

Distribution of orphan metabolic activities

Lifeng Chen and Dennis Vitkup

Center for Computational Biology and Bioinformatics and Department of Biomedical Informatics, Columbia University, 1130 Nicholas Ave., Irving Cancer Research Center, New York, NY 10032, USA

A significant fraction (30–40%) of known metabolic activities is currently orphan. Although orphan activities have been biochemically characterized, we do not know a single gene responsible for these reactions in any organism. The problem of orphan activities represents one of the major challenges of modern biochemistry. We analyze the distribution of orphans across biochemical space, through years of enzymatic characterization, and by biological organisms. We find that orphan metabolic activities have been accumulating for many decades. They are widely distributed across enzymatic functional space and metabolic network neighborhoods. Although orphans are relatively more abundant in less studied species, over half of orphan reactions have been experimentally characterized in more than one organism. Shrinking the space of orphan activities will likely require a close collaboration between computational and experimental laboratories.

Introduction

Rapid technological progress, spurred by genomic technologies, has revolutionized biological research [1,2]. The race to obtain complete functional datasets for various model organisms is under way [3]. The best known example is the exponential growth of complete genomic sequences (<http://www.genomesonline.org/>). There are multiple ongoing efforts to accumulate other 'omics datasets: transcriptome collections [4,5], collections of regulatory elements [6,7], cellular proteomics [8,9] and metabolomics data [10]. These resources will ultimately allow us to place genes and functions in context with their appropriate cellular networks [11–13].

The collection of all biochemical reactions catalyzed in nature represents one of the oldest and most fundamental pieces of biological knowledge. Over the last century, biochemical studies have identified several thousand enzymatic reactions catalyzed in various species. The International Congress of Biochemistry set up the Enzyme Commission (<http://www.chem.qmul.ac.uk/iubmb/enzyme/>) in 1955 to classify biochemical reactions. More than 4,000 known enzymatic activities are currently assigned unique identification numbers (EC numbers) by the Commission. Usually, only enzymes characterized in a published manuscript are assigned EC numbers. The EC numbers are organized hierarchically using a four-digit code: the first

digit indicates the general class of the catalyzed reactions (oxidoreductases, transferases, hydrolases, lyases, isomerases or ligases), the second usually specifies the type of the chemical bond the reaction acts on (for example, peptide bonds), the third specifies the general type of the reaction substrates (for example, polypeptides), and the fourth is a serial number that generally indicates the substrate specificity.

Paradoxically, as pointed out by Karp [14] and other researchers [15–19], for a significant fraction of the characterized metabolic activities no representative genes have been identified across all organisms. Such activities have been called 'orphan' to emphasize that they are not currently assigned to any gene. It is necessary to differentiate between global and local orphan metabolic activities. Global orphans represent activities without a single representative gene in any organism. By contrast, the term local orphan refers to a known activity without an assigned gene in a particular organism, while genes for the activity are known in other species. Global orphans are the main topic of this article and here we refer to them simply as orphans.

By analyzing available sequence and enzyme databases we have identified a total of 1360 orphan metabolic activities*. This constitutes ~34% of the 4003 active EC numbers in the ENZYME database (an electronic version of the EC classification, release of 12 December 2006, <http://ca.expasy.org/enzyme/>) [20]. No representative sequences for the orphan activities were found in the current releases of SWISSPROT (version 51.3), TrEMBL (version 34.3) [21], PIR (version 80.00) [22], or MetaCyc (version 10.5) [23]. Many of the identified orphan activities were characterized long ago. For example, Sacks *et al.* [24] first reported in 1951 that an enzyme mixture from corn kernels catalyzed the conversion of maleate to D-malate. The responsible enzyme – maleate hydratase (EC 4.2.1.31) – was then purified from rabbit kidneys [25] and from *Pseudomonas pseudoalcaligenes* [26]. In 1992, van der Werf *et al.* [27] screened >300 microbes for production of D-malate from maleate and found that 128 of the tested bacteria are capable of the conversion. Despite the wide distribution among organisms including bacteria, plants and mammals, the activity remains orphan. Another example is the nicotinamide-nucleotide (NMN) amidase (EC 3.5.1.42), which is involved in the pyridine nucleotide cycle and converts NMN

Corresponding author: Vitkup, D. (dv2121@columbia.edu). Available online 18 June 2007.

* Supplementary materials are available at http://vitkuplab.cu-genome.org/html/global_orphan/global_orphan_ec_2007.

to nicotinic acid mononucleotide (NaMN). This activity was first described in *Azotobacter vinelandii* [28]. The biochemical evidence for this enzyme was later found in *Salmonella typhimurium* [29], and in 1981, the enzyme was purified from *Escherichia coli* [30]. Nevertheless, currently there is no known gene for the activity in metabolic databases.

