Genome evolution reveals biochemical networks and functional modules

Christian von Mering, Evgeny M. Zdobnov, Sophia Tsoka, Francesca D. Ciccarelli,
Jose B. Pereira-Leal, Christos A. Ouzounis, Peer Bork

Supporting Online Material

1. Detailed Procedures and Data Sources

1.1. Input Data

Functional associations between orthologous groups of proteins were predicted using STRING ([1, 2], version 3.0). The predictions in STRING are based on benchmarked, quantitative integration of three different genomic context methods: common phylogenetic distribution, conserved gene neighborhood, and gene fusions. The predictions cover 260,023 proteins in 89 genomes, with a total of 1,908,210 binary links. Orthology information in STRING is based on the database 'Clusters of Orthologous Groups', COGs ([3], as of October 2002), which we extended to cover 89 species. The extension was by a two-step procedure: Wherever possible, genes from novel species were assigned to existing COGs, if at least three of their closest homologs within the COG database were found in the same COG. Remaining unassigned genes were grouped, where possible, into novel orthologous groups, by identifying and grouping triangles of reciprocal best matches in all-against-all Smith-Waterman similarity searches. These new groups were termed 'non-supervised orthologous groups' NOGs ([1], see the STRING website for details on the procedure [2]).

For the present study, only orthologous groups containing at least one protein from the respective target organism (*Escherichia coli* K12 or *Haemophilus influenzae*) were considered. Note that some orthologous groups contain several proteins per species, thereby limiting the resolution of genomic context analysis. For most of these cases, we accepted this limited resolution and assigned any predicted link concerning the group to be true for all of the proteins in the group. However, to limit errors introduced by excessively large groups, we completely excluded from the input data those orthologous groups which had on average more than 4 distinct genes per species (not counting quasi identical in-paralogs resulting from very recent duplications). This excluded 93 groups for *E. coli* and 64 groups for *H. influenzae.*

For *E. coli*, we additionally processed a limited number of large groups manually – essentially to enhance orthology resolution by splitting orthologous groups into two or more smaller groups. The splitting was done prior to the analysis presented in this study, and was guided solely by inherent sequence information, not by functional annotation. In total, we split 28 groups having between 2 and 12 proteins in *E. coli*; these groups were selected based on the appearance of phylogenetic trees, the availability of operon information, and their relevance to metabolism.

1.2. Clustering

The orthologous groups and their functional associations predicted in STRING define a network with undirected, weighted edges. We identified functional modules in this network, by the use of unsupervised hierarchical clustering [4]. Three algorithmically distinct approaches were applied - namely mean (UPGMA) clustering and single-linkage clustering, both implemented in the OC package [5], as well as Markov clustering [6, 7]. For the exploration of

parameter space, clustering was performed at several different cut-off values (for single-linkage and mean clustering) or inflation values (for Markov clustering).

1.3. Metabolic Pathways

As a trusted reference for functional modularity, we chose small-molecule metabolism in *E. coli*, which can be separated into metabolic 'pathways' as described in the database EcoCyc [8, 9], (version 6.5). In *E. coli* K12, this reference set consists of 583 proteins (enzymes or enzyme subunits) belonging to 144 pathways. The pathways are somewhat redundant: 187 proteins map to more than one pathway. This redundancy is partly due to historical reasons and/or for increased clarity (pathways often begin or end at 'important' compounds), but is also often due to enzymes actually having multiple functions, i.e. being annotated with distinct reactions in distinct pathways. Note that since metabolism is a single, large web of interconnected metabolites and enzymes, the separation into pathways is necessarily somewhat subjective.

1.4. Benchmarking

The above metabolic pathways were used to benchmark the functional modules we predicted here through the clustering of genomic context associations. The predicted modules cover many cellular functions, not only metabolism. However, because we focused on metabolism for benchmarking, we removed from the predicted modules prior to benchmarking all proteins not present in the EcoCyc pathways, and finally considered only modules remaining with at least two annotated enzymes (or enzyme subunits).

