Molecular Systems Biology 8; Article number 581; doi:10.1038/msb.2012.13
Citation: ***Molecular Systems Biology*** **8:581**

www.molecularsystemsbiology.com

molecular systems biology

Prediction and identification of sequences coding for orphan enzymes using genomic and metagenomic neighbours

Takuji Yamada[1], Alison S Waller[1], Jeroen Raes[2,3], Aleksej Zelezniak[1,4], Nadia Perchat[5,6,7], Alain Perret[5,6,7], Marcel Salanoubat[5,6,7], Kiran R Patil[1], Jean Weissenbach[5,6,7] and Peer Bork[1,8,*]

[1] Structural and Computational Biology Unit, European Molecular Biology Laboratory, Heidelberg, Germany, [2] Molecular and Cellular Interactions Department, VIB, Brussels, Belgium, [3] Vrije Universiteit Brussel, Brussels, Belgium, [4] Department of Systems Biology, Center for Microbial Biotechnology, Technical University of Denmark, Lyngby, Denmark, [5] Commissariat à l'Energie Atomique, Evry, France, [6] Centre National de la Recherche Scientifique, Evry, France, [7] Université d'Evry Val d'Essonne, boulevard François Mitterrand, Evry, France and [8] Max-Delbruck-Centre for Molecular Medicine, Berlin, Germany
* Corresponding author. Structural and Computational Biology Unit, European Molecular Biology Laboratory, Meyerhofstrasse 1, Heidelberg 69117, Germany. Tel.: +49 6221 3878526; Fax: +49 6221 3878517; E-mail: bork@embl.de

Received 21.12.11; accepted 24.3.12

Despite the current wealth of sequencing data, one-third of all biochemically characterized metabolic enzymes lack a corresponding gene or protein sequence, and as such can be considered orphan enzymes. They represent a major gap between our molecular and biochemical knowledge, and consequently are not amenable to modern systemic analyses. As 555 of these orphan enzymes have metabolic pathway neighbours, we developed a global framework that utilizes the pathway and (meta)genomic neighbour information to assign candidate sequences to orphan enzymes. For 131 orphan enzymes (37% of those for which (meta)genomic neighbours are available), we associate sequences to them using scoring parameters with an estimated accuracy of 70%, implying functional annotation of 16 345 gene sequences in numerous (meta)genomes. As a case in point, two of these candidate sequences were experimentally validated to encode the predicted activity. In addition, we augmented the currently available genome-scale metabolic models with these new sequence–function associations and were able to expand the models by on average 8%, with a considerable change in the flux connectivity patterns and improved essentiality prediction.
Molecular Systems Biology **8**: 581; published online 8 May 2012; doi:10.1038/msb.2012.13
Subject Categories: bioinformatics; metabolic and regulatory networks
Keywords: genomics; metabolic pathways; metagenomics; neighbourhood information; orphan enzymes

Introduction

Enzymes are the catalysts that fuel almost all of the chemical reactions necessary for life in the biological cell. Currently, more than 5000 unique enzymes have been sufficiently biochemically characterized that an Enzyme Commission (EC) number could be assigned; however, more than one-third of these lack a corresponding gene or protein sequence, and as such can be considered 'orphan enzymes' (Lespinet and Labedan, 2005; Pouliot and Karp, 2007). Even in this age of genome sequencing, the fraction of newly reported enzymes that are orphan has remained relatively stable with about 40% of the enzymes reported in the past decade being orphan (Supplementary Figure 1). These orphan enzymes participate in central metabolic pathways as well as peripheral ones, and cover all six enzymes classes. The fact that these enzymatic functions are not linked to their cognate sequences means that important biological functions are inaccessible through molecular data-driven studies. Orphan enzymes also render many approaches for functional characterization such as genome or proteome annotation, metabolic modelling, metabolic engineering and drug design incomplete and inaccurate. Closing this gap between biochemical and molecular knowledge will considerably improve the characterization of biological systems at the molecular level.

Many computational approaches have been developed to predict functional annotations for protein sequences. In addition to transferring annotations from homologous proteins, many genome-context methods exist (Huynen *et al*, 2003). Genome-context methods are based on the fact that in prokaryote genomes genes involved in the same metabolic pathway often co-occur in the same genome (Dandekar *et al*, 1998; Pellegrini *et al*, 1999; Yamada *et al*, 2006), are located in proximity to each other or occasionally fused together (Dandekar *et al*, 1998; Enright *et al*, 1999; Huynen *et al*, 2003) or share regulatory sites (Gelfand *et al*, 2000). In addition, information based on post-genomic associations such as

gene-expression profiles, protein–protein interaction data, phenotypic data, or three-dimensional (3D) structure predictions can also be combined with genome-context information to assign a function to a sequence (Hanson *et al*, 2010; Letunic *et al*, 2012). However, linking orphan enzymes to genomic information represents the reverse problem, which is, assigning a sequence to a function.

Several methods have already been developed to assign gene sequences to specific EC numbers for a particular species for which a genome exists and metabolic pathways have been reconstructed. During pathway reconstruction, 'gaps' occur when certain reactions must take place, but none of the genes in the genome are annotated to perform the reaction. The respective gaps are filled using a variety of homology and genome-context methods such as analysis of chromosomal clustering, protein fusion events, co-occurrence profiles, shared regulatory sites and co-expression profiles (Osterman and Overbeek, 2003; Green and Karp, 2004; Chen and Vitkup, 2006; Kharchenko *et al*, 2006). This species-centric approach is limited to the set of candidate genes in a given organism and requires the manual annotation of pathways 'gaps'. Here, we introduce a global search strategy for candidate sequences that encode orphan enzymes operating in known metabolic pathways. It utilizes genomic neighbourhood in genomes and metagenomes and reconciles it with pathway neighbourhood deduced from the KEGG database. Of the ca. 1700 orphan enzymes, 555 are known to operate in pathways and 350 have pathway neighbours that can be connected to genomic information (Figure 1). Here, we integrate genomic-context information (Huynen *et al*, 2003; Harrington *et al*, 2007) derived from 338 completely sequenced genomes and 63 metagenomes, with pathway adjacency to reliably predict candidate sequences for 131 orphan enzymes, more than a third of the tractable ones; as a proof of principle, two of these predictions were functionally validated. Applied to metabolic modelling, these novel gene–enzyme relationships lead to an on average 8% (up to 15%) increase in the enzymatic reaction content of all 120 genome-scale metabolic models probed in our study. The relevance of the addition of the novel orphan enzyme reactions to the metabolic models was attested by improved gene-essentiality predictions for the updated models and altered topology of the flux connectivity within these networks.

