Chemistry & Biology, Vol. 10, 1141–1150, December, 2003, ©2003 Elsevier Science Ltd. All rights reserved. DOI 10.1016/j.chembiol.2003.12.008



# How to Find "Missing" Genes

**Assigning function to "new" proteins is frequently the rate-determining step for deciphering metabolic pathways and regulatory networks. Osterman and Begley break down this barrier by demonstrating that comparative analyses of microbial genomes is a powerful strategy for identifying pathway components.**

Determining the function of proteins encoded by "new" genome sequences is a major challenge for postgenomic biology. Protein sequence databases are rapidly expanding and will continue to do so as the result of numerous genome projects. As of October 17, 2003, the TrEMBL nonredundant database contained sequences for 1,017,041 proteins. These sequences should be sufficient to describe and understand complex metabolic and regulatory networks within cells, tissues, and whole organisms *if* the molecular functions of *all* of the proteins in an organism are known. However, at least 50% of the proteins have uncertain, unknown, or incorrectly assigned function. Without this accurate information, the potential inherent in genome sequences cannot be realized. Today's challenge is therefore to develop strategies to facilitate functional assignment.

In some cases, sequence identity is sufficient to deduce function, i.e., orthologs (isofunctional homologs) usually share >35% sequence identity [1]. But, conservation of function sometimes occurs only with much greater (or sometimes even much lower) levels of sequence identity. Additionally, enzymes that catalyze the same reaction may be analogs, not homologs, i.e., the same function may be catalyzed by unrelated sequences and structures [2]. As a result, assignment of function can be very difficult. So, what other approaches can be adopted to uncover protein function?

Clues for functional identity also can be provided by (1) transcriptional analysis (gene chips), (2) identity of upstream DNA sequences that regulate transcription, (3) functions of multidomain proteins (coupled activities in a pathway sometimes are located in a multidomain protein), and (4) phenotypes of deletion/knockout mutants. However, when sequence homology is used as the primary clue for functional assignments, as it is in the annotation of newly sequenced genomes, the resulting functional assignments are often misleading at best or incorrect. This is particularly true for members of mechanistically diverse enzyme superfamilies for which neither the identity of the substrate nor the overall reaction is conserved [3].

Functional assignment of unknown proteins encoded by microbes (eubacteria and archeae) can be facilitated by information describing the gene's genomic context because enzymes in metabolic pathways are often encoded either by operons or clusters of genome-proximal genes that are transcriptionally coregulated. Such information can be used to assign potential functions to enzymes deduced by sequence homology and can be extended to decipher metabolic pathways. The author's laboratory has used this approach to both discover previously unknown metabolic pathways and to verify novel functions of enzymes in these pathways (these novel enzyme roles include conversion of the succinate to propionate in *Escherichia coli* by a novel methylmalonyl-CoA decarboxylase [4]; utilization of L-ascorbate by *E. coli* by a novel Mg(II)-dependent 3-keto-L-gulonate 6-phosphate decarboxylase that is a homolog of the metal-independent orotidine 5-monophosphate decarboxylase [5]; and catabolism of the murein peptide by *E. coli* by a novel L-Ala-D/D-Glu epimerase [6]). Although genomic context and sequence homology are usually sufficient to specify the types of reactions in a metabolic pathway, e.g., kinase, dehydrogenase, dehydratase, and aldolase in the catabolic pathway for a carbohydrate, these often are insufficient to predict the identity of the substrate for the pathway and, therefore, specify the exact reactions catalyzed by the enzymes in the pathway.

Operons in different species that encode the same metabolic pathway need not contain the same genes. Therefore, comparative analyses of genomic contexts in several organisms offers the potential to recognize and identify all of the genes that encode a metabolic pathway, even if analogous enzymes catalyze the same reaction [7]. With this information, it is possible to find a solution to the "missing gene problem" in which sequence homology is not sufficient to allow identification of the enzymes in a metabolic pathway in a newly sequenced organism.

Osterman and coworkers have advocated the use of a comparative genomics approach to reconstruct the metabolic pathways from genomic sequences. An early example of this strategy was the identification of the enzymes in the human pathway for coenzyme A biosynthesis [8]. Although *E. coli* and *Homo sapiens* both perform de novo biosynthesis of coenzyme A, the sequences of the *E. coli* proteins were insufficient to directly identify the human genes. However, by realizing that (1) the human genome encodes a bifunctional enzyme that catalyzes two steps in the pathway, whereas *E. coli* utilizes two separate enzymes; and (2) the sequence homology relating the *E. coli* and human enzymes is insufficient to assign function, but the mutual sequence identity to the *S. cerevisiae* enzyme provides the "missing link," the human genes that encode the steps in coenzyme A biosynthesis were identified.

This issue of *Chemistry & Biology* includes a paper by Osterman, Begley, and coworkers that provides a further example of the use of comparative genomics to discover the enzymatic components of a metabolic pathway, thereby making them available for detailed mechanistic analyses [12]. The pathway, the de novo synthesis of nicotinamide adenine dinucleotide ($NAD^+$), is well known (Figure 1). Quinolinate is always a precursor to the nicotinamide ring, but the precise steps by which this intermediate is synthesized are not universally conserved. In microbes, aspartate and dihydroxyacetone phosphate had been established as the precursors [9], but in eukaryotes tryptophan is the precursor