For each predicted module, the best-matching pathway was selected for comparison and the following measures were computed: specificity - defined as Tp/(Tp+Fp), sensitivity - Tp/(Tp+Fn) and overlap function - Tp/(Tp+Fp+Fn), where Tp denotes true positives, Fp – false positives and Fn – false negatives. The counting was done on the level of proteins, i.e. orthologous groups containing several *E.coli* proteins generated multiple counts. Enzyme subunits were counted as separate entities. The choice as to which pathway a predicted module should be compared to ('best matching pathway'), was by selecting the pathway with maximal overlap function.

We also computed the number of enzymes that failed to cluster with any other enzyme (singletons). A complementary measure, 'total coverage', was defined as the fraction of enzymes found clustered in predicted modules together with at least one other enzyme.

1.5. Random Background

We compared our benchmarking results with random expectations at two different levels of randomization. One very conservative random model was to (i) keep the cluster size distribution as predicted, (ii) keep the number of enzymes within each cluster fixed, and (iii) only swap enzyme identities across those fixed clusters. On average, this lead to a 2.2-fold reduction in observed specificity and to a 2.4-fold reduction in observed overlap.

A more realistic comparison to random expectation is to ask how many modules can be expected, by chance, to consist entirely of enzymes only. Given the known numbers of enzymes and non-enzymes in *E.coli*, we computed this expectation using the hypergeometric distribution (sampling without replacement). When comparing the actual predictions against this expectation, we observed a strong deviation from randomness (especially for the larger modules); in total the predicted clusters differed by more than one order of magnitude from the random expectation *(Table S1)*.

size	clusters	purely enzymatic	expected	observed	ratio
1	466	50	0.179054	0.107296	0.599239
2	206	19	0.032015	0.092233	2.880914
3	117	8	0.005716	0.068376	11.9616
4	72	4	0.001019	0.055556	54.50908
5	54	1	0.000181	0.018519	102.0515
6	42	1	3.23E-05	0.02381	737.9945
7	21	1	5.73E-06	0.047619	8313.63
8	24	1	1.02E-06	0.041667	41032.31
9	18	0	1.8E-07	0	0
10	10	0	3.18E-08	0	0

Table S1: Deviations from random expectation
Using the hypergeometric distribution, and the known frequencies of enzymes and non-enzymes in *E. coli* (according to EcoCyc small molecule metabolism), the actual number of purely enzymatic modules was compared to what can be expected by chance. For all module sizes >= 3, the observed deviation from random is larger than 10-fold.

1.6. Functional Categories

We systematically analyzed the functional composition of the predicted modules using Gene Ontology (GO) categories [10], as assigned to proteins in *E. coli* [11]. We focused on terms of the subcategory 'cellular processes', and reduced the number of terms by grouping related terms as follows. First, we checked the distribution of all *E. coli* proteins over the whole GO hierarchy (which is a directed acyclic graph), by traversing from assigned leaf terms through all possible paths up to the root term. Throughout this procedure, we marked all nodes visited at least 100 times as terms of sufficiently high generality. For any protein of interest, we then traversed from its assigned terms up the hierarchy, and stopped at the first encountered 'marked' node, thereby objectively grouping functional assignments at a medium level of detail.

In Figure 4 of the manuscript, for clarity, an even higher-level assignment of functional categories was needed. Here, Gene Ontology annotations proved impractical, and we instead chose the summary categories defined for orthologous groups at the COG-website [12].

1.7. Analysis of False Positives

We manually analyzed all 40 cases where a predicted metabolic module could not be fully mapped to a single pathway (i.e. the pathway specificity was less than one). We assigned each case to one of four categories (some modules had more than one false positive protein; if these fell into different categories, we assigned the module fractionally). The first and largest category contained cases in which two or more proteins were assigned to different pathways in EcoCyc, but to the same orthologous group in the COG database [3]. In these cases, our comparative genomics methods currently do not have sufficient resolution and have to assign both proteins to one module. A second category contained cases where the two pathways in question where either annotated as overlapping in EcoCyc, or where they were connected through a common metabolite (excluding trivial metabolites such as water or ATP). A third category contained cases where pathways had previously been suggested to be connected in recent literature. This could mean a genetic or experimental connection, and was determined by searching for reports dealing with both proteins, or for cases where both proteins had the same annotated reference in SwissProt (excluding whole genome sequencing papers). The fourth category finally contained all the remaining cases, i.e. for which we did not find any obvious explanation. This last category constitutes novel predicted inter-pathway links, but some may obviously also be false positives.