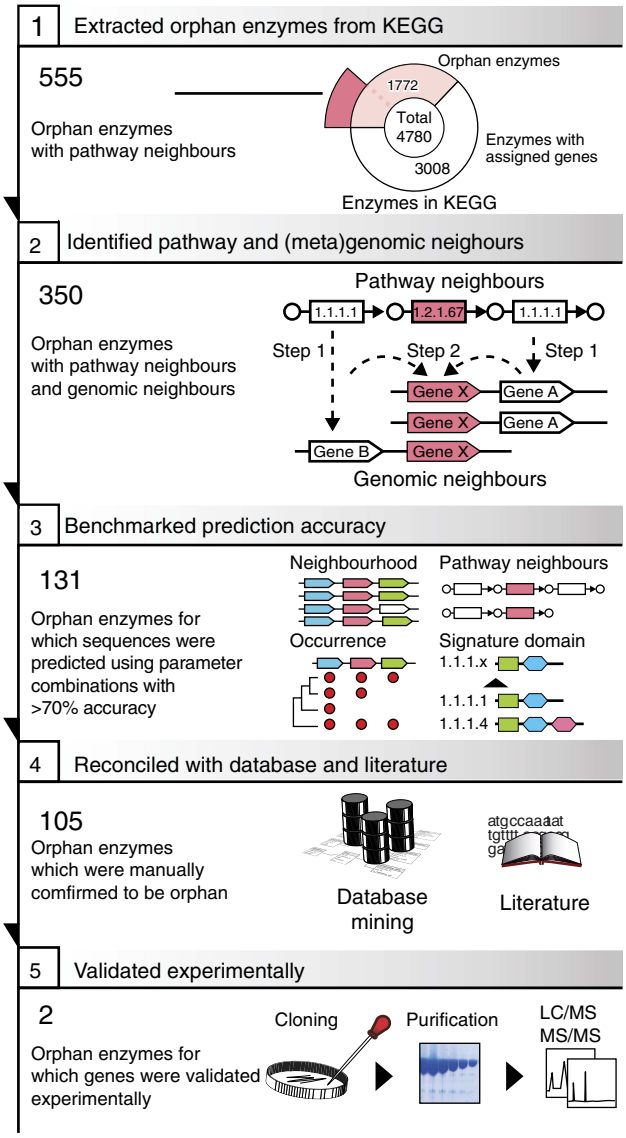


Figure 1 Schematic view of the sequence detection pipeline. A total of 555 enzymes without corresponding sequences were extracted as orphans in metabolic pathways from the 4780 enzymes stored in KEGG (Panel 1). The metabolic pathway neighbours of the orphan enzymes were extracted from KEGG, and the pathway neighbours were then mapped to meta/genomes through homology (Panel 2, Step 1). Genomic neighbours were obtained for 350 orphan enzymes. These genomic neighbours of the pathway neighbours were obtained as possible candidate sequences for the orphan enzymes (Panel2, Step 2). To determine the likelihood that a candidate sequence indeed encodes the orphan enzyme, a scoring scheme was developed involving four parameters 1: the intergenic distance between, and synteny of, genome neighbours, 2: the number of pathway neighbours, 3: the co-occurrence of genes across species and 4: the presence of enzyme class-specific signature domains. Benchmarking of the scoring scheme indicated that some parameter combinations yielded greater than 70% accuracy, resulting in a high-confidence set of predictions for 131 orphan enzymes (Panel 3). We manually confirmed the orphan status for all of the high-confidence predictions by searching for sequences in literature and other databases. About 105 out of the 131 orphans in the KEGG database were verified to be orphans (Panel 4). Finally, we experimentally validated the function of candidate sequences for two enzymatic reactions (Panel 5).

Results

Predicting candidate sequences for orphan enzymes based on (meta)genomic and metabolic pathway neighbours

We first identified 555 orphan enzymes that operate in metabolic pathways (i.e., connected to at least one other enzyme by a common compound) by analysing the KEGG database (Kanehisa *et al*, 2008) (Figure 1). After identifying the EC numbers of the pathway neighbours of these orphan ECs, we retrieved all genes with the same EC number from the 338 prokaryotic genomes of the STRING7 resource (von Mering *et al*, 2007). For the genes in the 63 metagenomes, EC numbers were assigned via a best BLAST match to KEGG orthologous groups (see Materials and methods and Supplementary Section 1). As neighbouring prokaryotic genes are often involved in the same metabolic pathway, we analysed the genomic neighbourhood and retrieved gene

sequences of relevant genome neighbours as candidate genes for the orphan enzymes. Using genomic data, we extracted 400 320 candidate genes and 97 343 from metagenomic data (Supplementary dataset 1).

To quantify the likelihood that a specific candidate gene performs the function of the orphan enzyme, we developed a scoring scheme based on four parameters: (1) The genome neighbourhood score (NBH), which measures the distance between two neighbouring genes as well as the evolutionary conservation of the synteny. This metric captures the biological phenomenon that functionally associated genes are usually clustered in conserved operon structures, (2) The co-occurrence score (COR), which measures how often two genes occur within the same genome. This metric reflects the tendency for members of the same pathway to appear in genomes together, (3) The pathway neighbour score (PNE), which normalizes for the varying numbers of pathway neighbours of the orphan enzyme and (4) The signature domain score (DOM), which indicates whether candidate proteins contain domain(s) that are unique to enzymes catalysing similar reactions to the orphan enzymes (having the same first 3 EC numbers).

Benchmarking revealed that high-confidence candidate sequences can be obtained for over 100 orphan enzymes

To assess the accuracy of our pipeline and to determine the best combination of the four scoring parameters, we benchmarked our predictions using 100 sets of 350 randomly selected enzymes from the KEGG database (that have corresponding sequences) (Figure 2). We considered each of these to be orphan enzymes, applied the newly developed pipeline and then assigned the candidate genes a set of four scores for each of the parameters (NBH, COR, PNE and DOM). We classified the predictions according to their four scores, and then, to estimate the accuracy of each scoring parameter, or combination of parameters, we calculated the proportion of the predictions that were assigned to the correct EC number. First, to understand the predictive power of each of the four

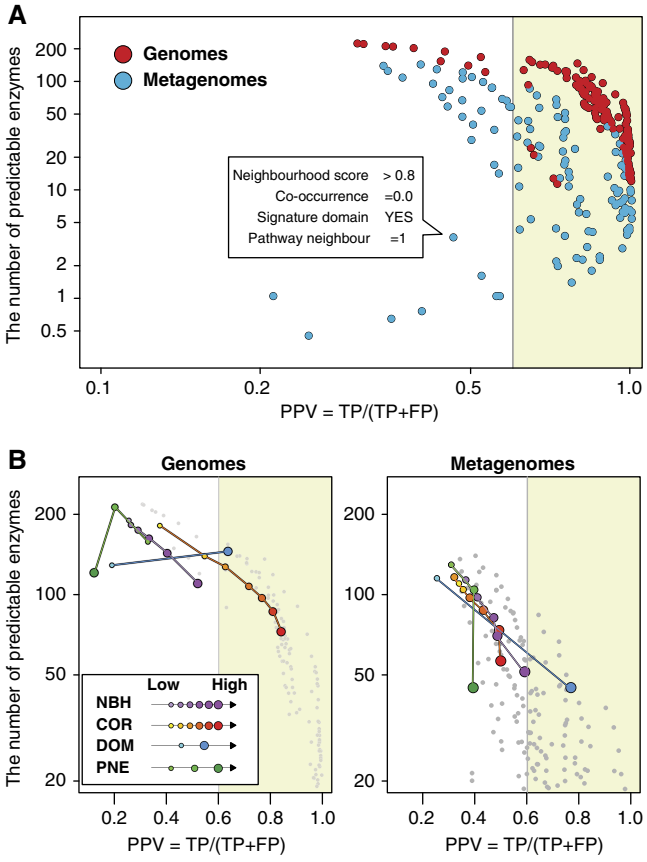


Figure 2 Benchmarking of the scoring parameters. (**A**) Accuracy plot derived from genomic (red) and metagenomic data (blue) using the combination of neighbourhood score (NBH), co-occurrence (COR), signature domains (DOM) and pathway neighbours (PNE). Each candidate gene/neighbouring gene pair was assigned a score for NBH and COR. Each candidate gene was also assigned a PNE and DOM score. The predictions were classified according to their four scores: NBH ($>0.4, >0.5, >0.6, >0.7, >0.8, >0.9$), COR ($>0.1, >0.2, >0.3, >0.4, >0.5, >0.6$), DOM (0 or 1) and PNE (1, 2 or more). Then for each combination of scoring parameters, the number of correct and incorrect EC number assignments was calculated in order to determine the accuracy of each parameter combination. In total, 100 randomized datasets were generated to benchmark the prediction pipeline. Each point represents all predictions from a specific combination of the four parameters (center). The horizontal axis indicates the positive predicted values (PPV), which is calculated as the number of true positives (TP) over the summation of TP and false positives (FP). The vertical axis indicates the number of predictable enzymes. The yellow-shaded area represents the high-confidence set of predictions that was assembled from the union of all points yielding greater than 70% accuracy. (**B**) Accuracy plot for each separate parameter calculated using genomic or metagenomic data. The colour and size of the points represents the intensity of the scores. The grey dots indicate the combined plot in (A).

scoring parameters, we benchmarked each parameter separately, using the genomic and metagenomic data (Figure 2B). Predictions from the genome data illustrate that the co-occurrence score is the best predictor and correlates most strongly with the overall accuracy. The parameter COR in metagenomic data also works well, but for more than 30% of the metagenomic sequences, phylogenetic profiles could not be constructed due to a lack of sequence similarity to currently available data. Here, the signature domains allowed many predictions (Figure 2B). Second, we performed benchmarking for each combination of the four scoring parameters. Although each individual scoring parameter works to some extent, benchmarking clearly shows that integration of the four parameters is better than any one parameter used in isolation (Figure 2A). Finally, we assembled a set of high-confidence predictions from all of the parameter combinations that yielded an accuracy greater than 70% (Figure 2A), resulting in predicted sequences for 131 orphan enzymes (Supplementary Table 2 and Supplementary datasets 2 and 3). For some of the parameter combinations, even more than 90% accuracy is expected.

We then manually investigated the 131 orphan enzymes with high-confidence predictions in more detail. Reconciliation with additional databases and literature searches revealed that 26 out of these 131 already have a sequence deposited in the curated Swissprot database or literature (Supplementary Figure 4 and Supplementary Tables 3 and 4). For 17 of the 26 (65%) database sequences there was homology to sequences from EC numbers that agreed up to at least the first digit (Supplementary Figure 5). Our candidate sequences that have no orthology to the sequences in the database may represent alternative orthologous groups catalysing the same reaction, as about 70% of the EC numbers in KEGG are encoded by more than one orthologous group (Supplementary Figure 3A). Therefore, we do not consider these as mispredictions, but they can no longer be called orphan enzymes, although none of these sequences are indicated in the enzyme-specific databases ExPASy-ENZYME or KEGG. The activities of the remaining 105 orphan enzymes range from core metabolism, such as nucleotide metabolism, to peripheral pathway (Figure 3A, Supplementary Figure 6) and we could assign over 16 000 sequences to these.

