Open Access

Open letter

Call for an enzyme genomics initiative

Peter D Karp

Address: Bioinformatics Research Group, SRI International, 333 Ravenswood Ave, Menlo Park, CA 94025, USA. E-mail: pkarp@ai.sri.com

Published: 30 July 2004

Genome ***Biology*** 2004, **5:**401

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2004/5/8/401

I propose an Enzyme Genomics Initiative, the goal of which is to obtain at least one protein sequence for each enzyme that has previously been characterized biochemically. There are 1,437 enzyme activities for which Enzyme Commission (EC) numbers have been assigned but no sequence can be found in public protein-sequence databases.

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. Here, I explore the inverse problem: functions with no associated sequence. I propose an Enzyme Genomics project whose goal is to find at least one amino-acid sequence for every biochemically characterized enzyme activity for which there is currently no known sequence.

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

My proposal for an Enzyme Genomics Initiative is based on a different part of the gap between genomics and biochemical function, and I suggest it as a fourth priority area in addition to the three suggested by Roberts. Elucidation of protein sequences corresponding to enzyme activities is important because of the many applications of metabolic enzymes in areas ranging from metabolic engineering to antimicrobial drug discovery to metabolic diseases. Finding enzyme sequences may also be easier than the projects listed by Roberts, because in many cases significant biochemical knowledge about these enzymes (such as purification procedures and assays) 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 sequenced 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.

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 are available for this enzyme in the public sequence databases.

As shown by the following analysis, for 38% 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 of 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 [4]. Version 33.0 of ENZYME contains 4,208 distinct EC numbers, of which 472 have been deleted or transferred to new numbers; it therefore lists 3,736 different biochemically characterized enzyme activities. I wrote programs to query BioWarehouse in such a way as 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 results are as follows.

The SWISS-PROT database (version 42.6) [5,6] references 1,899 distinct EC numbers. The TrEMBL database (version 25.4) [6] references 239 EC numbers beyond those referenced in SWISS-PROT. The PIR database (PIR-PSD version 78.03) [7] references 100 EC numbers beyond those referenced in SWISS-PROT and TrEMBL (which is curious, given that 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 [8] references an additional 19 EC numbers beyond those referenced in SWISS-PROT, TrEMBL, and PIR. The BioCyc (version 7.6) database collection [9] references an additional 42 EC numbers beyond those referenced in SWISS-PROT, TrEMBL, PIR, and CMR. In total, therefore, these databases reference 2,299 distinct EC numbers, or 62% of all known EC numbers. And, for 1,437 (3,736 - 2,299) EC numbers (38% of the 3,736 total), no protein sequence for that enzyme activity is known. A list of these 1,437 EC numbers is included as an additional data file with the complete version of this article, online.

There are two qualifications to the preceding analysis. 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 [10,11] alone describes 890 enzyme activities that have no associated EC number. The true number of biochemically characterized enzymes is therefore probably 5,000 to 6,000, and the preceding analysis based on EC numbers is a lower bound on the number of unsequenced enzymes. The proposed initiative should include all enzymes, whether they have been assigned EC numbers or not. Second, there might be incompletely annotated entries in PIR [7] and SWISS-PROT [5,6] that have not been assigned EC numbers, but which, if fully annotated, would provide sequences for some of these enzymes. When I 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 [10,11], I found fewer than 110 sequences for any EC number that previously lacked a sequence.

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 multiple genomes. This project should be considerably easier than the one proposed by Roberts, who advocates choosing a sequenced gene and attempting 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, my project calls for close collaboration between bioinformaticians and wet-lab biologists. One can expect that, in some cases, the genes encoding the relevant enzymes have already been sequenced by genome projects, but we simply do not know which sequences correspond to the enzyme functions we seek. Bioinformatic analyses can suggest which sequenced gene corresponds to a given enzyme function. For example, 124 of the unsequenced enzymes identified here participate in known metabolic pathways defined in MetaCyc [10,11]. Computational techniques are available that will postulate other genes whose products act within the same pathway as a set of input genes; these techniques could be used to generate candidates for wet-lab investigation [12-14].

I envisage 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 possible strategy to approach this task would be as follows. Consider an enzyme activity E that was reported in the biochemical literature 20 years ago. Imagine that the enzyme was isolated from an organism whose genome has now been completely sequenced, such as *Saccharomyces cerevisiae*. Imagine further that the 20-year-old 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 for those 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 website. For that matter, the same bioinformatics consortium should be able to provide analysis services and coordination for both projects. Future developments in this project will be available at [15].