2. Supplementary Data

2.1. Detailed Benchmarking Results

As mentioned above, we performed unsupervised hierarchical clustering on the network of genomic context associations, using three different algorithms and a wide range of clustering parameters. The resulting clusters were benchmarked against metabolic pathways as described (section 1.4). The following is a summary of the benchmarking results for the individual clustering results (see the accompanying website for the complete list of annotated clusters. The address is http://www.bork.embl-heidelberg.de/Docu/String_modules/index.html).

UPGMA mean clustering

clustering cutoff	Total coverage	Average sensitivity	average specificity	average overlap	predicted clusters	singleton clusters	valid clusters	average cluster size
100	0.8336	0.5112	0.7373	0.3948	170	68	102	4.8
150	0.8130	0.4907	0.7656	0.3881	188	80	108	4.4
200	0.7839	0.4966	0.8136	0.4214	209	97	112	4.1
250	0.7650	0.5000	0.8252	0.4264	223	108	115	3.9
300	0.7496	0.4933	0.8328	0.4238	235	118	117	3.7
350	0.7444	0.4887	0.8411	0.4257	240	121	119	3.6
400	0.7358	0.4931	0.8425	0.4292	245	126	119	3.6
450	0.7256	0.4934	0.8433	0.4291	250	132	118	3.6
500	0.7033	0.5007	0.8439	0.4292	260	145	115	3.6
550	0.6947	0.4928	0.8524	0.4257	266	150	116	3.5
600	0.6844	0.4892	0.8553	0.4241	272	156	116	3.4
650	0.6707	0.4872	0.8633	0.4234	280	164	116	3.4
700	0.6621	0.4820	0.8620	0.4196	285	169	116	3.3

Single linkage clustering

clustering cutoff	total coverage	Average sensitivity	average specificity	average overlap	predicted clusters	singleton clusters	valid clusters	average cluster size
100	0.9520	1.0000	0.0577	0.0577	1	0	1	555.0
150	0.9520	1.0000	0.0577	0.0577	1	0	1	555.0
200	0.9485	0.7000	0.5290	0.2290	4	2	2	276.5
250	0.9417	0.7000	0.5293	0.2293	8	6	2	274.5
300	0.9314	0.6339	0.7644	0.4061	16	12	4	135.7
350	0.9160	0.5199	0.8398	0.4129	30	21	9	59.3
400	0.9022	0.4739	0.8244	0.3730	41	29	12	43.8
450	0.8851	0.4937	0.8226	0.3951	55	39	16	32.2
500	0.8473	0.5127	0.8362	0.4147	84	61	23	21.5
550	0.8250	0.4981	0.8289	0.4155	108	74	34	14.1
600	0.8010	0.5106	0.8331	0.4310	131	88	43	10.9
650	0.7753	0.4880	0.8363	0.4062	150	103	47	9.6
700	0.7358	0.4659	0.8478	0.3997	186	126	60	7.1

Markov clustering

inflation parameter	total coverage	Average sensitivity	average specificity	average overlap	predicted clusters	singleton clusters	valid clusters	average cluster size
1.5	0.9485	0.4000	0.6095	0.1510	6	2	4	138.2
1.7	0.9417	0.4950	0.6518	0.3300	17	6	11	49.9
2.0	0.8559	0.5129	0.7120	0.3658	118	57	61	8.2
2.3	0.6672	0.4674	0.8162	0.3938	271	170	101	3.9
2.5	0.6000	0.4597	0.8363	0.3969	309	204	105	3.3
2.7	0.5506	0.4497	0.8471	0.3933	339	239	100	3.2
3.0	0.5009	0.4287	0.8573	0.3859	367	270	97	3.0
3.2	0.4889	0.4302	0.8664	0.3888	376	279	97	2.9
3.5	0.4700	0.4144	0.8760	0.3742	388	291	97	2.8

Table S2: Benchmarking results.
Comparison of predicted functional modules against pathway definitions in small molecule metabolism.