Experimental confirmation of the predicted enzymatic function for two candidate sequences

After determining that our pipeline can reveal high-confidence predictions for candidate sequences for orphan enzymes, we

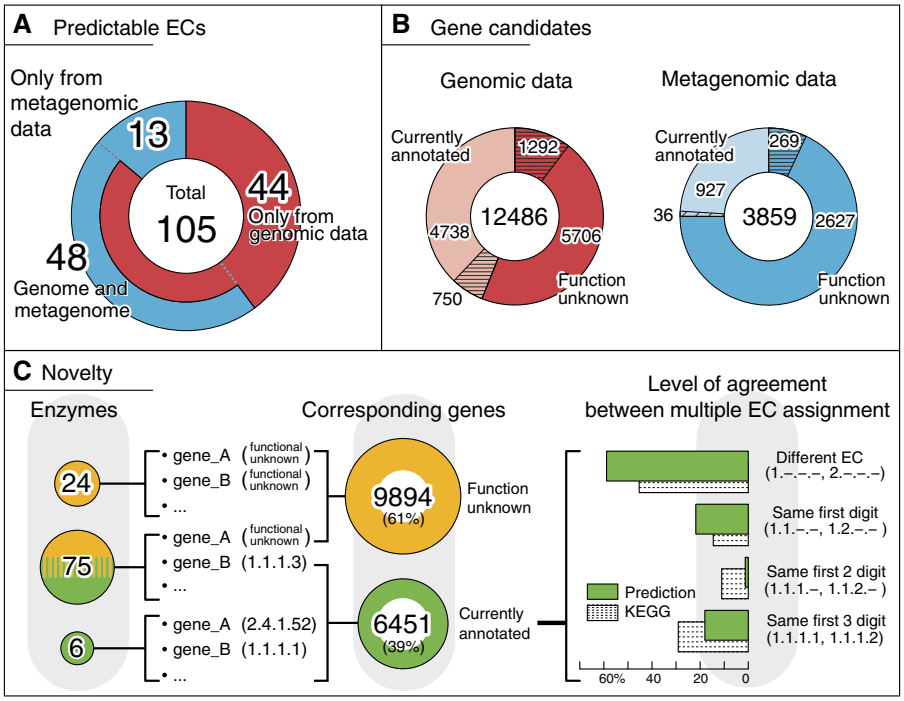


Figure 3 Breakdown of the predicted enzymes. (**A**) The number of EC numbers for which candidate genes can be predicted using parameter combinations with greater than 70% accuracy. Red indicates candidate genes that were derived from only genomic data, and blue indicates candidate genes that were derived only from metagenomic data. (**B**) The pie charts represent the proportion of the gene candidates that have an unknown function versus a current annotation for genes from genomic (red) and metagenomic (blue) data. The striped area represents genes that were detected only in genomic or metagenomic data, whereas the genes represented by the solid colours were identified in both genomic and metagenomic data. (**C**, left) The novelty of the predictions is illustrated at the enzyme level and the gene level. The enzymes were categorized into three categories: (1) all candidate genes for that enzyme are currently annotated as functionally unknown (yellow), (2) some (usually most) of the candidate genes for the enzyme are functionally unknown while others are annotated with an EC number (yellow/green) and (3) all of the candidate genes for that enzyme have a current EC annotation (green). The candidate genes are then divided into functionally unknown (yellow) and currently annotated (green). (C, right) For those 40% of the candidate genes that are currently annotated, we illustrate the level of agreement between our predicted EC number and the current annotation. We overlaid this with similar data from KEGG, as over 30% of the OGs in KEGG are assigned to multiple EC numbers (Supplementary Figure 2). White bars represents multifunctionality of enzymatic activity in KEGG original data and green the currently annotated candidate genes.

performed experimental confirmations. We assessed the ease of experimental validation for some of the high-confidence predictions (e.g., access to gDNA); out of 45 corresponding EC numbers, 15 sequences were amenable to cloning and 7 were chosen for functional validation based on the commercial availability of the reactants as well as the ability to monitor the substrates and products using available analytical methods. Of the six proteins that were successfully heterologously expressed, the proposed function was verified for two enzymes (Supplementary Section 5).

We succeeded in experimentally verifying the correct function of candidate sequences for EC 2.6.1.14 (asparagine oxo-acid transaminase, Figure 4A left) and EC 2.6.1.38 (histidine transaminase, Figure 4A right), showing the reliability of this prediction pipeline. Using the prediction pipeline, we retrieved candidate sequences for these two enzymes using genomic (EC 2.6.1.14) or genomic and metagenomic data (EC 2.6.1.38) (Figure 4B). Candidate sequences were heterologously expressed, and in assays containing the purified candidate proteins and the substrates, the expected reaction products were unambiguously identified using a combination of LC/MS and MS/MS (Figure 4C and Supplementary Figures 11–17; see Supplementary Section 5 for details). Concerning the four other candidate proteins (for EC 2.1.1.19, 2.1.1.68, 2.3.1.32 and 2.7.1.28), neither product formation nor substrate consumption was detected in enzymatic assays through LC/MS. For EC 2.7.1.28, a peak of very slight intensity with a m/z consistent with the one of the products, D-glyceraldehyde-3-phosphate, could be detected. Nevertheless, LC/MS analyses could not lead us to conclude the predicted activity, as the substrate D-glyceraldehyde could never be detected, and neither ATP consumption nor ADP formation could be established. In addition, two different continuous spectrophotometric assays were set up to try to confirm the predicted activity. In the first one, the production of ADP was coupled to the consumption of NADH, using commercial pyruvate kinase and lactate dehydrogenase, along with phosphoenolpyruvate. In the second one, the production of glyceraldehyde-3-phosphate was coupled to the production of NADH using commercial glyceraldehyde-3-phosphate dehydrogenase. In both cases, the assays were inconclusive. However, as detailed in the Discussion, there can be many difficulties in the experimental process to validate an enzymes' function therefore absence of evidence is not necessarily evidence of absence.

Assessing functional novelty and multifunctionality for the candidate sequences

After the benchmarking and experimental validations showed the reliability of the pipeline, we examined the validated orphan enzymes and their corresponding genes in more detail. As expected from the benchmarking, the number of enzymes for which candidate sequences can be predicted was greater for genomic than for metagenomic data (Figure 3A). This is due in part to the short length of contigs in metagenomic data, as this reduces the number of genomic neighbours that are available for the first screen of our pipeline. For 48 enzymes, candidate sequences were predicted from both metagenomic and genomic data. However, for 13 orphan enzymes we found candidate sequences only in metagenomic data, exemplifying the ability of this pipeline to detect sequences from bacteria in environmental samples. One example is biotin CoA synthetase (6.2.1.11) found in the gut metagenomes. This prediction is supported by the fact that bacterial synthesis and degradation of biotin is known to be important in the human large intestines (Said, 2009; Arumugam *et al*, 2011).

As many as 9884 of the individual candidate sequences (about 60%) are annotated as 'function unknown', 'hypothetical' or similar (Figure 3B), and assigning them to orphan activities thus provides functional annotations that can be further propagated into newly sequenced genomes through the use of homology-based annotation methods. An even higher fraction of unannotated sequences predicted to code for orphan enzymes can be found in metagenomics data (Figure 3B).

Overall, 40% of the candidate sequences are already annotated with an EC number (Figure 3C). We believe that the vast majority of these imply multifunctionality, as this is a common attribute of enzymes (Nobeli *et al*, 2009). Indeed, over 30% of the genes in the KEGG database are assigned to more than one EC number (Supplementary Figure 3B). Of these multifunctional enzymes in KEGG, about 30% are assigned to EC numbers that agree up to 3 digits, while another 50% have no agreement between the different EC numbers. Our candidate sequences that have a current annotation and are potentially multifunctional have a similar trend in the level of agreement between the assigned and predicted EC numbers (Figure 3C). It is therefore plausible that these genes with current annotations represent multifunctional enzymes, although we cannot rule out either mispredictions from our pipeline nor errors in the current annotations due to the automatic nature of most genome annotations.

In addition to coupling unannotated sequences to specific functions, our pedictions also provided putative functions for certain Domains of Unknown Function (DUF domains). The prediction pipeline led to the identification of five DUF domains that are unique to candidates of orphan enzymes. For example, DUF2254 is only present in genes predicted to encode the orphan EC 2.4.2.15, guanosine phosphorylase (Supplementary Table 5). As a byproduct of our pipeline, we also identified 150 DUF domains that are unique to specific non-orphan EC numbers yet had not been annotated so far (Supplementary Table 6), and should improve various studies that use domain databases like Pfam or SMART (Finn *et al*, 2010; Letunic *et al*, 2012).

