Call for an Enzyme Genomics Initiative

Peter D. Karp
SRI International
333 Ravenswood Ave, Menlo Park, CA 94025
Tel (650) 859-4358, Fax (650) 859-3735
pkarp@ai.sri.com

Abstract

I propose an Enzyme Genomics Initiative whose goal is to obtain at least one protein sequence for all enzymes that have previously been characterized biochemically. For 36% of those enzyme activities for which EC numbers have been assigned, no sequence can be found in public protein-sequence databases.

1 Introduction

A recent essay by Roberts [1] called for an effort by the scientific community to experimentally determine functions for unidentified genes in microbial genomes. Put another way, the essay focused on sequences with no associated function. This essay explores the inverse problem: functions with no associated sequence. I propose an Enzyme Genomics project whose goal is to elucidate at least one amino-acid sequence for every biochemically characterized enzyme activity for which no known sequence exists.

Roberts identifies three classes of genes whose functions would be most valuable to obtain: hypothetical genes with homologs in multiple organisms (conserved hypotheticals), unconserved hypothetical genes, and misannotated genes. Roberts proposes that a consortium of bioinformaticians post functional predictions for these genes to a central web site. Biologists would choose candidates and test the predicted functions in the lab, with results — both positive and negative – added to the same web site. Roberts also proposes that the initial list of target genes be chosen from an experimentally tractable organism such as *E. coli*, with the recognition that some experiments might be performed on homologs from other organisms.

The proposed Enzyme Genomics Initiative is based on a different gap between genomics and biochemical function, and I suggest it as another priority area both because of the many applications of metabolic enzymes in areas ranging from metabolic engineering to anti-microbial drug discovery to metabolic diseases, and because it may be easier to pursue because in many cases significant biochemical knowledge about these enzymes (such as purification procedures) is already in hand.

Consider two implications of the many characterized enzymes for which no sequence exists. We cannot identify in a newly sequenced genome any of the enzyme activities for which no sequence exists, because to identify these enzyme functions in a new genome, we require at least one sequence in a public sequence database to match against in the newly sequence genome. This consideration limits both the completeness of genome annotations, and our ability to infer the metabolic pathway complement of an organism from its genome using methods such as the PathoLogic program [2]. A second implication is that we cannot genetically engineer any of these enzymes into a new organism to accomplish a metabolic engineering goal, because we do not know which gene(s) to insert to provide the needed enzyme activity.

2 No Sequence Has Been Determined for Many Known Enzymes

Consider the enzyme D-mannitol oxidase, which was isolated from the snail digestive gland and assigned the EC number 1.1.3.40. Although the activity of this enzyme was characterized biochemically and published in 1986 [3], no amino-acid or nucleotide sequences exist for this enzyme or its gene in the public sequence databases.

As shown by the following analysis, for 36% of the enzyme activities that have been characterized biochemically, no corresponding amino-acid sequence is known. Consider the Enzyme Nomenclature System of the International Union of Biochemistry and Molecular Biology (commonly called the EC system), which is a catalog of many (but not all) biochemically characterized enzyme activities. For what fraction of those enzyme activities is at least one sequence known in a public protein sequence database? Unless otherwise stated, all the following statistics refer to database versions available as of December 2003, and were calculated with the help of SRI's BioWarehouse system for integration of bioinformatics databases.

The ENZYME database is an electronic version of the EC system. Version 33.0 of ENZYME contains 4208 distinct EC numbers, of which 472 have been deleted or transferred to new numbers. The EC system therefore lists 3736 different biochemically characterized enzyme activities.

We wrote programs to query BioWarehouse to determine how many of those EC numbers are referenced in different protein sequence databases, as a way of determining for how many of those enzymes at least one sequence is known.

- The Swiss-Prot database (version 42.6) references 1899 distinct EC numbers.

- The TrEMBL database (version 25.4) references 309 EC numbers beyond those referenced in Swiss-Prot.

- The PIR database (PIR-PSD version 78.03) references 104 EC numbers beyond those referenced in Swiss-Prot and TrEMBL (which is curious since version 42.6 of Swiss-Prot is the first UniProt release, which integrates Swiss-Prot and PIR).
- The CMR (Comprehensive Microbial Resource, version April-2003) database references an additional 24 EC numbers beyond those referenced in Swiss-Prot, TrEMBL, and PIR.
- The BioCyc (version 7.6) database collection references an additional 44 EC numbers beyond those referenced in Swiss-Prot, TrEMBL, PIR, and CMR.

In total, these databases reference 2380 distinct EC numbers, or 64% of all known EC numbers.

