whether these insertions have any functional consequences. In this regard, it will be particularly interesting to define the mutations responsible for changes in the regulation or function of genes, such as *tga1* and *tb1*, and so learn what molecular magic caught the eye of ancient teosinte farmers some 7000 years ago.

References

1 Kato, T.A. (1976) *Research Bulletin Massachusetts Agricultural Experiment Station* 635
2 Doebley, J. (1990) *Econ. Bot.* 44 (Suppl. 3) 6–27
3 King, M-C. and Wilson, A.C. (1975) *Science* 188, 107–116
4 Britten, R.J. and Davidson, E.H. (1971) *Quart. Rev. Biol.* 46, 111–133
5 Hake, S. and Walbot, V. (1980) *Chromosoma* 79, 251–270
6 Freeling, M. (1984) *Annu. Rev. Plant Physiol.* 35, 277–298
7 Springer, P.S., Edwards, K.J. and Bennetzen, J.L. (1994) *Proc. Natl. Acad. Sci. U. S. A.* 91, 863–867
8 Bennetzen, J.L. *et al.* (1994) *Genome* 37, 565–576
9 SanMiguel, P. *et al.* (1996) *Science* 274, 765–768
10 Bureau, T.E. and Wessler, S.R. (1992) *Plant Cell* 4, 1283–1294
11 Bureau, T.E. and Wessler, S.R. (1994) *Proc. Natl. Acad. Sci. U. S. A.* 91, 1411–1415
12 Bureau, T.E. and Wessler, S.R. (1994) *Plant Cell* 6, 907–916
13 White, S.E., Habera, L.F. and Wessler S.R. (1994) *Proc. Natl. Acad. Sci. U. S. A.* 91, 11792–11796
14 Neuffer, M.G. (1966) *Genetics* 53, 541–549
15 Bradshaw, V.A. and McEntee, K. (1989) *Mol. Gen. Genet.* 218, 465–474
16 Pouteau, S., Grandbastien, M.A. and Boccara, M. (1994) *Plant J.* 5, 535–542
17 Mottinger, J.P., Dellaporta, S.L. and Keller, P.B. (1984) *Genetics* 106, 751–767
18 Bregliano, J.C. and Kidwell, M.G. (1983) in *Mobile Genetic Elements* (Shapiro, J.A., ed.), pp. 363–410, Academic Press
19 Laurie, D.A. and Bennett, M.D. (1985) *Heredity* 55, 307–313
20 Whitkus, R., Doebley, J. and Lee, M. (1992) *Genetics* 132, 1119–1132
21 Moore, G., Devos, K.M., Wang, Z. and Gale, M.D. (1995) *Curr. Biol.* 5, 737–739
22 Avramova, Z. *et al.* (1996) *Plant J.* 10, 1163–1168
23 Helentjaris, T., Weber, D. and Wright, S. (1988) *Genetics* 118, 353–363
24 Gaut, B.S. and Doebley, J.F. (1997) *Proc. Natl. Acad. Sci. U. S. A.* 94, 6809–6814
25 Ahn, S. and Tanksley, S.D. (1993) *Proc. Natl. Acad. Sci. U. S. A.* 90, 7980–7984
26 Doebley, J., Stec, A., Wendel, J. and Edwards, M. (1990) *Proc. Natl. Acad. Sci. U. S. A.* 87, 9888–9892
27 Doebley, J. and Stec, A. (1993) *Genetics* 134, 559–570
28 Dorweiler, J., Stec, A., Kermicle, J. and Doebley, J. (1993) *Science* 262, 233–235
29 Doebley, J., Stec, A. and Gustus, C. (1995) *Genetics* 141, 333–346
30 Dorweiler, J.E. and Doebley, J.F. (1997) *Am. J. Bot.* 84, 1313–1322
31 Burnham, C. (1959) *Maize Genet. Coop. Newslett.* 33, 74
32 Doebley, J., Stec, A. and Hubbard, L. (1997) *Nature* 386, 485–488

S. White *and* ***J. Doebley*** *are at the Department of Plant Biology, University of Minnesota, St Paul, MN 55108, USA.*

Genetwork

Using metabolic pathway databases for functional annotation

Newly obtained nucleotide and protein sequences are searched routinely against databases, and the World Wide Web has made such queries simple to perform[1]. Improved searching and scoring methods detect more subtle similarities than ever before, often allowing a researcher to make reasonable guesses about the possible role(s) of new gene sequences. Unfortunately, functional annotation of gene sequences can be fraught with difficulties, the most pernicious of which can be erroneous descriptions of database entries[2]. Therefore, the results of any database search need careful examination, and it is essential to understand the functions of the matched proteins. Metabolic pathway databases can help in providing this understanding and also offer the context for further explorations of a functional assignment. Here, we describe what you might do when you find database matches that suggest your new protein has some similarity to, say, ketol-acid reductoisomerase and you have little idea what these words even mean.