2.2. Biological Discovery (I): *pathway extensions proven correct by recent literature.*

Genomic context associations are objective, unbiased and independent of prior knowledge. This enables the discovery of enzymes which are still lacking in the current description of a specific pathway. As proof of principle, we list below some example cases where an enzyme is still annotated as missing/uncharacterized in EcoCyc (version 6.5), and for which our predicted modules predict suitable candidates which are confirmed by recent literature. Note that metabolism in *E. coli* is one of the best studied cellular systems. We do still find pathway extensions in *E .coli* today, but we expect that genomic context methods should uncover even more novelty in the majority of other organisms/systems, which are less well-studied.

a) alpha-ribazole-5'-phosphatase

This enzyme is thought to be part of cobalamin biosynthesis, but according to EcoCyc it is not yet known in *E.coli*. In our predicted clusters, four known cobalamin biosynthesis proteins (CobU, CobT, CobS and BtuR) are predicted together with the protein P52086, a candidate for alpha-ribazole-5'-phosphatase (the cluster also contains an apparent false-positive, GpmB). P52086 is very similar to an enzyme characterized in *Salmonella typhimurium* (COBC_SALTY, to which P52086 shows 74% sequence identity over the full length of 198 amino acids). COBC_SALTY has recently been shown *in vitro* to indeed catalyze a step in cobalamin biosynthesis [13], effectively confirming the genomic context prediction.

b) 1-hydroxy-2-methyl-2-(E)-butenyl 4-diphosphate synthase

c) 1-hydroxy-2-methyl-2-(E)-butenyl 4-diphosphate reductase

In EcoCyc, these two enzymes catalyzing the last steps of the nonmevalonate isoprenoid biosynthesis pathway are annotated as unknown. However, using genomic context analysis and genetics, both enzymes have recently been identified and characterized experimentally [14-16]. In the genomic context modules presented here, the enzymes are also correctly annotated, being grouped together in clusters with other, known enzymes of the pathway (Figure S1). Note that the two enzymes show no sequence similarity to each other, so they constitute two separate predictions.

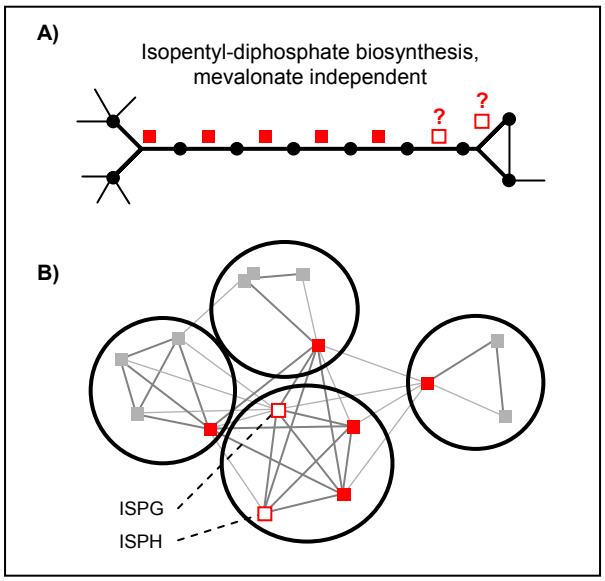

Figure S1: a confirmed pathway extension.
A) schematic view of the nonmevalonate isoprenoid biosynthesis pathway in *E. coli*. Black dots represent metabolites, red squares denote enzymes. Open squares are enzymes listed as not yet known in EcoCyc.
B) subset of the genomic context network. Black circles indicate the subdivision into functional modules, as defined by the unsupervised clustering. The labelled open squares are recently confirmed additions to the pathway. Note that in this case, the type of genomic context evidence which has contributed most is common phylogenetic distribution (gene co-occurrence).

d) Nitrate/nitrite transporters NarU / NarK

This is an example for a pathway extension that does *not* consist of enzymes, but of accessory proteins which are functionally associated to a metabolic pathway. The two proteins NarU and NarK are both known to function as transporters in nitrate uptake and nitrite extrusion [17]. In our predicted modules, they are correctly associated with a total of 10 subunits of the two nitrate reductase complexes NarGHJI and NarZYWV.

2.3. In E Biological Discovery (II): *predicted pathway extensions.*

a) a novel enzyme predicted to be associated with ubiquinone biosynthesis.