High-confidence predictions yield putative sequences for enzymes with commercial and biotechnological applications

Some orphan enzymes from our high-confidence predictions have potential commercial or medical applications, for example EC 2.8.1.5, thiosulphate—dithiol sulphurtransferase, involved in sulphur metabolic pathways that are essential in many pathogenic bacteria, but not present in humans, and could therefore provide drug targets. In addition, four of the orphan enzymes with very high scores could be utilized for

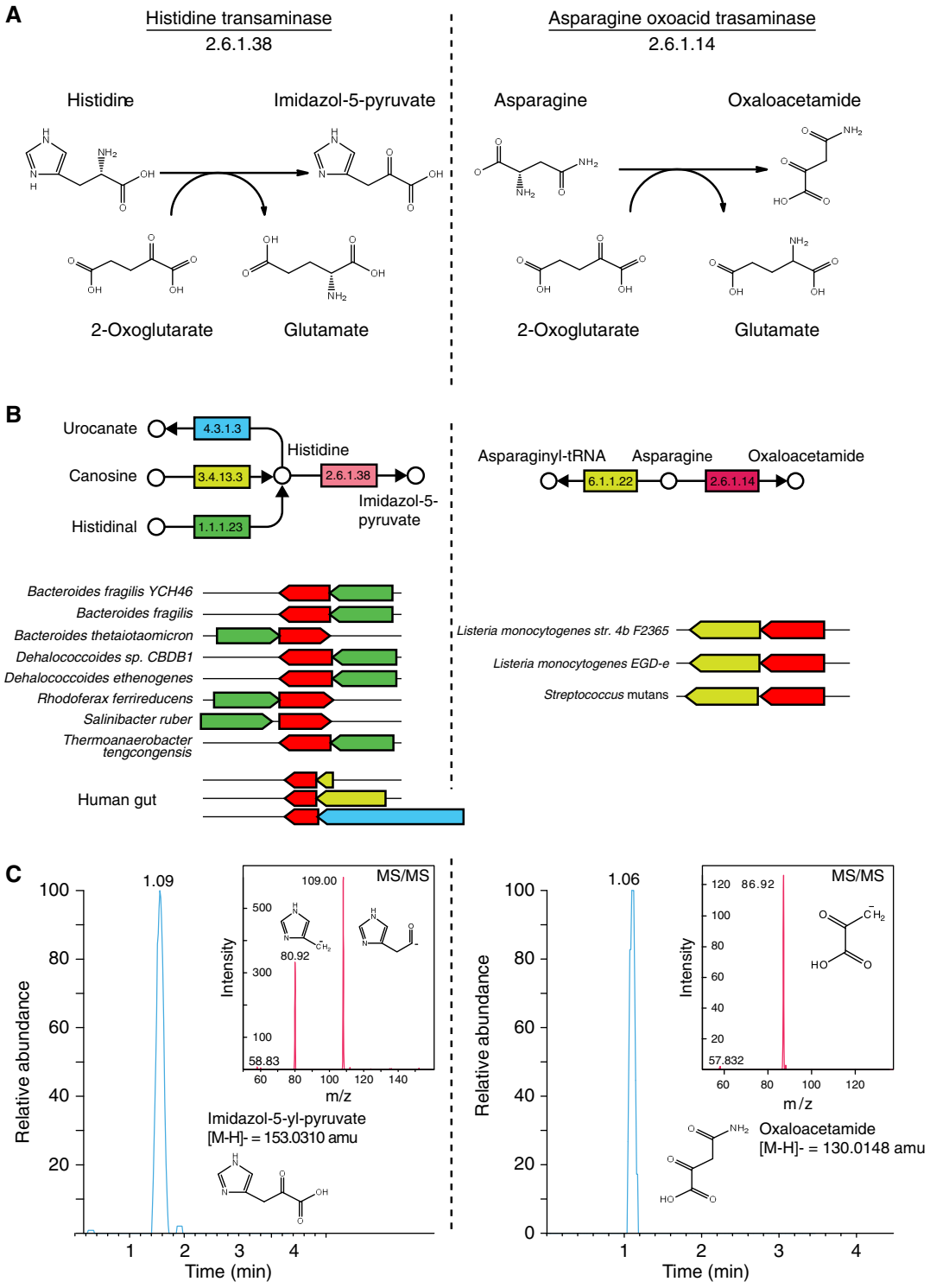


Figure 4 Orphan enzymes with experimental validation. (**A**) The chemical reactions catalysed by the two orphan enzymes for which candidate sequences were experimentally validated (**B**) Metabolic pathway neighbours and genome neighbours of the orphan enzymes. (**C**) Extracted ion chromatogram (EIC) and MS/MS plots supporting the identity of the expected reaction products.

the synthesis of commercially available nutraceuticals, one could be used in the food industry and another two have applications in bioremediation (Supplementary Table 7). Furthermore, candidate genes were predicted for phenylpyruvate decarboxylase (EC 4.1.1.43), using a parameter combination with 80% accuracy, that converts phenylpyruvate to phenylacetaldehyde, which is the first and crucial step in the synthesis of branched-chain higher alcohols as biofuels (Atsumi *et al*, 2008). The genes that our analysis linked to phenylpyruvate decarboxylase represent a valuable repertoire

for efficient production of biofuels. All of the predictions and sequences are available at our website (http://www.bork.embl.de/~yamada/orphan_enzymes/).

Orphan enzyme reactions improve the accuracy of genome-scale metabolic models

To measure the impact of our findings on genome-scale metabolic models, we analysed reactions represented by the 120 metabolic models obtained from the Model SEED database (Henry *et al*, 2010) (Supplementary Table 8) and determined if any of them contained orphan enzymes for which we have reliable predictions. For most of the metabolic models, the reactions encoded by the orphan enzymes were not included, and thereby represent novel reactions. For each model, there were around 40 novel reactions averaging about 5–10% of total reactions (Figure 5). Interestingly, this trend was observed for manually reconstructed models as well as for automatically reconstructed models. For example, in the most recent reconstruction for *Escherichia coli* (Orth *et al*, 2011), 49 novel reactions (from parameter combinations with estimated accuracy >70%) could be added to the model while only 1 reaction in the current model represents one of these orphan enzymes (Supplementary Table 9). The fact that these orphan enzymes are not represented in the metabolic models shows that the completeness of these reconstructions is heavily reliant on the current annotation quality, and thus considerably affected by orphan enzymes.

To estimate the impact of the novel reactions on flux simulations using these models, we performed flux coupling analysis (FCA) (Burgard *et al*, 2004), before and after adding the corresponding novel orphan enzyme reactions into the models. Comparative FCA helped us to systematically elucidate the effects of adding new reactions on the topology of flux connectivity at the whole-network scale (see Materials and methods). In the case of the latest (manually curated) *E. coli* model (Orth *et al*, 2011), a large fraction (16%) of dependency relationships between the fluxes were altered following the addition of 49 novel reactions (Supplementary Figure 9). In general, the addition of the new reactions led to a decrease in the number of coupled reactions. For example, changes were detected in vitamin biosynthesis pathways where the addition of the orphan reactions led to a decrease in the number of fully coupled reactions (reaction pairs for which the corresponding fluxes are directly proportional to each other). This trend shows that the new reactions are relatively well embedded within the existing network and provide additional branches for flux routing.

Then to establish if adding the orphan enzyme reactions to the current models improves their accuracy, we determined if the updated models were better in predicting gene essentiality. For ~80% of the 72 SEED models tested, there was at least one gene for which the prediction changed from essential to non-essential, with the largest change being 26 genes in the case of *Salmonella typhimurium*. For the rest ~20% of the models, no change in essentiality predictions was observed following the addition of the orphan enzyme reactions (Supplementary Figure 10). Addition of new reactions to a model can change the existing predictions in two different ways; (i) false essential predictions can then be correctly predicted as non-essential, and/or, (ii) some of the true essential predictions are later wrongly predicted as non-essential. To determine if the observed changes in essentiality predictions were biologically meaningful, we compared the experimentally determined essentiality status of the genes to the essentiality status predicted from the models with and without the orphan enzyme reactions. Four of the species probed in our study had genome-wide gene-essentiality data available. For the *Bacillus subtilis* model, no changes were predicted for gene essentiality following the addition of the corresponding orphan enzyme reactions. However, for the other three species, *E. coli K-12*,