The SWISS-PROT database[3], maintained by Amos Bairoch, is the most complete general resource for information about individual proteins. SWISS-PROT annotations have descriptions of the function of a protein, its domain structure, post-translational modifications, variants, reactions catalyzed by this protein, active site residues, similarities with other sequences and more. The database entries are linked to the ENZYME database[4], which contains short descriptions of each enzyme and the reaction it catalyzes. ENZYME is the primary reference point for the Enzyme Classification (EC) numbers and, unlike SWISS-PROT, includes enzymes that have not yet been sequenced.

To put an enzyme name into a biochemical perspective, it is valuable to consider the metabolic pathways to which it contributes. Perhaps the most familiar way to do this is using the popular poster of biochemical pathways distributed by the Boehringer Mannheim[5], which is now available on the WWW[6]. This online map can be searched for both the enzyme and the metabolite names, and it links to the ENZYME database. If you still prefer the paper version, you can request it by sending an e-mail message to biochemts_us@bmc.boehringer-mannheim.com. The Kyoto Encyclopedia of Genes and Genomes (KEGG)[7] was developed especially for the Web and offers the additional ability to focus on the metabolic reactions in specific organisms. This frequently updated site presents a comprehensive set of metabolic pathway charts, both general and specific for each of the completely sequenced genomes, as well as for *Caenorhabditis elegans*, *Drosophila* and human. Before getting links to the pathways for a specific organism, it is necessary to step down through the text hierarchy. However, on the charts, the enzymes that

PII: S0168-9525(98)01523-6

have been identified in a particular organism are color-coded, so that one can easily trace the pathways that are likely to be present or absent. KEGG has links to many resources, including information about the enzymes, substrates, products and even protein structure.

KEGG is rather conservative in its functional assignments and would not indicate that an enzyme is present in a genome unless it has a highly statistically significant match with previously known enzymes. If the KEGG chart for a particular organism does not show the enzyme you are interested in, but you have reasons to believe that it must be present, you can scan the possible candidates for this role by using the WIT database[8]. WIT requires you to enter a username, but the system is easy to use. For experts, this database allows searches of unannotated proteins in each of the completely sequenced genomes, to help find the ones most likely to have the apparently missing function. Another way to consider pathways and look for missing functions is to use the COG database[9], which organized proteins into clusters of orthologs. There is a searching tool called Cognitor, and use of COG can illuminate pathways and their evolution (Fig. 1)

Escherichia coli is certainly the most completely studied organism, so it is unsurprising that it has the most complete metabolic pathway database. The EcoCyc database is most convenient to use when downloaded, but can also be accessed on the Web[10]. It is necessary to formally register, but the database is free to academics. This database is essentially specific to *E. coli*, but contains extensive and detailed information about its known metabolic pathways and the reactions they embody.

The molecules in a complete organism are linked to different molecules for different reasons, and databases are beginning to explore these connections. In addition to metabolic pathway databases, there are also databases of cell-cycle regulation, developmental regulation, gene regulation and interaction, cell-signaling and general protein interactions. Because much of the work describing these pathways is still just beginning, the databases are in considerable flux and new ones are emerging all the time. There are also databases of small molecules and the ways in which they interact with proteins as signals, substrates, and co-factors. A page at KEGG contains links to many of these databases[11].