The two examples above illustrate the presence of known orphan activities in sequenced organisms. The genes responsible for the orphan activities are among the sequenced open reading frames. The problem is that the activities and the responsible genes are not matched to each other: the genes are likely to be marked as hypotheticals and the activities as orphan.

Historical accumulation of orphan activities

To understand the historical accumulation of all and orphan activities we calculated the number of enzymes characterized every year using the earliest publication reference for each biochemical activity on the Enzyme Commission website (<http://www.chem.qmul.ac.uk/iubmb/enzyme/>) (Figure 1a). Only activities reported before the year 2000 were included in the analysis as a significant number of recently characterized activities have yet to be classified. Interestingly, the discovery rate of new enzymes decreased in the 1970s and 1990s. These decreases might be related to shifting trends in funding and research directions, such as genomics in the 1990s. Alternatively, the decreases might be related to changes in the criteria used to define new activities.

For comparison, we also calculated the fractions of all activities characterized in different years which still remain orphan (Figure 1b). Generally, activities characterized in earlier years are less likely to remain orphan. In the 1990s, the fraction of newly characterized orphan activities decreased as sequencing technologies improved

and it became difficult to publish a manuscript without reporting the sequenced gene for a new activity.

Distribution of orphan activities across the biochemical space

What is the distribution of orphan activities across the enzymatic functional space? One possibility is that orphan activities are primarily clustered in less studied regions of the EC classification space. However, our results demonstrate that this is unlikely to be the case (Figure 2). Figure 2a presents the fractions of orphan activities at different levels of the EC classification organized in a tree hierarchy. For example, at the top level of the EC hierarchy, we find that 403 (36.5%) of oxidoreductases, 436 (37.6%) of transferases, 326 (29.8%) of hydrolases, 116 (33.0%) of lyases, 51 (32.0%) of isomerases, and 28 (20.5%) of ligases are orphan. Similar patterns are also observed at other levels of the EC hierarchy. The number of orphan activities in a given enzymatic class is proportional (Pearson's $R = 0.93$, $P < 0.0001$) to the total number of metabolic activities in the class (Figure 2b), whereas the fraction of orphan activities is not correlated ($R = 0.098$, $P = 0.4$) with the class size (Figure 2b, inset). Consequently, metabolic activities in large metabolic classes are, on average, as likely to be orphan as activities in smaller classes.

In many cases, enzymes neighboring each other in metabolic networks catalyze different types of biochemical reactions (i.e. they are not similar in terms of the EC classification). Consequently, the distribution of orphans in the context of metabolic networks is different from their distribution across the EC space. We asked whether activities neighboring orphans in metabolic networks are more likely to be orphans themselves. Following the previously described approach [31,32], we defined network neighbors as activities (reactions) that share the metabolites either as substrates or products. All metabolic reactions