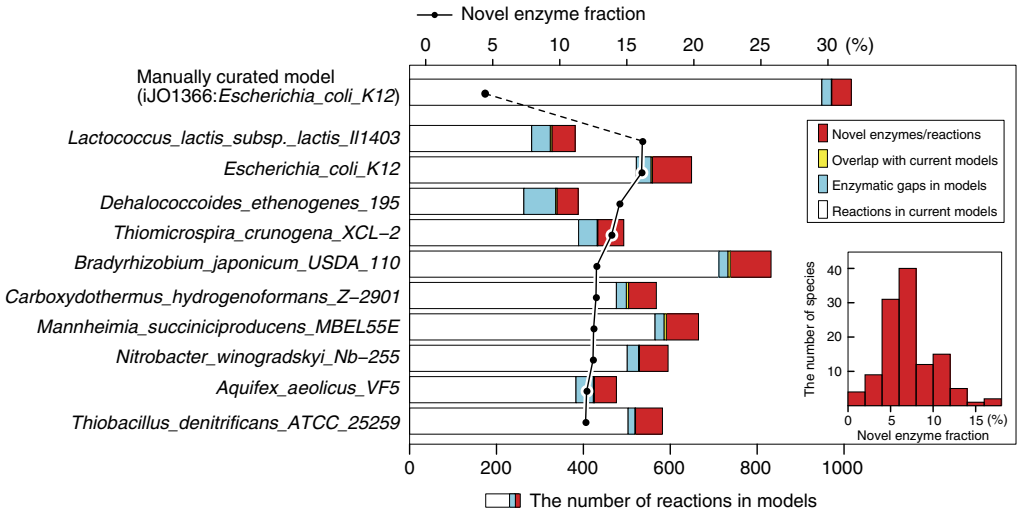
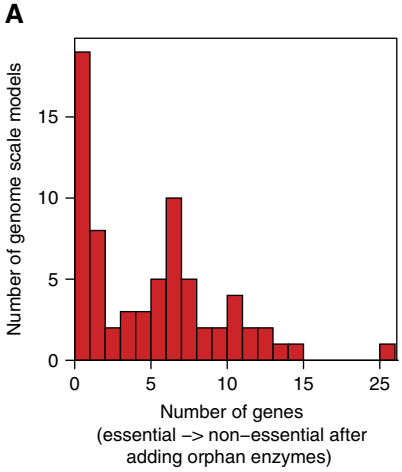


Figure 5 Enrichment of genome-scale metabolic models by orphan enzymes. The barplot shows the number of reactions in 120 publically available genome-scale metabolic models from Model Seed (Henry *et al*, 2010) (white) and novel enzymatic reactions for these models predicted by our pipeline with over 70% accuracy (red). Current gaps in terms of enzyme-catalysed reactions are also shown (blue). The line graph plots the fraction of novel enzymes contributed by orphan enzymes. Only the 10 models with the highest fraction of novel reactions are shown. The histogram in the lower right shows the distribution of the novel fraction for 120 seed models used in this study (Supplementary Table 7).



B

Species	Model	Gene for which essentiality changed	Prediction with original model	Prediction with orphan enzyme reactions	Experimental validation
Bacillus subtilis subsp. subtilis str. 168	Seed224308.1	–	–	–	–
Escherichia coli K12	Seed83333.1	b3939	E	N	N
		b4013	E	N	N
Campylobacter jejuni subsp. Jejuni NCTC 11168	Seed192222.1	Cj1727C	E	N	N
		Cj0013	E	N	N
		Cj0269C	E	N	N
		Cj0531	E	N	N
		Cj0574	E	N	N
		Cj0575	E	N	N
		Cj1503C	E	N	N
		Cj0790	E	N	N
		Cj0088	E	N	N
		Cj1619	E	N	N
Helicobacter pylori J99	Seed85963.1	HP0121	E	N	N
		HP0171	E	N	N
		HP0183	E	N	N

N: Non-essential E: Essential

Figure 6 Gene-essentiality predictions for genome-scale metabolic models including orphan enzymes. (**A**) Distribution of the number of genes for which the computational prediction changed from essential to non-essential across 72 genome-scale metabolic models (Supplementary Table 8). (**B**) Comparison of the gene-essentiality predictions from the models with/without orphan enzymes to essentiality derived from experimental data. Only genes for which addition of the orphan enzymes altered the existing predictions are shown.

Campylobacter jejuni subsp. Jejuni NCTC 11168 and *Helicobacter pylori J99*, predictions for a total of 15 genes changed to non-essential due to addition of the orphan enzyme reactions. All of these changes to non-essential were then found to be consistent with the results from experimental genome-wide knock-out data, illustrating that the addition of the orphan enzyme reactions to the metabolic models made them more accurate for gene knock-out analyses (Figure 6B).

Discussion

Here we have described a global strategy to predict candidate sequences for orphan enzymes. Candidate sequences were obtained using a combination of metabolic pathway adjacency and genomic neighbourhood information. Overall, a lower proportion of candidate sequences were obtained using metagenomic data, than genomic data, but this might only be due to the restrictions we had to impose: Sanger and 454 samples that have a low coverage of the respective genomes. Although many novel enzymes and organisms may be represented in metagenomic samples, the human gut and marine metagenomes that we used are complex communities with hundreds of species (Qin *et al*, 2010), and a long tail of low-abundance organisms (Arumugam *et al*, 2011), thereby limiting the coverage of each individual genome and thus the extent of assembly. Consequently, the majority of the contigs that we analysed only contained two genes, thus limiting the number of neighbour gene pairs that can be detected (Supplementary Figures 7 and 8). Although some available metagenomic datasets have a large number of long contigs, these are usually dominated by a few genomes and thus would not offer access to an increased number of genomes (Tyson *et al*, 2004; Garcia Martin *et al*, 2006). In the future, contigs will become longer, due to increases in read lengths and improvements in assembly algorithms, therefore enhancing the ability of this pipeline to make predictions from metagenomic data allowing greater access to novel activities of hidden environmental samples.

In addition to the benchmarking, we supported our predictions with the experimental validation of the proposed enzymatic function for two out of six heterologously expressed candidate proteins. The ratio of experimental successes is lower than the 70% expected accuracy. However, we would not expect the ratio of experimental successes to be equivalent to the theoretical prediction accuracy. The experimental process to validate a specific enzymatic function is a very complex process involving many variables. First, an enzyme can be purified in a soluble form but will become inactive during the purification process due to improper handling or exposure to unfavourable conditions such as oxygen. In addition, the proteins purified in this study were tagged with a histidine (his-tagged), as many heterologously expressed proteins are. The addition of a terminal his-tag can dramatically decrease the activity of a protein (Kadas *et al*, 2008) or render it totally inactive (Albermann *et al*, 2000; Halliwell *et al*, 2001). Moreover, there are many variables to optimize for the enzymatic activity tests. Only by adjusting the buffer type, buffer pH, cofactors, time of incubation, temperature of incubation or the analytical methods used might a certain assay become successful. For example, in assay optimization trials for EC 2.6.1.38 we changed the mobile phase for the LC/MS from 10 mM ammonium acetate to water and the peak area of the product glutamate was increased more than 11 times (Supplementary Figure 16). However, there is a practical limit to how many permutations of experimental conditions can be attempted, and only if the initial screening assay is close to the optimal conditions further optimization is feasible. Yet, the two validations in hand are a proof of principle for our approach and even without further experimental validation the benchmarks indicated high-accuracy candidate sequences for 131 orphan enzymes, more than a third of the tractable enzymes stored in pathway databases.

Then to assess the impact of this expanded enzyme knowledge on systems biology, we compared the currently available genome-scale metabolic models with and without the addition of the orphan enzymes with high-confidence predictions. Subsequently, gene-knockout simulations

showed that some genes considered to be essential in the current models became non-essential after the addition of the orphan enzymes. The addition of these orphan enzymes increased the accuracy of the models as all genes for which gene essentiality changed now agree with the experimentally determined essentiality status of the gene. Interestingly, several of the reactions for which the essential to non-essential predictions changed were reactions introduced by the automated gap-filling procedure during the reconstruction process. This observation suggests that the orphan enzyme reactions will not only influence the model simulations but also likely affect the gap-filling procedure, and thereby the reaction content of the final model, beyond simple addition of few new reactions. Taken together, the percentage of novel reactions, FCA and improved gene-essentiality predictions mean that our findings will improve the automatic as well as the manual reconstruction process for genome-scale metabolic models and applications thereof (Oberhardt *et al*, 2009).

About 70% of the orphan enzymes in KEGG do not have pathway neighbours and are thus not amenable to our current pipeline (Figure 1). However, in the future, our candidate gene identification pipeline could be modified to identify other genes that might be functionally related to the orphan enzymes through the integration of genome-scale functional data, such as gene lethality screens (Nichols *et al*, 2011), genetic interactions (Costanzo *et al*, 2010) or gene-expression profiles. This should enable one to retrieve candidate genes by searching the gene neighbourhood of the orthologs of these genes that are functionally related to the orphan enzymes. Furthermore, the current pipeline is only applicable to prokaryotic genomes. However, it could be extended to partially analyse fungal genomes as certain secondary metabolite pathways are known to be organized in gene clusters (Regueira *et al*, 2011).