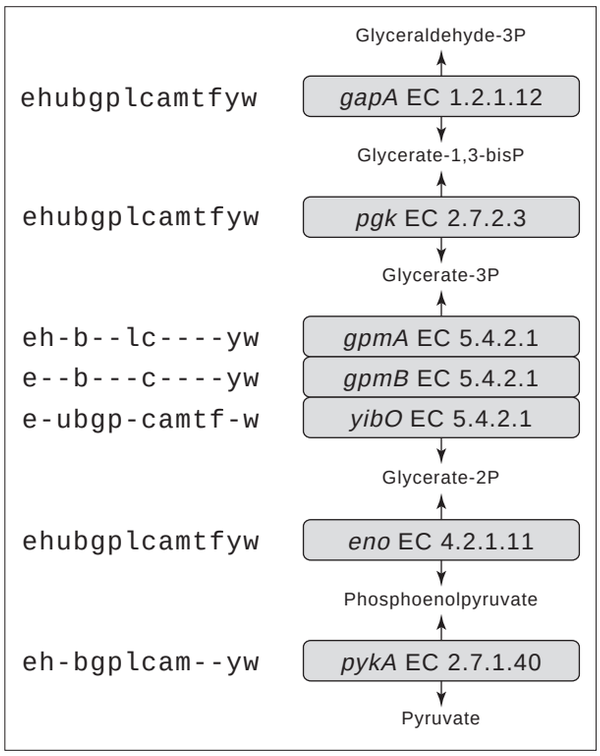

FIGURE 1. Genes and enzymes of the tri-carbon part of glycolysis. The compound names and EC numbers are from KEGG database, gene names and phylogenetic patterns are from COG. The organism symbols are: e, *E. coli*; h, *H. influenzae*; u, *H. pylori*; b, *B. subtilis*; g, *M. genitalium*; p, *M. pneumoniae*; l, *B. burgdorferi*; c, *Synechocystis sp.*; a, *A. aeolicus*; m, *M. jannschii*; t, *M. thermoautotrophicum*; f, *A. fulgidus*; y, *S. cerevisiae*; w, *C. elegans*. Most reactions of this trunk pathway are clearly catalyzed by orthologous enzymes in all organisms. There are two different types of phosphoglycerate mutase, though, and while free-living organisms (*E. coli*, *B. subtilis*, *Synechocystis sp.* and *C. elegans*) have both enzyme types (one of them in two parologous forms), the parasitic ones have either one form or the other.

Returning to ketol-acid reductoisomerase, what can be learned? First, we find it in SWISS-PROT as ILVC_ECOLI. The annotation states that it catalyzes the reaction:

(*R*)-2,3-dihydroxy-3-methylbutanoate + $NADP^+ \leftrightarrow$ (*S*)-2-hydroxy-2-methyl-3-oxobutanoate + NADPH

This is good to know, but for many of us, it is more valuable to read the description indicating that this enzyme is the second step in valine and isoleucine biosynthesis. The EC number (1.1.1.86) for this entry links to the ENZYME database. From here, there are a variety of options; one click goes to the relevant section of the Boehringer Mannheim map where the pathway and reactants are more readily understood. Details about this pathway in specific species can be obtained from the links to EcoCyc, WIT and KEGG. Thus, by using pathway databases, it is easy to learn not only what specifically a protein does, but also in what biological context it acts. This crucial knowledge can allow better evaluation of a protein's function and suggest directions for further exploration.

References

1 Brenner, S.E. (1995) *Trends Genet.* 11, 330–331
2 Doerks, T., Bairoch, A. and Bork, P. (1998) *Trends Genet.* 14, 248–250
3 Bairoch, A. and Apweiler, R. (1998) *Nucleic Acids Res.* 26, 38–42; http://www.expasy.ch/sprot/sprot-top.html
4 http://www.expasy.ch/sprot/ enzyme.html
5 http://biochem.boehringer-mannheim.com
6 http://www.expasy.ch/cgi-bin/search-biochem-index
7 Kanehisa, M. (1997) *Trends Genet.* 13, 375–376; http://www.genome.ad.jp/kegg/kegg2.html
8 Selkov, E., Jr, Grechkin, Y., Mikhailova, N. and Selkov, E. (1998) *Nucleic Acids Res.* 26, 43–45; http://wit.mcs.anl.gov/WIT2/ wit.html
9 Tatusov, R.L., Koonin, E.V. and Lipman, D.J. (1997) *Science* 278, 631–637; http://www. ncbi.nlm.nih.gov/COG
10 Karp, P.D. *et al.* (1998) *Nucleic Acids Res.* 26, 50–53; http://ecocyc.panbio.com/ecocyc/
11 http://www.genome.ad.jp/keg/kegg4.html

Michael Y. Galperin
galperin@ncbi.nlm.nih.gov

National Center for Biotechnology Information, 8600 Rockville Pike, Bethesda, MD 20894, USA.

Steven E. Brenner
brenner@hyper.stanford.edu

Department of Structural Biology, Stanford University, Fairchild Building, Stanford, CA 94305–5400, USA.