We observed several uncharacterized proteins clustering together with the enzyme UbiC, which catalyzes the first step in ubiquinone biosynthesis. Of these, one familiy is particularly strongly linked: it has in the COG database a phylogenetic distribution exactly matching that of UbiC, and is a direct genomic neighbor of UbiC in one of the genomes studied (*Ralstonia solanacearum*). Together, these two conceptually independent types of evidence constitute a rather significant genomic context association, and predict that at least one member of this uncharacterized family might be functionally associated with the metabolism of ubiquinone. The family has two representatives in *E. coli*, YqiA and YcfP, which are both annotated as 'hypothetical proteins' in the SwissProt knowledgebase. We performed fold-prediction for one of them (YqiA), employing several independent algorithms through a metaserver [18]. The results are summarized in Figure S2: YqiA is predicted to have significant structural similarity to known enzymes of the alpha/beta-hydrolase superfamily, with conserved residues shown to form a catalytic triad. While this finding strongly suggest that YqiA may indeed have enzymatic activity, the actual function of this enzyme and the nature of its possible association to ubiquinone metabolism are more difficult to predict and require experimental clarification.

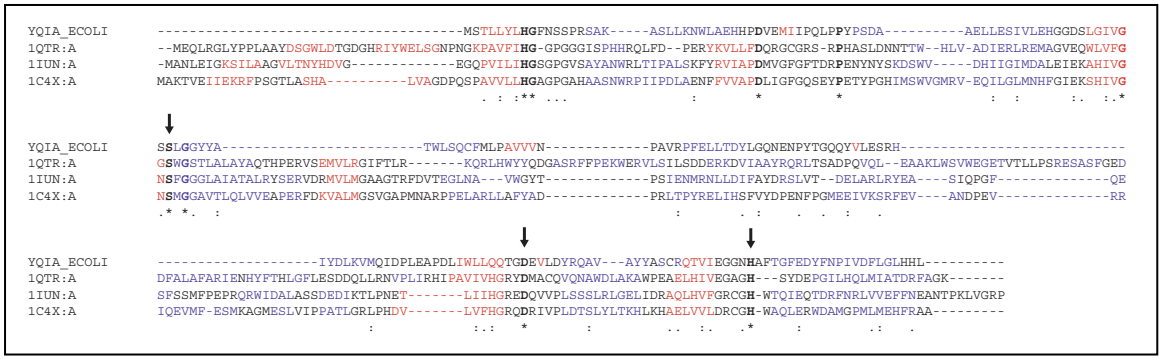

Figure S2: YqiA is a putative alpha/beta hydrolase.
Multiple sequence alignment, showing the similarity of the YqiA protein to alpha/beta hydrolases of known structure. The colors represent secondary structure (blue: helical, red: beta-sheet. For YqiA, the secondary structure is a prediction using the program sam-t99-2d). Arrows denote the conserved catalytic triad [19].

b) a predicted transcriptional regulator of riboflavin biosynthesis.

Pathway extensions predicted by genomic context are not limited to novel enzymes. An example for a typical non-enzymatic extension is contained in the predicted module no. 988 (Figure S3). This module contains five known enzymes of the riboflavin biosynthesis pathway, together with the hypothetical protein YbaD. The riboflavin biosynthesis pathway has one unassigned enzyme (pyrimidine phosphatase), but YbaD does not appear to be a candidate for this missing enzyme, as YbaD does not present any detectable similarity to other phosphatases. YbaD is, however, similar to a protein from *Phosphobacterium phosphoreum* (81% identity). The gene coding for this protein has been proposed to be associated with the *rib* operon in this species and although not experimentally characterized, the presence of a high proportion of basic

aminoacids has led to the suggestion that the protein may be a regulator of gene expression for riboflavin biosynthesis genes [20]. To further support this claim, YbaD possesses an ATP-cone domain [21] and a Zinc-ribbon domain, suggesting a regulatory, DNA-binding function. Based on this, YbaD is proposed to be a regulator of the riboflavin pathway genes in *E. coli*.