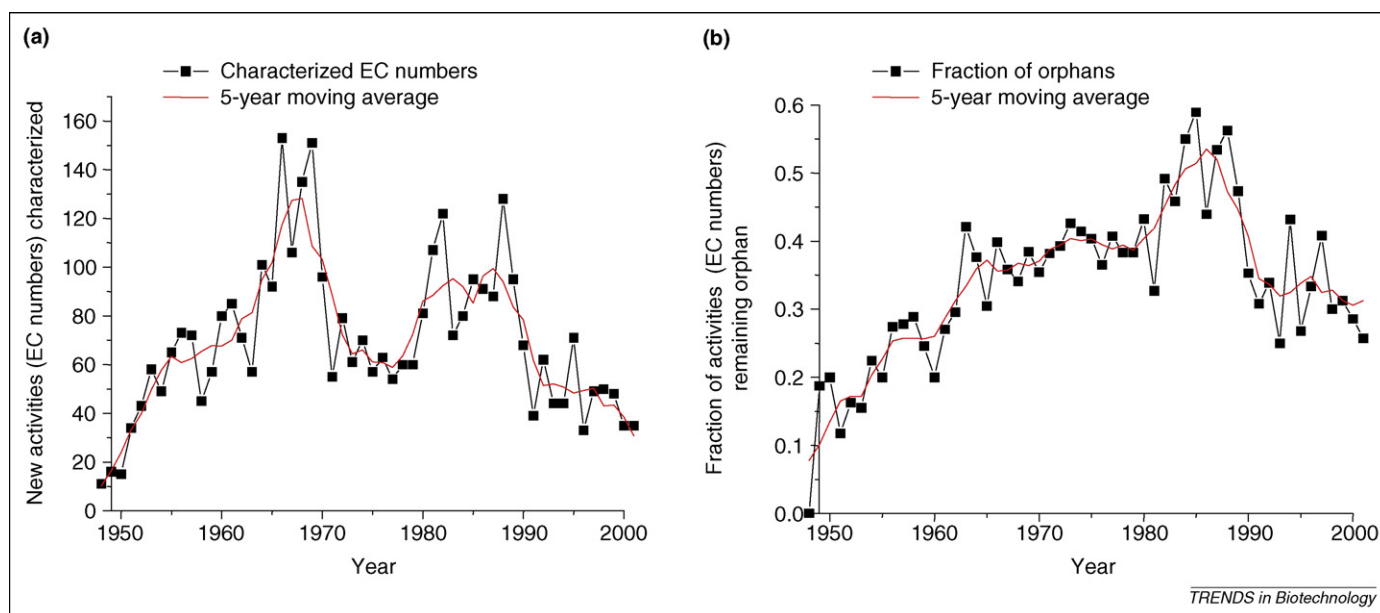


Figure 1. Experimental characterization of biochemical activities over the years. Only data before the year 2000 are shown owing to the lag in classification of recently characterized EC numbers. (a) The total number of unique biochemical activities (EC numbers) characterized each year (1950–2000). (b) The fractions of activities characterized in different years which still remain orphan. Activities characterized in earlier years are more likely to be assigned genes. The fractions of orphans decreased after the mid-1980s (probably owing to developments in sequencing technologies).