The linkage of sequences to these orphan functions implies that these functions can be utilized in genome-, transcriptome- and proteome-based methods. Here we illustrated the impact on genome-scale metabolic models. This benefit will be propagated into many different biological systems as these sequences will act as bait so that the newly sequenced genomes can be ascribed these functions through homology-based annotation methods. This is the first systematic approach to retrieve sequences for many orphan enzymes, and the developed computational framework can be applied to additional genomes and metagenomes as they get sequenced.

Materials and methods

Construction of genomic and metagenomic datasets

For genome data, the 338 fully sequenced prokaryote genomes stored in the STRING v7 database (von Mering *et al*, 2007) were used. For metagenomic data, we obtained sequencing data from 37 metagenomes from the human gut and 26 metagenomes from the ocean (Supplementary Table 1). The human gut metagenomes were sequenced by Sanger sequencing, and assembled with the Arachne assembler using SMASHcommunity (Arumugam *et al*, 2010). The specific samples consist of samples from 22 Europeans (Arumugam *et al*, 2011), 13 Japanese (Kurokawa *et al*, 2007) and 2 Americans (Gill *et al*, 2006). The majority of the ocean metagenomes were from the Global Ocean Sampling Expeditions (Venter *et al*, 2004; Rusch *et al*, 2007). Specifically, sequences were obtained for 18 stations: GOS_GS000c, GOS_GS001c, GOS_GS004, GOS_GS007, GOS_GS008, GOS_GS009, GOS_GS010, GOS_GS013, GOS_GS015, GOS_GS016, GOS_GS019, GOS_GS022, GOS_GS023, GOS_GS049, GOS_GS112a, GOS_GS116, GOS_GS121 and GOS_GS122a. Additional polar metagenomes were added one from an Arctic sample (pyrosequencing (Alonso-Saez *et al*, submitted—sequences will be available upon request)), and four from the Antarctic (NCBI project IDs 30009, 30011). The reads from these metagenomes were assembled with the Celera assembler using SMASHcommunity default settings (Arumugam *et al*, 2010).

Enzyme data and candidate sequence extraction

The KEGG pathway database (v57) was queried and all EC numbers without any associated sequence were identified as orphan EC numbers. Next, pathway information about adjacent enzymes was extracted from XML/KGML data and parsed by in-house ruby scripts. Pathway neighbours were defined as enzymes that are connected to each other through a common substrate. After identifying the EC numbers of the pathway neighbours of the orphan ECs, we retrieved all genes with the same EC number from the 338 prokaryotic genomes of the STRING7 resource (von Mering *et al*, 2007). In order to map the pathway neighbours to genes in the 63 metagenomes, we first assigned the metagenomic genes to KOs using the best hit from a BlastP against the KEGG proteins (>60 bits), using the SMASHcommunity pipeline (Arumugam *et al*, 2010). Finally, genes adjacent to the genes for the pathway neighbours of the orphan enzymes were then extracted as candidate genes for orphan ECs. Only genes closer than 300 bps were considered genomic neighbours.

Neighbourhood score (NBH)

The neighbourhood score indicates the probability that neighbouring genes participate in the same metabolic pathway, it is based on the intergenic distance as well as the conservation of synteny across species. For genomic data, we utilized the neighbourhood score from the STRING database (v7) (von Mering *et al*, 2007). For metagenomic data, the probability was derived from 2D histograms of gene distance and conservation rate of the synteny (Harrington *et al*, 2007). As such, pairs of genes are assigned a neighbourhood score between 0 and 1.

Co-occurrence score (COR)

For genomic data, co-occurrence scores were taken from the STRING database (v7) (von Mering *et al*, 2007). For metagenomic data, phylogenetic profiles for each gene (vectors composed of 1 and 0 representing presence and absence of genes) were constructed by blasting against 338 fully sequenced prokaryotic genomes (blast bit score $\geqslant 60$). Then for each pair of genes, the pearson correlation coefficient was calculated between each pair of phylogenetic profiles and used as the co-occurence score. As such, pairs of genes are assigned a co-occurrence score between 0 and 1.

Signature domain score (DOM)

Signature domains represent unique domains for each EC sub-subclass (all ECs having the same first 3 digits). Domain information for enzyme genes was derived from the KEGG ENZYME database (v57). This domain list was then clustered to identify domain(s) that were unique for each EC sub-subclass (Supplementary datasets 4 and 5). For the candidate sequences, domains were identified by HMMER3 search (Eddy, 2009) against PFAM database (Finn *et al*, 2010). The domains in the candidate sequences were then checked against the list of sub-subclass-specific domains. The DOM score thus represents a binary score indicating if the candidate sequence contains the domain(s) that are unique to the EC sub-subclass.

Pathway neighbour score (PNE)

The pathway neighbour score indicates the number of adjacent enzyme genes on the pathway. After MCL clustering of candidate

genes using their homology (bit score >60, $l=1.1$) (Enright *et al*, 2002), we counted the number of adjacent genes encoding pathway neighbours of the orphan enzyme. Candidate sequences were thus assigned a PNE score of one, two or three or more.

Benchmarking with randomized data

In order to estimate the predictive power of the four scoring parameters, we benchmarked our prediction pipeline using data from enzymes with assigned sequences in KEGG pathways. About 350 non-orphan enzymes were randomly extracted from the KEGG pathway database v. 57. This number is the same as that of orphan enzymes in KEGG pathway. In addition, these enzymes were chosen so that the distribution of node degree (network structure) was the same as for the orphan enzymes. These enzymes were then treated as orphan enzymes and candidate sequences were generated using the computational pipeline described above, and each prediction was assigned a set of four scores (NBH, COR, DOM and PNE). The predictions were classified according to their four scores. For genomic data: NBH ($>0.4, >0.5, >0.6, >0.7, >0.8, >0.9$), COR ($>0.1, >0.2, >0.3, >0.4, >0.5, >0.6$), DOM (0 or 1) and PNE (1, 2 or more). Due to the different distribution of the COR score in metagenomic data, the COR score was classified as COR ($>0.2, >0.4, >0.6$, or not determined). Due to the lack of sequence homology with current genomes, co-occurrence scores could not be determined for more than 30% of the genes. To estimate the accuracy for each combination of scoring parameters (120 parameter combinations in total for metagenomic data and 144 parameter combinations for genomic data), the number of correct and incorrect EC number assignments was calculated. In total, 100 randomized datasets were generated to benchmark the scoring parameters. Then to obtain a high-confidence set of candidate sequences, we took the union from all of the parameter combinations that yielded an accuracy of $>70\%$. Finally, to estimate the overall accuracy of the high-confidence set we made a non-redundant set of predictions from the union (accuracy $>70\%$), and then calculated the number of correct and incorrect predictions in this set for each randomized set. For genomic data the mean accuracy was above 85% and for metagenomic data the mean accuracy was above 70% (Supplementary Figure 2). In addition, to examine the predictive power of the individual scoring parameters, we performed a similar benchmarking protocol except that the predictions were only classified according to a single scoring parameter, each in turn, and the binning was more fine-grained than for the combination of the scoring parameters.

Genome-scale metabolic models

About 120 publically available metabolic models were downloaded from Model Seed (Overbeek *et al*, 2005) (Supplementary Table 8) as SBML files. In the case where sequence candidates for orphan enzymes were identified in a species, chemical reactions corresponding to them were compared with reaction list from the model for the same organism in order to identify novel reactions.

Flux coupling analysis

FCA was performed by using the algorithm proposed by Burgard *et al* (2004). The algorithm was implemented in C++ with IBM ILOG CPLEX Optimizer. FCA categorizes relationships between two reactions into three categories, according to the nature of dependency between the fluxes through these reactions. That is, (1) fully coupled, (2) directionally coupled and (3) partially coupled. Two fluxes are fully coupled if the activity of one fully determines the activity of the other and vice versa. A reaction pair is directionally coupled if flux activity of one implies that of the other but not the other way around. A reaction pair is partially coupled if each flux implies activity of the other, however, still allowing a certain degree of flexibility in their flux values. FCA for the *E. coli* model iJO1366 was performed under glucose minimal medium conditions as stated in the original publication (Orth *et al*, 2011), following a preprocessing step to remove blocked reactions. For the novel metabolites introduced in the model, a drain reaction to the extracellular environment was also added. The number of coupled reaction pairs in the *E. coli* model considerably reduced after adding novel reactions, suggesting that these additional reactions provide alternative routes for supplying substrates and for consuming products of the existing reactions.