Additional data file

A table (Additional data file 1) listing EC numbers for which no sequence was found in SWISS-PROT, TrEMBL, PIR, CMR, or BioCyc as of December 2003 is provided with the online version of this article.

Acknowledgements

This work was partly supported by grant GM70065 from the NIH National Institute for General Medical Sciences.

Richard J Roberts responds:

Peter Karp proposes a project that would greatly aid the annotation of sequenced genomes. It is both complementary to and would be synergistic with the project I proposed to assign function to unidentified genes in microbial genomes [1]. I support it heartily. One interesting question that arises is how many different ways are there to provide any given biological function? For instance, if we can identify a gene encoding a particular enzyme activity, will that automatically lead us to all of the homologs or merely to one of many families of homologs? Just how diverse is protein space?

At New England Biolabs we have already embarked on a project of this sort. There are more than 240 different discrete recognition sequences for restriction endonucleases. We now have sequences for enzymes able to recognize more than two thirds of these specificities. In many cases we have sequences for more than one example of each recognition sequence. For restriction enzymes that recognize GATC, we find that there are at least four different families of protein sequences that can recognize and cleave this sequence. Because we do not currently have three dimensional structures for any of these GATC enzymes, our estimate of the number of families is based strictly on sequence similarity – or rather the lack thereof. We cannot at this stage exclude the possibility that the families are all very similar structurally, but even that would not help unless we become much more proficient at the *de novo* prediction of protein structures from sequence.

Thus, we face the distinct possibility that for the 1,437 enzyme activities noted by Karp, for which no gene sequence is available, there might be four or more times that number of distinct gene families encoding enzymes with those activities. This combined with the large numbers of enzyme activities that are not presently represented by EC numbers means that the task ahead is daunting. As always biology is wonderfully complex and poses great challenges to both the bioinformaticians and the biochemists. But here at least is an area where small science carried out in parallel in many experimental and computational laboratories will lead to big results - and the costs could be remarkably modest!

Richard J Roberts

New England Biolabs, 32 Tozer Road, Beverly, MA 01915, USA. E-mail: roberts@neb.com

References

1. Roberts RJ: **Identifying protein function - a call for community action.** *PLoS Biol* 2004, **2:**E42. [http://www.plosbiology.org/plosonline/?request=get-document&doi=10.1371%2Fjournal.pbio.0020042]
2. Karp PD, Paley S, Romero P: **The pathway tools software.** *Bioinformatics* 2002, **18:**S225-S232.
3. Vorhaben JE, Smith DD, Campbell JW: **Mannitol oxidase: partial purification and characterisation of the membrane-bound enzyme from the snail *Helix aspersa*.** *Int J Biochem* 1986, **18:**337-344.
4. **ENZYME - Enzyme nomenclature database** [http://www.expasy.org/enzyme/]
5. Boeckmann B, Bairoch A, Apweiler R, Blatter M, Estreicher A, Gasteiger E, Martin MJ, Michoud K, O'Donovan C, Phan I, *et al.*: **The Swiss-Prot protein knowledgebase and its supplement TrEMBL in 2003.** *Nucleic Acids Res* 2003, **31:**365-370.
6. **SWISS-PROT/TrEMBL** [http://www.expasy.org/sprot/]
7. **PIR-International Protein Sequence Database** [http://pir.georgetown.edu/pirwww/dbinfo/pirpsd.html]
8. **Comprehensive Microbial Resource (CMR)** [http://www.tigr.org/tigr-scripts/CMR2/CMRHomePage.spl]
9. **BioCyc Database Collection** [http://biocyc.org/]
10. Krieger CJ, Zhang P, Mueller LA, Wang A, Paley S, Arnaud M, Pick J, Rhee SY, Karp PD: **MetaCyc: a multiorganism database of metabolic pathways and enzymes.** *Nucleic Acids Res* 2004, **32 Database issue:**D438-D432.
11. **MetaCyc** [http://metacyc.org/]
12. Galperin MY, Koonin EV: **Who's your neighbor? New computational approaches for functional genomics.** *Nat Biotechnol* 2000, **18:**609-613.
13. Yanai I, Mellor JC, DeLisi C: **Identifying functional links between genes using conserved chromosomal proximity.** *Trends Genet* 2002, **18:**176-179.
14. Zheng Y, Roberts RJ, Kasif S: **Genomic functional annotation using co-evolution profiles of gene clusters.** *Genome Biol* 2002, **3:**research0060.1-0060.9.
15. **Index of enzyme genomics** [http://bioinformatics.ai.sri.com/enzyme-genomics/]