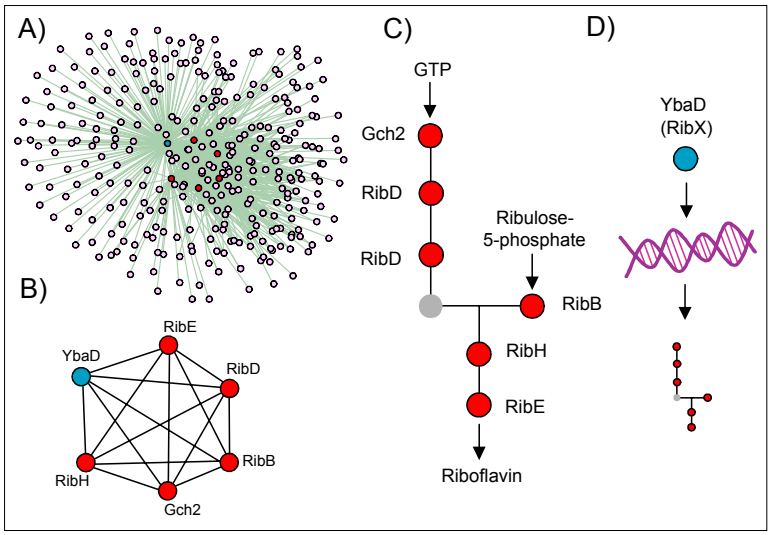

Figure S3: YbaD, a predicted transcription regulator possibly involved in riboflavin biosynthesis
A) the proteins in predicted module no. 988 are shown (colored) with their immediate neighbors (white) in the genomic context network. B) the predicted module and the identity of its proteins. C) EcoCyc representation of the riboflavin biosynthesis pathway. One enzymatic step (pyrimidine phosphatase, shown in grey) is uncharacterized, the enzyme responsible is not yet known. D) suggested function for the module member YbaD. In the model shown here, YbaD is a putative upstream transcriptional regulator (possibly a repressor) of the other pathway proteins.

2.4. Biological Discovery (III): *novel connections between known pathways.*

Some of the functional modules predicted here are grouping together enzymes which are not part of the same pathway in EcoCyc; these modules are thus representing predicted associations *between* pathways. In the benchmarking procedure, all of these were counted as 'false positives', and indeed in at least a number of cases these predictions are due to technical artifacts (mostly insufficient resolution in orthology detection).

In other cases, however, the modules appear to reflect actual functional links between pathways (known and novel links), and thereby reveal a higher-order connectivity among biological processes. An example for a novel link is a predicted connection between the biosynthesis pathway for coenzyme A and the metabolism of nucleotides. The prediction is supported by several independent observations, both within and outside of the predicted modules: First, the coenzyme A biosynthesis enzyme CoaA/B (phosphopantothenoylcysteine synthetase/decarboxylase) is in a predicted module together with dUTPase, being its direct neighbor in 14 distinct genomes. Secondly, CoaA/B has a phylogenetic distribution that closely matches that of several other enzymes involved in nucleotide metabolism (PyrF, GuaA and Pur2), and it matches those closer than it does any enzyme of coenzyme A metabolism. Thirdly, and probably most significant, the gene encoding PanC (pantoate-beta-alanine ligase, also involved in coenzyme A biosynthesis) is found fused to the gene encoding cytidylate kinase in the *Nostoc sp.* genome, specifying a single polypeptide that can simultaneously function in both nucleotide metabolism and coenzyme A biosynthesis.

Together, the above observations make a strong case for a functional connection between the two pathways. The precise nature of the predicted link remains unknown, but it should be noted that the base adenine is a structural component of coenzyme A, and that the reaction catalyzed by

CoaA/B requires the nucleotide CTP as an unusual energy source, pointing to possible connections between the pathways at the substrate level. Alternatively, one of the enzymes in coenzyme A biosynthesis might have a second, additional function in nucleotide metabolism. In this regard it is tempting to search for structural and mechanistic similarities among reactions in both pathways (see Figure S4 for an example). Intriguing support for a functional link also comes from experimental data: the initial phenotype reported for mutants in the gene encoding CoaA/B was a defect in DNA synthesis [22]; and in an unrelated context, CoaA/B was recently found unexpectedly to co-purify with one of its partners predicted here, dUTPase [23].