Therefore, for 1356 (= 3736 − 2380 =) EC numbers (36%), no protein sequence for that enzyme activity is known.

Two qualifications to the preceding analysis should be stated. First, the EC system is incomplete in that it does not yet include a number of enzymes whose biochemical activities have been characterized. The MetaCyc database [4] alone describes 890 enzyme activities that have no associated EC number. The true number of biochemically characterized enzymes is probably 5000 to 6000, thus the preceding analysis based on EC numbers is a lower bound on the number of unsequenced enzymes. The initiative should include all enzymes, whether they have been assigned EC numbers or not.

Second, there might be incompletely annotated entries in PIR and Swiss-Prot that have not been assigned EC numbers, which, if fully annotated, would provide sequences for some of these enzymes. However, when we searched the protein names and synonyms for 1.1 million proteins in UniProt that lack EC numbers against the enzyme name synonyms stored in MetaCyc, we found less than 110 sequences for any EC numbers that previously lacked them.

3 Enzyme Genomics: Sequence an Enzyme for Each Enzyme Activity

I propose a project to systematically isolate and sequence at least one enzyme for each enzyme activity that lacks any known sequence. The knowledge gained from each newly sequenced enzyme will immediately ricochet across previously sequenced genomes, as sequence similarity is used to identify its homologs in those genomes.

This project should be easier than that proposed by Roberts to choose a sequenced gene, and attempt to assign a function to it, because biochemical assays already exist for the enzyme functions in question, and purification procedures for many of these proteins have already been published.

As in Roberts' proposal, our project calls for close collaboration between bioinformaticians and wet-lab biologists. We expect that in some cases, the genes encoding these enzymes have already been sequenced by genome projects, but we simply do not know which sequences correspond to the enzyme functions we seek. Bioinformatics analyses can suggest which sequenced gene corresponds to a given enzyme function. For example, 124 of the unsequenced enzymes participate in a known metabolic pathway defined in MetaCyc. Computational techniques exist that will postulate other genes whose products act within the same pathway as a set of input genes; those techniques could be used to generate candidates for wet-lab investigation [5, 6, 7].

I envision that a number of possible experimental strategies will be used concurrently to pursue this project, and I hope that high-throughput strategies will be devised. One example strategy to approach this task would be as follows. Consider an enzyme activity E that was reported in the biochemical literature twenty years ago. Imagine that the enzyme was isolated from an organism whose genome has been completely sequenced, such as *Saccharomyces cerevisiae*. Imagine further that the paper reported a molecular weight for the protein as a whole, and molecular weights for three trypsin-cleaved fragments of the protein. An investigator searching for this enzyme activity would search the *S. cerevisiae* genome computationally for all proteins of that molecular weight, and that contained three trypsin cleavage sites that would yield fragments of approximately the observed sizes. All such proteins would be cloned, over-expressed, and assayed for the enzyme activity E.

I support many of the procedures proposed by Roberts, which should be equally applicable to the Enzyme Genomics project, such as low-overhead proposals for wet-lab funding, prioritization of targets, and project status tracking through a central database and Web site. For that matter, the same bioinformatics consortium should be able to provide analysis services and coordination for both projects.

References

[1] R. Roberts. Identifying protein function - a call for community action. *PLoS Biol.*, 2(3):E42, 2004.

[2] P.D. Karp, S. Paley, and P. Romero. The Pathway Tools Software. *Bioinformatics*, 18:S225–S232, 2002.

[3] J.E. Vorhaben, J.F. Scott, D.D. Smith, and J.W. Campbell. Mannitol oxidase: partial purification and characterisation of the membrane-bound enzyme from the snail *helix aspersa*. *Int. J. Biochem.*, 18:337–44, 1986.

[4] Cynthia J. Krieger, Peifen Zhang, Lukas A. Mueller, Alfred Wang, Suzanne Paley, Martha Arnaud, John Pick, Seung Y. Rhee, and Peter D. Karp. MetaCyc: a multiorganism database of metabolic pathways and enzymes. *Nuc. Acids Res.*, 32:D438–432, 2004.

[5] MY Galperin and EV Koonin. Who's your neighbor? new computational approaches for functional genomics. *Nature Biotechnology*, 18:609–613, 2000.

[6] I. Yanai, J.C. Mellor, and C. DeLisi. Identifying functional links between genes using conserved chromosomal proximity. *Trends Genet.*, 18:176–9, 2002.

[7] Y. Zheng, R.J. Roberts, and S. Kasif. Genomic functional annotation using co-evolution profiles of gene clusters. *Genome Biology*, 3(11):1–9, 2002.