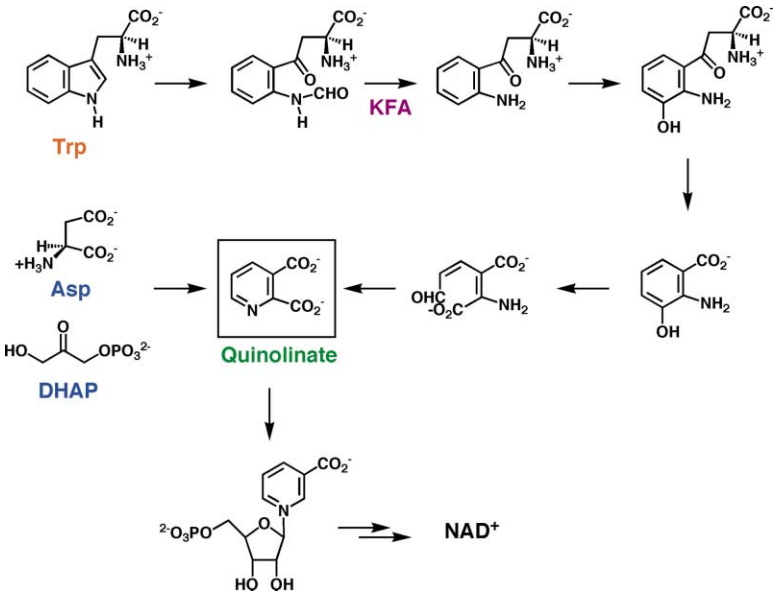


Figure 1. The Pathways for Synthesis of Quinolinate, an Intermediate in the Biosynthesis of $NAD^+$

[10]. The enzymology of the microbial pathway for the biosynthesis of quinolinate is well studied; however, the molecular details of the conversion of tryptophan to quinolinate in eukaryotes are vague, primarily due to difficulties in obtaining the purified enzymes.

Access to the complete sequences of >400 microbial genomes (public and proprietary) allowed the authors to discover that several species encode orthologs of three of the enzymes in the eukaryotic pathway [11]. As detailed in their report, the genes encoding four of the five enzymes in the quinolinate pathway were proximal in several genomes. The gene encoding the fifth enzyme, kyneurenine formamidase (KFA), could not be located in these genomes despite the availability of the sequence for the human enzyme that had been characterized as a serine protease. However, the comparative genomic analysis identified a candidate for the gene encoding an analogous microbial KFA that clustered with the quinolinate biosynthesis enzymes identified by sequence homology; surprisingly, the predicted protein contained a metal-dependent hydrolase signature sequence. Biochemical analyses of the purified protein confirmed KFA activity. Identifying the bacterial KFA as an analog of the previously characterized eukaryotic enzyme is a powerful example of how comparative genomics can be used to assign function to unknown proteins discovered in genome sequencing projects. Taken one gene at a time, as is the usual practice in annotation of genomic sequences, the bacterial protein would have been annotated as a "hypothetical" protein.

The discovery that the tryptophan-derived pathway for $NAD^+$ biosynthesis is not restricted to eukaryotes now allows for detailed structural and functional characterization of the five enzymes that convert tryptophan to quinolinate: orthologs of the eukaryotic enzymes that could not be purified are now readily available from bacterial sources. This also emphasizes that the availability of many genomic sequences could provide insights into the evolution of metabolic pathways that could not be gleaned from focused biochemical studies. The long-held notion that microbial and eukaryotic pathways for $NAD^+$ biosynthesis could be separately compartmentalized according to the strategy for quinolinate biosynthesis is now shown to be incorrect, suggesting that two microbial strategies for quinolinate biosynthesis evolved in parallel, with one of these passed to the eukaryotic progeny.

**John A. Gerlt**
Department of Biochemistry
University of Illinois
415 Roger Adams Laboratory
600 S. Mathews Avenue
Urbana, Illinois 61801

**Selected Reading**

1. Gerlt, J.A., and Babbitt, P.C. (2000). Genome Biol *1*, REVIEWS0005.
2. Galperin, M.Y., Walker, D.R., and Koonin, E.V. (1998). Genome Res. *8*, 779–790.
3. Gerlt, J.A., and Babbitt, P.C. (2001). Annu. Rev. Biochem. *70*, 209–246.
4. Haller, T., Buckel, T., Retey, J., and Gerlt, J.A. (2000). Biochemistry *39*, 4622–4629.
5. Yew, W.S., and Gerlt, J.A. (2002). J. Bacteriol. *184*, 302–306.
6. Schmidt, D.M., Hubbard, B.K., and Gerlt, J.A. (2001). Biochemistry *40*, 15707–15715.
7. Osterman, A., and Overbeek, R. (2003). Curr. Opin. Chem. Biol. *7*, 238–251.
8. Daugherty, M., Polanuyer, B., Farrell, M., Scholle, M., Lykidis, A., de Crecy-Lagard, V., and Osterman, A. (2002). J. Biol. Chem. *277*, 21431–21439.
9. Begley, T.P., Kinsland, C., Mehl, R.A., Osterman, A., and Dorrestein, P. (2001). Vitam. Horm. *61*, 103–119.
10. Magni, G., Amici, A., Emanuelli, M., Raffaelli, N., and Ruggieri, S. (1999). Adv. Enzymol. Relat. Areas Mol. Biol. *73*, 135–182.
11. Overbeek, R., Larsen, N., Walunas, T., D'Souza, M., Pusch, G., Selkov, E., Jr., Liolios, K., Joukov, V., Kaznadzey, D., Anderson, I., et al. (2003). Nucleic Acids Res. *31*, 164–171.
12. Kurnasov, O., Goral, V., Colabroy, K., Gerdes, S., Anantha, S., Osterman, A., and Begley, T.P. (2003). Chem. Biol. *10*, this issue, 1195–1204.