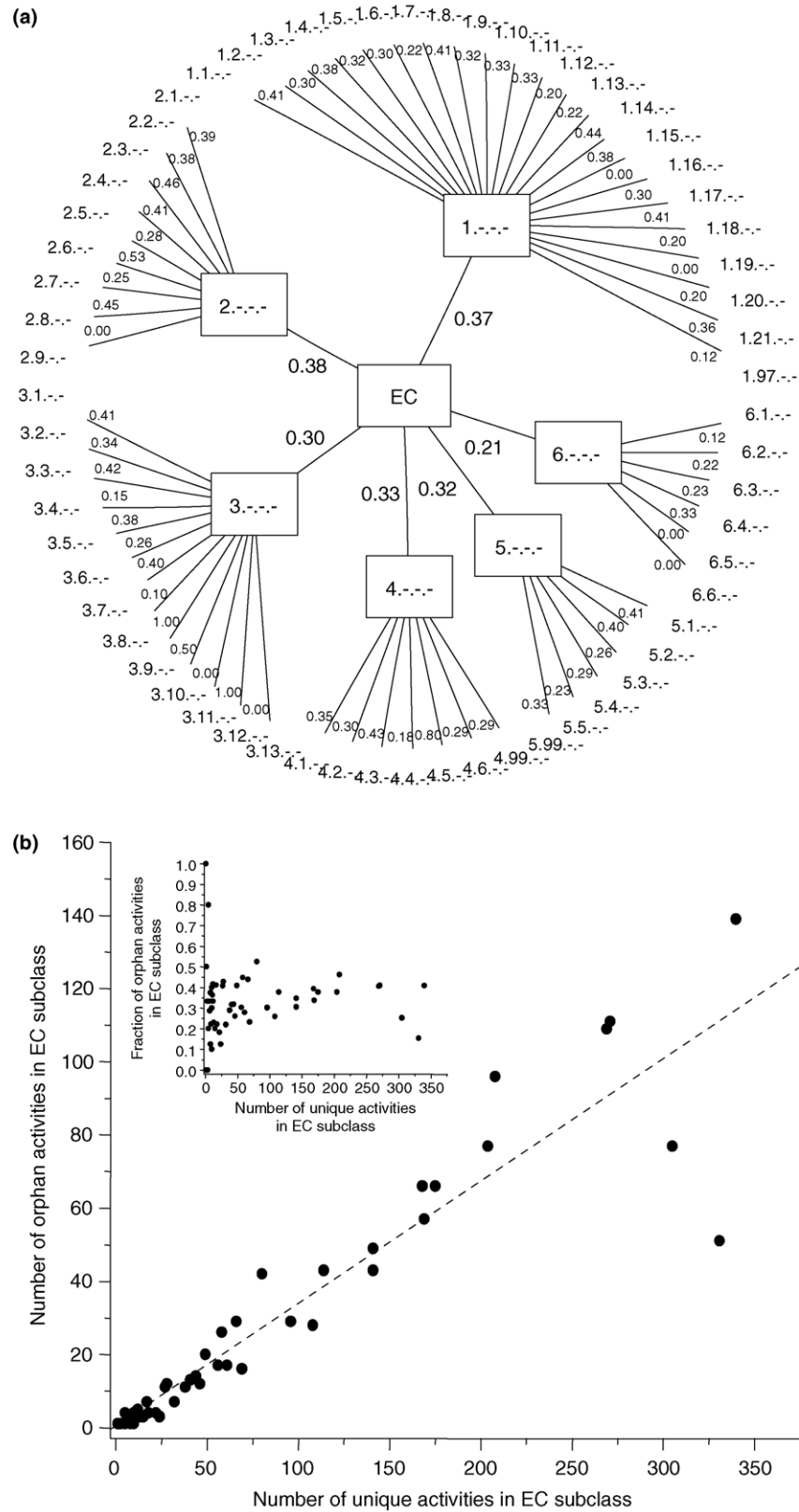


Figure 2. Distribution of orphan activities across the EC classification. **(a)** The EC classification is organized in a tree hierarchy and the fractions of orphans for each class and sub-class are shown on the tree branches. **(b)** The correlation (Pearson's $R = 0.93$) between the numbers of all and orphan activities in different subclasses of the EC hierarchy. Inset shows the fraction of orphan activities versus the size of the subclass. EC numbers in large enzymatic classes are, on average, as likely to be orphan as EC numbers in smaller classes.

described in the ENZYME database [20] were used to build a metabolic network in which network nodes represent metabolic activities (orphan or non-orphan EC numbers) and edges represent connections between the activities established by shared metabolites. Common metabolites such as water or ammonia are not likely to represent functional linkages. Therefore, the 15 most frequent metabolites – ATP, ADP, AMP, CO₂, CoA, glutamate, H, NAD, NADH, NADP, NADPH, NH₃, GLC, orthophosphate and pyrophosphate – were removed before the network was reconstructed. Our results are not sensitive to the exact number of removed metabolites. The analysis of the constructed metabolic network demonstrates only a marginal difference in the neighborhood composition between orphan and non-orphan activities. For example, 32% of the network neighbors of orphan activities are orphans themselves, compared with 29% for neighbors of non-orphan activities.

Organism distribution of orphan activities

How are orphan activities distributed across organisms? To answer this question, we used information available in the BRENDA database [33]. The database references species in which orphan and non-orphan activities were experimentally characterized. Sequence homology to other genes responsible for a non-orphan activity can be used to search for an orthologous gene in a target organism. Such analysis might be followed by genetic or biochemical experiments on the ‘suspect’ gene to validate the presence of the activity. By contrast, experimental characterization of orphan activities in a target organism cannot rely on a similar set of candidate genes and is significantly more difficult. Consequently, the information available in the

BRENDA database will significantly underestimate the true extent of orphan distribution across species in comparison with non-orphan activities. Nevertheless, we found that each orphan activity is, on average, experimentally characterized in more than three organisms (median of the distribution is 2), compared with ~12 organisms (median = 8) for a non-orphan activity.

The number of orphan activities experimentally characterized in an organism correlates well with the total number of activities characterized in that organism (Spearman’s $R = 0.69$, $P < 0.001$; Figure 3). Organisms with a smaller number of experimentally characterized activities usually have a higher fraction of orphans (Spearman’s $R = -0.86$, $P < 0.001$; Figure 3, inset). This suggests that orphan activities are relatively more abundant in less studied species. However, as the total number of experimentally characterized activities increases, the fraction of orphans remains at around 10–15% independently of the network size (Figure 3, inset).

We investigated the original source organisms for known orphan activities using the earliest publication about each activity referenced in the BRENDA database. The fractions of orphan and non-orphan metabolic activities described first in different organisms are shown in Figure 4. For both non-orphan and orphan activities there is a wide distribution of source organisms. The largest fraction of non-orphan activities has been described in *E. coli* (historically, the main biochemical model organism). By contrast, the mammalian species *Homo sapiens* and *Rattus norvegicus* are currently the largest sources of orphan activities. Orphan activities are less likely to be described in several well-studied organisms (see also Figure 3, inset) compared with non-orphan activities.