Prediction of gene essentiality

Each model obtained from Model SEED database (http://seed-viewer.theseed.org) (Henry *et al*, 2010) was constrained for LB-rich medium with glucose (Oh *et al*, 2007) (Supplementary Table 9). Some of the SEED models were unable to have a non-zero biomass flux under these conditions and required the presence of specific ions/vitamins in the environment. To account for these special requirements, we determined the minimum number of additional media components required for each model by using a Mixed Integer Linear programming (MLP) (Klitgord and Segre, 2010) (Supplementary Table 10). To avoid LP artifacts, upper bounds for the additional media compounds were constrained to 100 mmol gDW-1h-1. A fraction of the models were subsequently excluded and the remaining models that were able to have non-zero biomass flux were used for gene-essentiality predictions (Supplementary Table 8). Simulations for the *E.coli* K-12 MG1655 model (iJR1366) were carried out as described in the original publication (Orth *et al*, 2011) under glucose minimal medium conditions. Simulations for *B. subtilis subsp. subtilis str. 168* (Seed224308.1) were performed under rich medium conditions (Oh *et al*, 2007).

To simulate the effects of gene knockouts, all genes were knocked out one at a time and maximal growth was computed. Gene deletions resulting in close to zero growth predictions ($<10^{-7}$) were considered as computationally essential. Experimental data for *E. coli* gene essentiality was obtained from PEC database (http://www.shigen.nig.ac.jp/ecoli/pecplus/index.jsp) (Kato and Hashimoto, 2007). Gene-essentiality data for *B. subtilis subsp. subtilis str. 168, C. jejuni subsp. Jejuni NCTC 11168* and *H. pylori J99* were obtained from genome-scale knockout studies (Chalker *et al*, 2001; Oh *et al*, 2007; Stahl and Stintzi, 2011).

Heterologous expression of candidate genes

For EC 2.6.1.14, protein Q8DTM1 (STRING id) was PCR amplified from *Streptococcus* mutans (DSM 20523) gDNA. For EC 2.6.1.38, protein Q8R5Q4 (STRING id) was PCR amplified from *Thermoanaerobacter tengcongensis* (DSM 15242) gDNA (see Supplemental Methods for PCR primers and more details). The PCR-amplified genes were cloned into pET22 modified for the purpose of ligation-independent cloning (LIC). The modified expression vector was transformed into *E. coli* BL21 DE3. Isopropyl beta-D-thiogalactopyranoside (IPTG) was added to induce protein production, and the cells were further grown at 20 °C overnight. After centrifugation, cells were washed and suspended in lysis buffer and sonicated using an ultrasonic processor. After centrifugation, the supernatant was loaded onto an 1-ml HisTrap FF column (GE Healthcare) and the protein was eluted with the lysis buffer containing 250 mM imidazole. Buffer exchange was performed using a HiPrep 26/10 Desalting column (GE Healthcare) with a mobile phase composed of 50 mM Tris/HCl, pH 8.0; 50 mM NaCl; 10% glycerol; and 1 mM DTT. The protein for EC 2.6.1.38 was further purified by ion exchange using a MonoQ 5/50 GL column (GE Healthcare). The protein was eluted with a NaCl gradient ranging from 50 mM to 1 M over 100 column volumes. The purified protein was stored at -80 °C. The samples were analysed by SDS–PAGE using the Invitrogen NuPAGE system. More detailed information is available in Supplementary Methods.

Enzymatic assays

For EC 2.6.1.14, 3.5 μg of the candidate protein was incubated with 5 mM L-asparagine, 20 mM 2-oxoglutarate and 10 μM PLP in 50 mM Tris/HCl pH 9.0, 25 mM KCl. The products of the reaction (L-glutamate and oxoaloacetamid) were detected using high-resolution LC/MS/MS

(LTQ-Orbitrap, Thermo Scientific). Oxaloacetamid was also detected by UV spectrophotometry at 290 nm (Cooper, 1977).

For EC 2.6.1.38, 3 μg of the candidate protein was incubated with 5 mM L-histidine, 10 mM 2-oxoglutarate and 5 μM PLP in 50 mM Tris/HCl pH 8.0. The products of the reaction (L-glutamate and (imidazol-5-yl)-pyruvate) were detected using LC/MS/MS. (Imidazol-5-yl)-pyruvate could also be detected by UV spectrophotometry at 284 nm (Hacking and Hassall, 1975).

For EC 2.1.1.68, 3 μg of candidate protein was incubated for 30 min at 30°C in 20 mM Tris/HCl pH 8.0 containing 10 mM MgCl2 and 40 mM sodium ascorbate, in the presence of 5 mM caffeate and 5 mM *S*-adenosylmethionine (SAM), and the reaction was monitored using LC/MS as described in Supplementary Materials.

For EC 2.3.1.32, 3 μg of candidate protein was incubated for 30 min at 30°C in 10 mM Tris/HCl pH 8.0 in the presence of 5 mM L-lysine and 10 mM acetyl phosphate, and the reaction was monitored using LC/MS as described in Supplementary Materials.

For EC 2.1.1.19, 3 μg of candidate protein was incubated for 60 min at 30°C in 20 mM Tris/HCl pH 8.0 containing 1 mM DTT and 30 mM sodium ascorbate in the presence of 2 mM tetrahydrofolate and 2 mM trimethylsulphonium, and the reaction was monitored using LC/MS as described in Supplementary Materials.

For EC 2.7.1.28, 3 μg of candidate protein was incubated for 30 min at 30°C in 10 mM Tris/HCl pH 8.0 containing 10 mM MgCl2 in the presence of 2 mM D-glyceraldehyde and 3 mM ATP, and the reaction was monitored using LC/MS as described in Supplementary Materials.

Supplementary information

Supplementary information is available at the *Molecular Systems Biology* website (www.nature.com/msb).

Acknowledgements

We are grateful to Eoghan Harrington and members of the Bork group at EMBL for helpful discussions and assistance. We also thank Sergej Andrejev for the help in implementing FCA. We wish to thank Jean-Louis Petit for excellent technical assistance and Sabine Tricot for her expertize in LC/MS. We thank also Véronique de Berardinis for giving us access to the Genoscope bacterial strain collection. The research leading to these results has received funding from the Natural Sciences and Engineering Research Council of Canada (NSERC), CEA, CNRS, the University of Evry and EMBL. We are thankful to Y. Yuan and EMBL IT service for the technical and computational support

Author contributions: PB designed and supervised this study. TY conducted all bioinformatics analyses. TY and ASW analysed the data. PB, TY and ASW wrote the manuscript. JR contributed to the design of this study. KRP and AZ contributed to the metabolic modelling. NP, AP and MS performed the experiments. JW gave technical support and conceptual advice.

Conflict of interest

The authors declare that they have no conflict of interest.

References

Albermann C, Distler J, Piepersberg W (2000) Preparative synthesis of GDP-beta-L-fucose by recombinant enzymes from enterobacterial sources. *Glycobiology* **10:** 875–881

Arumugam M, Harrington ED, Foerstner KU, Raes J, Bork P (2010) SmashCommunity: a metagenomic annotation and analysis tool. *Bioinformatics* **26:** 2977–2978

Arumugam M, Raes J, Pelletier E, Le Paslier D, Yamada T, Mende DR, Fernandes GR, Tap J, Bruls T, Batto JM, Bertalan M, Borruel N, Casellas F, Fernandez L, Gautier L, Hansen T, Hattori M, Hayashi T, Kleerebezem M, Kurokawa K *et al* (2011) Enterotypes of the human gut microbiome. *Nature* **473:** 174–180

Atsumi S, Hanai T, Liao JC (2008) Non-fermentative pathways for synthesis of branched-chain higher alcohols as biofuels. *Nature* **451:** 86–89

Burgard AP, Nikolaev EV, Schilling CH, Maranas CD (2004) Flux coupling analysis of genome-scale metabolic network reconstructions. *Genome Res* **14:** 301–312

Chalker AF, Minehart HW, Hughes NJ, Koretke KK, Lonetto MA, Brinkman KK, Warren PV, Lupas A, Stanhope MJ, Brown JR, Hoffman PS (2001) Systematic identification of selective essential genes in Helicobacter pylori by genome prioritization and allelic replacement mutagenesis. *J Bacteriol* **183:** 1259–1268

Chen L, Vitkup D (2006) Predicting genes for orphan metabolic activities using phylogenetic profiles. *Genome Biol* **7:** R17

Cooper AJ (1977) Asparagine transaminase from rat liver. *J Biol Chem* **252:** 2032–2038

Costanzo M, Baryshnikova A, Bellay J, Kim Y, Spear ED, Sevier CS, Ding H, Koh JL, Toufighi K, Mostafavi S, Prinz St J, Onge RP, VanderSluis B, Makhnevych T, Vizeacoumar FJ, Alizadeh S, Bahr S, Brost RL, Chen Y, Cokol M *et al* (2010) The genetic landscape of a cell. *Science* **327:** 425–431

Dandekar T, Snel B, Huynen M, Bork P (1998) Conservation of gene order: a fingerprint of proteins that physically interact. *Trends Biochem Sci* **23:** 324–328

Eddy SR (2009) A new generation of homology search tools based on probabilistic inference. *Genome Inform* **23:** 205–211

