.
- Kinch, L. N., and N. V. Grishin. 2002. Evolution of protein structures and functions. *Curr. Opin. Struct. Biol.* **12**:400–408.
- Kinoshita, K., and H. Nakamura. 2003. Protein informatics towards function identification. *Curr. Opin. Struct. Biol.* **13**:396–400.
- Laskowski, M. J., K. A. Dreher, M. A. Gehring, S. Abel, A. L. Gensler, and I. M. Sussex. 2002. FOR1, a novel primary auxin-response gene, encodes a flavin mononucleotide-binding quinone reductase. *Plant Physiol.* **128**:578–590.
- Mesecar, A. D., B. L. Stoddard, and D. E. Koshland, Jr. 1997. Orbital steering in the catalytic power of enzymes: small structural changes with large catalytic consequences. *Science* **277**:202–206.
- O'Brien, P. J., and D. Herschlag. 1999. Catalytic promiscuity and the evolution of new enzymatic activities. *Chem. Biol.* **6**:R91–R105.
- Patridge, E. V., and J. G. Ferry. 2006. WrbA from *Escherichia coli* and *Archaeoglobus fulgidus* is an NAD(P)H:quinone oxidoreductase. *J. Bacteriol.* **188**:3498–3506.
- Poole, F. L., II, B. A. Gerwe, R. C. Hopkins, G. J. Schut, M. V. Weinberg, F. E. Jenney, Jr., and M. W. Adams. 2005. Defining genes in the genome of the hyperthermophilic archaeon *Pyrococcus furiosus*: implications for all microbial genomes. *J. Bacteriol.* **187**:7325–7332.
- White, R. H. 2004. L-Aspartate semialdehyde and a 6-deoxy-5-ketohexose 1-phosphate are the precursors to the aromatic amino acids in *Methanocaldococcus jannaschii*. *Biochemistry* **43**:7618–7627.
- White, R. H. 2003. The biosynthesis of cysteine and homocysteine in *Methanocaldococcus jannaschii*. *Biochim. Biophys. Acta* **1624**:46–53.
- Williams, R. E., and N. C. Bruce. 2002. New uses for an old enzyme—the old yellow enzyme family of flavoenzymes. *Microbiology* **148**:1607–1614.
- Yang, W., L. Ni, and R. L. Somerville. 1993. A stationary-phase protein of *Escherichia coli* that affects the mode of association between the *trp* repressor protein and operator-bearing DNA. *Proc. Natl. Acad. Sci. USA* **90**:5796–5800.
- Yew, W. S., J. Akana, E. L. Wise, I. Rayment, and J. A. Gerlt. 2005. Evolution of enzymatic activities in the orotidine 5'-monophosphate decarboxylase superfamily: enhancing the promiscuous D-arabino-hex-3-ulose 6-phosphate synthase reaction catalyzed by 3-keto-L-gulonate 6-phosphate decarboxylase. *Biochemistry* **44**:1807–1815.

The views expressed in this Commentary do not necessarily reflect the views of the journal or of ASM.