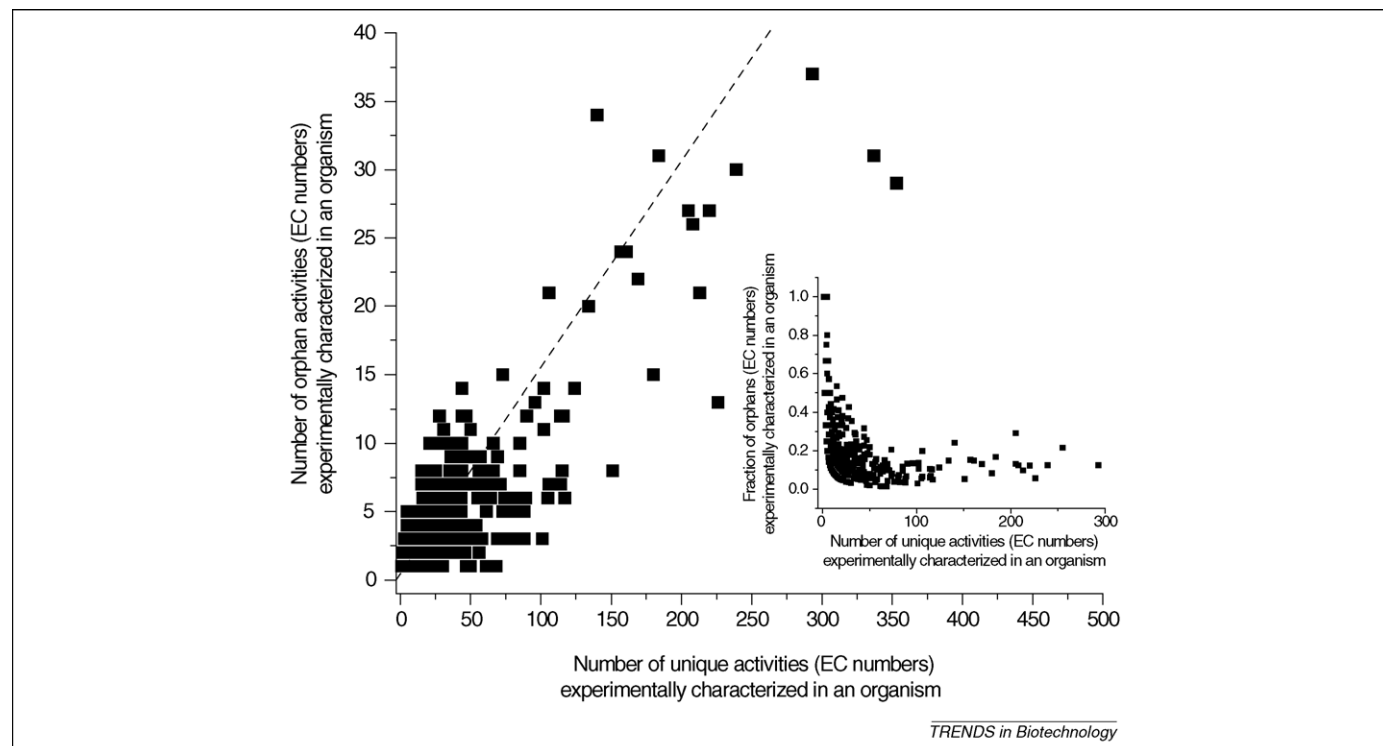


Figure 3. Distribution of orphan activities across organisms. The number of experimentally characterized orphan activities is correlated to the total number of distinct activities in the organisms (Spearman’s $R = 0.69$). Inset shows the fraction of orphans as a function of the total number of distinct, experimentally characterized activities. The fraction of orphans in an organism remains around 10–15% as the number of all experimentally characterized activities increases.