Enright AJ, Iliopoulos I, Kyrpides NC, Ouzounis CA (1999) Protein interaction maps for complete genomes based on gene fusion events. *Nature* **402:** 86–90

Enright AJ, Van Dongen S, Ouzounis CA (2002) An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res* **30:** 1575–1584

Finn RD, Mistry J, Tate J, Coggill P, Heger A, Pollington JE, Gavin OL, Gunasekaran P, Ceric G, Forslund K, Holm L, Sonnhammer EL, Eddy SR, Bateman A (2010) The Pfam protein families database. *Nucleic Acids Res* **38:** D211–D222

Garcia Martin H, Ivanova N, Kunin V, Warnecke F, Barry KW, McHardy AC, Yeates C, He S, Salamov AA, Szeto E, Dalin E, Putnam NH, Shapiro HJ, Pangilinan JL, Rigoutsos I, Kyrpides NC, Blackall LL, McMahon KD, Hugenholtz P (2006) Metagenomic analysis of two enhanced biological phosphorus removal (EBPR) sludge communities. *Nat Biotechnol* **24:** 1263–1269

Gelfand MS, Novichkov PS, Novichkova ES, Mironov AA (2000) Comparative analysis of regulatory patterns in bacterial genomes. *Brief Bioinform* **1:** 357–371

Gill SR, Pop M, Deboy RT, Eckburg PB, Turnbaugh PJ, Samuel BS, Gordon JI, Relman DA, Fraser-Liggett CM, Nelson KE (2006) Metagenomic analysis of the human distal gut microbiome. *Science* **312:** 1355–1359

Green ML, Karp PD (2004) A Bayesian method for identifying missing enzymes in predicted metabolic pathway databases. *BMC Bioinformatics* **5:** 76

Hacking AJ, Hassall H (1975) The purification and properties of L-histidine–2-oxoglutarate aminotransferase from Pseudomonas testosteroni. *Biochem J* **147:** 327–334

Halliwell CM, Morgan G, Ou CP, Cass AE (2001) Introduction of a (poly)histidine tag in L-lactate dehydrogenase produces a mixture of active and inactive molecules. *Anal Biochem* **295:** 257–261

Hanson AD, Pribat A, Waller JC, de Crecy-Lagard V (2010) 'Unknown' proteins and 'orphan' enzymes: the missing half of the engineering parts list–and how to find it. *Biochem J* **425:** 1–11

Harrington E, Singh A, Doerks T, Letunic I (2007) Quantitative assessment of protein function prediction from metagenomics shotgun sequences. *Proc Natl Acad Sci USA* **104:** 13913–13918

Henry CS, DeJongh M, Best AA, Frybarger PM, Linsay B, Stevens RL (2010) High-throughput generation, optimization and analysis of genome-scale metabolic models. *Nat Biotechnol* **28:** 977–982

Huynen MA, Snel B, von Mering C, Bork P (2003) Function prediction and protein networks. *Curr Opin Cell Biol* **15:** 191–198

Kadas J, Boross P, Weber IT, Bagossi P, Matuz K, Tozser J (2008) C-terminal residues of mature human T-lymphotropic virus type 1 protease are critical for dimerization and catalytic activity. *Biochem J* **416:** 357–364

Kanehisa M, Araki M, Goto S, Hattori M, Hirakawa M, Itoh M, Katayama T, Kawashima S, Okuda S, Tokimatsu T, Yamanishi Y (2008) KEGG for linking genomes to life and the environment. *Nucleic Acids Res* **36:** D480–D484

Kato J, Hashimoto M (2007) Construction of consecutive deletions of the Escherichia coli chromosome. *Mol Syst Biol* **3:** 132

Kharchenko P, Chen L, Freund Y, Vitkup D, Church GM (2006) Identifying metabolic enzymes with multiple types of association evidence. *BMC Bioinformatics* **7:** 177

Klitgord N, Segre D (2010) Environments that induce synthetic microbial ecosystems. *PLoS Comput Biol* **6:** e1001002

Kurokawa K, Itoh T, Kuwahara T, Oshima K, Toh H, Toyoda A, Takami H, Morita H, Sharma VK, Srivastava TP, Taylor TD, Noguchi H, Mori H, Ogura Y, Ehrlich DS, Itoh K, Takagi T, Sakaki Y, Hayashi T, Hattori M (2007) Comparative metagenomics revealed commonly enriched gene sets in human gut microbiomes. *DNA Res* **14:** 169–181

Lespinet O, Labedan B (2005) Orphan enzymes?. *Science* **307:** 42

Letunic I, Doerks T, Bork P (2012) SMART 7: recent updates to the protein domain annotation resource. *Nucleic Acids Res* **40:** D302–D305

Nichols RJ, Sen S, Choo YJ, Beltrao P, Zietek M, Chaba R, Lee S, Kazmierczak KM, Lee KJ, Wong A, Shales M, Lovett S, Winkler ME, Krogan NJ, Typas A, Gross CA (2011) Phenotypic landscape of a bacterial cell. *Cell* **144:** 143–156

Nobeli I, Favia AD, Thornton JM (2009) Protein promiscuity and its implications for biotechnology. *Nat Biotechnol* **27:** 157–167

Oberhardt MA, Palsson BO, Papin JA (2009) Applications of genome-scale metabolic reconstructions. *Mol Syst Biol* **5:** 320

Oh YK, Palsson BO, Park SM, Schilling CH, Mahadevan R (2007) Genome-scale reconstruction of metabolic network in Bacillus subtilis based on high-throughput phenotyping and gene essentiality data. *J Biol Chem* **282:** 28791–28799

Orth JD, Conrad TM, Na J, Lerman JA, Nam H, Feist AM, Palsson BO (2011) A comprehensive genome-scale reconstruction of Escherichia coli metabolism–2011. *Mol Syst Biol* **7:** 535

Osterman A, Overbeek R (2003) Missing genes in metabolic pathways: a comparative genomics approach. *Curr Opin Chem Biol* **7:** 238–251

Overbeek R, Begley T, Butler RM, Choudhuri JV, Chuang HY, Cohoon M, de Crecy-Lagard V, Diaz N, Disz T, Edwards R, Fonstein M, Frank ED, Gerdes S, Glass EM, Goesmann A, Hanson A, Iwata-Reuyl D, Jensen R, Jamshidi N, Krause L *et al* (2005) The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. *Nucleic Acids Res* **33:** 5691–5702

Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO (1999) Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc Natl Acad Sci USA* **96:** 4285–4288

Pouliot Y, Karp PD (2007) A survey of orphan enzyme activities. *BMC Bioinformatics* **8:** 244

Qin J, Li R, Raes J, Arumugam M, Burgdorf KS, Manichanh C, Nielsen T, Pons N, Levenez F, Yamada T, Mende DR, Li J, Xu J, Li S, Li D, Cao J, Wang B, Liang H, Zheng H, Xie Y *et al* (2010) A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* **464:** 59–65

Regueira TB, Kildegaard KR, Hansen BG, Mortensen UH, Hertweck C, Nielsen J (2011) Molecular basis for mycophenolic acid biosynthesis in Penicillium brevicompactum. *Appl Environ Microbiol* **77:** 3035–3043

Rusch DB, Halpern AL, Sutton G, Heidelberg KB, Williamson S, Yooseph S, Wu D, Eisen JA, Hoffman JM, Remington K, Beeson K, Tran B, Smith H, Baden-Tillson H, Stewart C, Thorpe J, Freeman J, Andrews-Pfannkoch C, Venter JE, Li K *et al* (2007) The Sorcerer II Global Ocean Sampling expedition: northwest Atlantic through eastern tropical Pacific. *PLoS Biol* **5:** e77

Said HM (2009) Cell and molecular aspects of human intestinal biotin absorption. *J Nutr* **139:** 158–162

Stahl M, Stintzi A (2011) Identification of essential genes in C. jejuni genome highlights hyper-variable plasticity regions. *Funct Integr Genomics* **11:** 241–257

Tyson GW, Chapman J, Hugenholtz P, Allen EE, Ram RJ, Richardson PM, Solovyev VV, Rubin EM, Rokhsar DS, Banfield JF (2004) Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* **428:** 37–43

Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, Eisen JA, Wu D, Paulsen I, Nelson KE, Nelson W, Fouts DE, Levy S, Knap AH, Lomas MW, Nealson K, White O, Peterson J, Hoffman J, Parsons R, Baden-Tillson H *et al* (2004) Environmental genome shotgun sequencing of the Sargasso Sea. *Science* **304:** 66–74

von Mering C, Jensen LJ, Kuhn M, Chaffron S, Doerks T, Kruger B, Snel B, Bork P (2007) STRING 7–recent developments in the integration and prediction of protein interactions. *Nucleic Acids Res* **35:** D358–362

Yamada T, Kanehisa M, Goto S (2006) Extraction of phylogenetic network modules from the metabolic network. *BMC Bioinformatics* **7:** 130