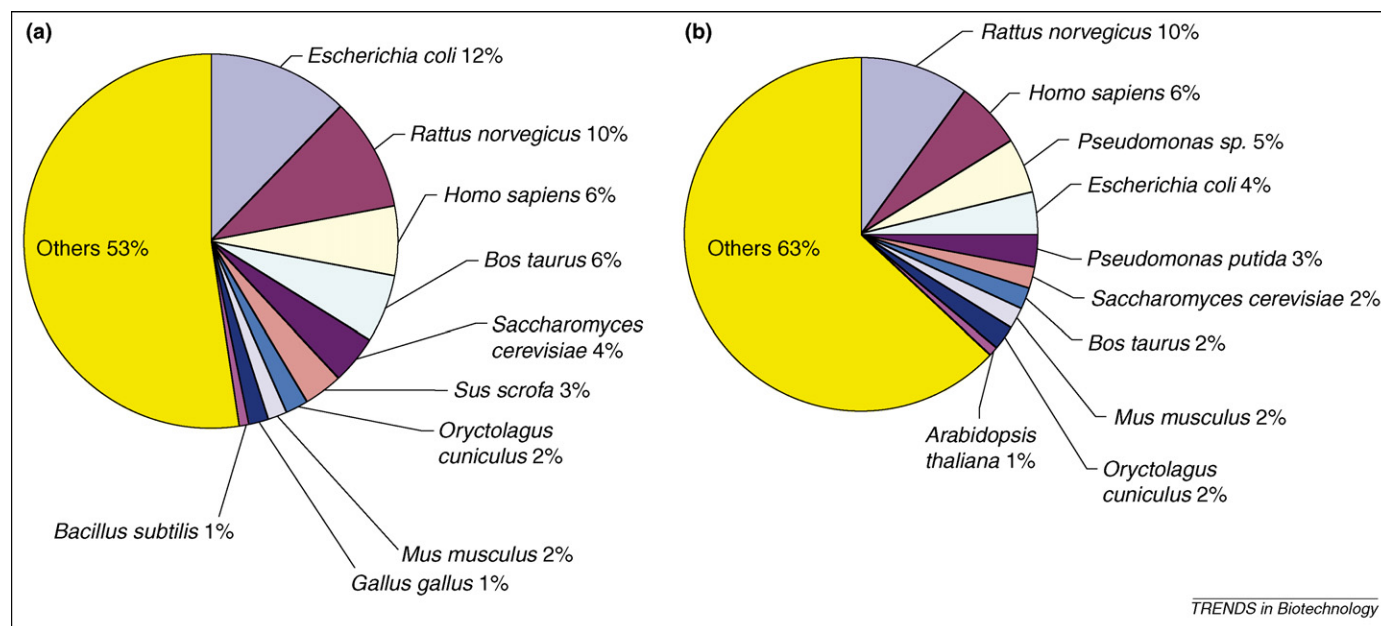


Figure 4. Original source organisms for (a) non-orphan and (b) orphan metabolic activities. The percentages reflect the fractions of known non-orphan (a) and orphan (b) activities first described in various organisms. Orphan activities are relatively less likely to be characterized in well-studied organisms compared with non-orphan activities.

For example, 47% of non-orphan activities were first described in the top 10 source organisms, compared with 37% for orphan activities (Figure 4a,b).

Concluding remarks

One important caveat of our analysis is the incompleteness of the EC classification. Many known, biochemically characterized, enzymatic reactions have yet to be assigned EC numbers [14]. The addition of these, currently unclassified, reactions is likely to increase significantly the fraction of orphan activities. For example, we identified 1289 reactions in the KEGG [34] database and 1439 reactions in the MetaCyc database [35] with no or incomplete EC classification. About 60–75% of these reactions are orphan. Therefore, it is likely that up to 50% of all known biochemical reactions are currently orphan.

One possible cause for orphan activities is an incomplete annotation of the known enzymes in existing databases. For example, an activity identified as orphan might have been already assigned a gene under a rarely used name. To explore this further, we identified official names and synonyms for all orphan activities and searched them against the UNIPROT database [21] using string matching. In agreement with the previous study [14] we found that only a small fraction (~5%) of the orphan activities could be assigned representative genes using this approach. Another possible source of orphan activities might be spurious reactions. For example, an incorrect interpretation of experimental results could lead researchers to postulate a biochemical activity that does not really exist. Although careful consideration of associated experimental evidence will be required to identify such spurious activities, the observation that a significant fraction of orphans have been experimentally characterized in several organisms makes it unlikely that they are dominated by spurious reactions.

Based on the discussion above, it is likely that a majority of identified metabolic orphans represent real enzymatic activities missing associated genes. Some of the responsible

genes are likely to be annotated as hypotheticals or incorrectly assigned a different function. We also envision that some known enzymes could ‘moonlight’ [36] and catalyze orphan activities in addition to their established functions.

In our view, the scale of the orphan activities problem makes it one of the major challenges of modern biochemistry. The orphan problem should be addressed by a combined computational–experimental effort. Several computational and experimental approaches can help to accelerate the identification of genes for metabolic orphans. Although homology-based computational methods [37–40] will not be accurate without reference sequence information, genomic context correlations such as protein fusion events [41], gene neighborhood [42], phylogenetic profiles [43] and mRNA co-expression data [44], can be used to suggest likely candidates [45,46]. These methods can be combined with the partially known structure of metabolic networks to significantly improve the prediction accuracy [16,31,47–49]. The likely candidates identified by computational approaches could then be tested in a high-throughput fashion using phenotype profiling in multiple environmental conditions [50] and parallel metabolomics [10] and fluxomics [51] measurements. Such a combined computational–experimental strategy will help to shrink efficiently the space of orphan activities.

Acknowledgements

We thank Michiel de Hoon, Peter Kharchenko and Andrey Rzhetsky for valuable discussions.

References

- 1 Eisenberg, D. *et al.* (2000) Protein function in the post-genomic era. *Nature* 405, 823–826
- 2 Collins, F.S. *et al.* (2003) A vision for the future of genomics research. *Nature* 422, 835–847
- 3 Selinger, D.W. *et al.* (2003) On the complete determination of biological systems. *Trends Biotechnol.* 21, 251–254
- 4 Carninci, P. *et al.* (2005) The transcriptional landscape of the mammalian genome. *Science* 309, 1559–1563

- 5 Caron, H. *et al.* (2001) The human transcriptome map: clustering of highly expressed genes in chromosomal domains. *Science* 291, 1289–1292
- 6 Perez-Rueda, E. and Collado-Vides, J. (2000) The repertoire of DNA-binding transcriptional regulators in *Escherichia coli* K-12. *Nucleic Acids Res.* 28, 1838–1847
- 7 Harbison, C.T. *et al.* (2004) Transcriptional regulatory code of a eukaryotic genome. *Nature* 431, 99–104
- 8 Ho, Y. *et al.* (2002) Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature* 415, 180–183
- 9 Aebersold, R. and Mann, M. (2003) Mass spectrometry-based proteomics. *Nature* 422, 198–207
- 10 Nielsen, J. and Oliver, S. (2005) The next wave in metabolome analysis. *Trends Biotechnol.* 23, 544–546
- 11 Covert, M.W. *et al.* (2004) Integrating high-throughput and computational data elucidates bacterial networks. *Nature* 429, 92–96
- 12 Fraser, A.G. and Marcotte, E.M. (2004) A probabilistic view of gene function. *Nat. Genet.* 36, 559–564
- 13 Joyce, A.R. and Palsson, B.O. (2006) The model organism as a system: integrating 'omics' data sets. *Nat. Rev. Mol. Cell Biol.* 7, 198–210
- 14 Karp, P.D. (2004) Call for an enzyme genomics initiative. *Genome Biol.* 5, 401 DOI: 10.1186/gb-2004-5-8-401 (<http://genomebiology.com>)
- 15 Lespinet, O. and Labeledan, B. (2005) Orphan enzymes? *Science* 307, 42
- 16 Chen, L. and Vitkup, D. (2006) Predicting genes for orphan metabolic activities using phylogenetic profiles. *Genome Biol.* 7, R17 DOI: 10.1186/gb-2006-7-2-r17 (<http://genomebiology.com/2006/7/2/R17>)
- 17 Lespinet, O. and Labeledan, B. (2006) ORENZA: a web resource for studying ORphan ENZYme activities. *BMC Bioinformatics* 7, 436 DOI: 10.1186/1471-2105-7-436 (<http://www.biomedcentral.com>)
- 18 Lespinet, O. and Labeledan, B. (2006) Puzzling over orphan enzymes. *Cell. Mol. Life Sci.* 63, 517–523
- 19 Lee, T.J. *et al.* (2006) BioWarehouse: a bioinformatics database warehouse toolkit. *BMC Bioinformatics* 7, 170 DOI: 10.1186/1471-2105-7-170 (<http://www.biomedcentral.com>)
- 20 Bairoch, A. (2000) The ENZYME database in 2000. *Nucleic Acids Res.* 28, 304–305
- 21 Bairoch, A. *et al.* (2005) The Universal Protein Resource (UniProt). *Nucleic Acids Res.* 33, D154–D159
- 22 Wu, C.H. *et al.* (2003) The Protein Information Resource. *Nucleic Acids Res.* 31, 345–347
- 23 Karp, P.D. *et al.* (2005) Expansion of the BioCyc collection of pathway/genome databases to 160 genomes. *Nucleic Acids Res.* 33, 6083–6089
- 24 Sacks, W. and Jensen, C.O. (1951) Malease, a hydrazase from corn kernels. *J. Biol. Chem.* 192, 231–236
- 25 Dreyer, J.L. (1985) Isolation and biochemical characterization of maleic-acid hydratase, an iron-requiring hydrolyase. *Eur. J. Biochem.* 150, 145–154
- 26 van der Werf, M.J. *et al.* (1993) Purification and characterization of maleate hydratase from *Pseudomonas pseudoalcaligenes*. *Appl. Environ. Microbiol.* 59, 2823–2829
- 27 van der Werf, M.J. *et al.* (1992) Screening for microorganisms producing D-malate from maleate. *Appl. Environ. Microbiol.* 58, 2854–2860
- 28 Imai, T. (1973) Purification and properties of nicotinamide mononucleotide amidohydrolase from *Azotobacter vinelandii*. *J. Biochem. (Tokyo)* 73, 139–153
- 29 Foster, J.W. *et al.* (1979) Pyridine nucleotide cycle of *Salmonella typhimurium*: isolation and characterization of pncA, pncB, and pncC mutants and utilization of exogenous nicotinamide adenine dinucleotide. *J. Bacteriol.* 137, 1165–1175
- 30 Hillyard, D. *et al.* (1981) The pyridine nucleotide cycle. Studies in *Escherichia coli* and the human cell line D98/AH2. *J. Biol. Chem.* 256, 8491–8497
- 31 Kharchenko, P. *et al.* (2004) Filling gaps in a metabolic network using expression information. *Bioinformatics* 20, 1178–1185
- 32 Kharchenko, P. *et al.* (2005) Expression dynamics of a cellular metabolic network. *Mol. Syst. Biol.* 1, E1–E6
- 33 Schomburg, I. *et al.* (2004) BRENDA, the enzyme database and major new developments. *Nucleic Acids Res.* 32, D431–D432
- 34 Kanehisa, M. *et al.* (2002) The KEGG database at GenomeNet. *Nucleic Acids Res.* 30, 42–46
- 35 Caspi, R. *et al.* (2006) MetaCyc: a multiorganism database of metabolic pathways and enzymes. *Nucleic Acids Res.* 34, D511–D516
- 36 Jeffery, C.J. (1999) Moonlighting proteins. *Trends Biochem. Sci.* 24, 8–11
- 37 Rost, B. (2002) Enzyme function less conserved than anticipated. *J. Mol. Biol.* 318, 595–608
- 38 Tian, W. and Skolnick, J. (2003) How well is enzyme function conserved as a function of pairwise sequence identity? *J. Mol. Biol.* 333, 863–882
- 39 Wilson, C.A. *et al.* (2000) Assessing annotation transfer for genomics: quantifying the relations between protein sequence, structure and function through traditional and probabilistic scores. *J. Mol. Biol.* 297, 233–249
- 40 Henikoff, J.G. *et al.* (1997) Recent enhancements to the Blocks Database servers. *Nucleic Acids Res.* 25, 222–225
- 41 Enright, A.J. *et al.* (1999) Protein interaction maps for complete genomes based on gene fusion events. *Nature* 402, 86–90
- 42 Overbeek, R. *et al.* (1999) The use of gene clusters to infer functional coupling. *Proc. Natl. Acad. Sci. U. S. A.* 96, 2896–2901
- 43 Pellegrini, M. *et al.* (1999) Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc. Natl. Acad. Sci. U. S. A.* 96, 4285–4288
- 44 DeRisi, J.L. *et al.* (1997) Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* 278, 680–686
- 45 Bowers, P.M. *et al.* (2004) Prolinks: a database of protein functional linkages derived from coevolution. *Genome Biol.* 5, R35 (<http://genomebiology.com/2004/5/5/R35>)
- 46 von Mering, C. *et al.* (2003) Genome evolution reveals biochemical networks and functional modules. *Proc. Natl. Acad. Sci. U. S. A.* 100, 15428–15433
- 47 Kharchenko, P. *et al.* (2006) Identifying metabolic enzymes with multiple types of association evidence. *BMC Bioinformatics* 7, 177 DOI: 10.1186/1471-2105-7-177 (<http://www.biomedcentral.com>)
- 48 Karp, P.D. *et al.* (2002) The Pathway Tools software. *Bioinformatics* 18, S225–S232
- 49 Green, M.L. and Karp, P.D. (2004) A Bayesian method for identifying missing enzymes in predicted metabolic pathway databases. *BMC Bioinformatics* 5, 76 DOI: 10.1186/1471-2105-5-76 (<http://www.biomedcentral.com>)
- 50 Dudley, A.M. *et al.* A global view of pleiotropy and phenotypically derived gene function in yeast. *Mol. Syst. Biol.* (in press) DOI:10.1038/msb4100004 (www.nature.com)
- 51 Sauer, U. *et al.* (1999) Metabolic flux ratio analysis of genetic and environmental modulations of *Escherichia coli* central carbon metabolism. *J. Bacteriol.* 181, 6679–6688

Free journals for developing countries

In 2002, the WHO and six medical journal publishers launched the Health InterNetwork Access to Research Initiative, which enabled nearly 70 of the world's poorest countries to gain free or reduced-cost access to biomedical literature through the internet. Currently more than 70 publishers are participating in the program, providing access to over 2000 journals.

Gro Harlem Brundtland, former director-general for the WHO, said that this initiative was "perhaps the biggest step ever taken towards reducing the health information gap between rich and poor countries".

For more information, visit www.who.int/hinari