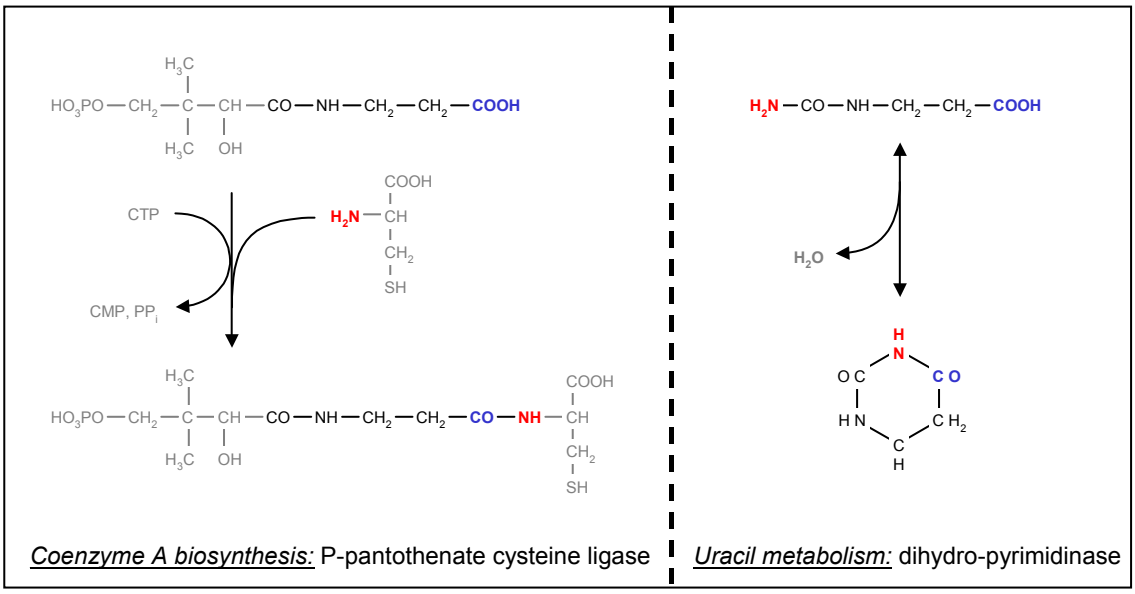

Figure S4: mechanistically and structurally related reactions in two different pathways.
The two reactions shown here both involve the formation/breaking of an amide bond, and there is some structural similarity in the substrates (marked in black/color). Additionally, both reactions are only a few steps away from the metabolite beta-alanine, hinting at the possibility of shared regulation or enzyme re-use. Both observations may offer an explanation for the predicted functional association between the two pathways.

In other cases, the inter-pathway connections are extending the putative functions of proteins previously thought to play an exclusive role in a single pathway. An example is phospho-glyceratephosphatase A (PgpA), which is thought to be involved in phospholipid biosynthesis [24]. However, in the functional modules predicted here, it is grouped together with thiamin monophosphate kinase (ThiL), which catalyzes the last step in thiamin biosynthesis. The two genes are immediate neighbors in ten different genomes, forming a strong genomic context association. The exact biochemical nature of this association remains unknown, but it is intriguing to note that PgpA is not essential for phospholipid biosynthesis *in vivo*, as strains of *E. coli* with the gene disrupted can still produce phospholipids [25]. Furthermore, several steps in the thiamine biosynthesis pathway remain poorly characterized, and at least some organisms are thought to possess a phosphatase [26] which dephosphorylates thiamine-monophosphate, the substrate of ThiL. In *E.coli*, such a phosphatase is not known, but could potentially be useful in regulating the uptake and/or intracellular homeostasis of thiamin-derivatives, in analogy to a scheme proposed in eukaryotes [27].

2.5. Biological discovery (IV): *entirely novel functional modules*

We observe a large number of predicted modules consisting largely or entirely of uncharacterized proteins. Each of these represents an opportunity to uncover/characterize a novel

functional process, defined by conserved associations across species. As an example, consider the predicted module 1071 in *E. coli*. It consists of four proteins, all of which are annotated as uncharacterized in SwissProt. For three of them (YeaG, YcgB and YeaH), the genomic context links are quite strong, defining a well conserved and evolutionarily widespread functional unit (Figure S5). The fourth protein (YfbU) is small and only loosely associated (being a direct neighbor in only one genome). Notice that in *E. coli*, only two of the three genes are forming an apparent operon, the third gene is found elsewhere in the genome. This illustrates the importance of using comparative genomics before defining a functional unit. The module receives further support by the phylogenetic occurrence of the genes (they always occur together, in a non-trivial species pattern, Figure S5); this shows the benefit of integrating genomic context methods.

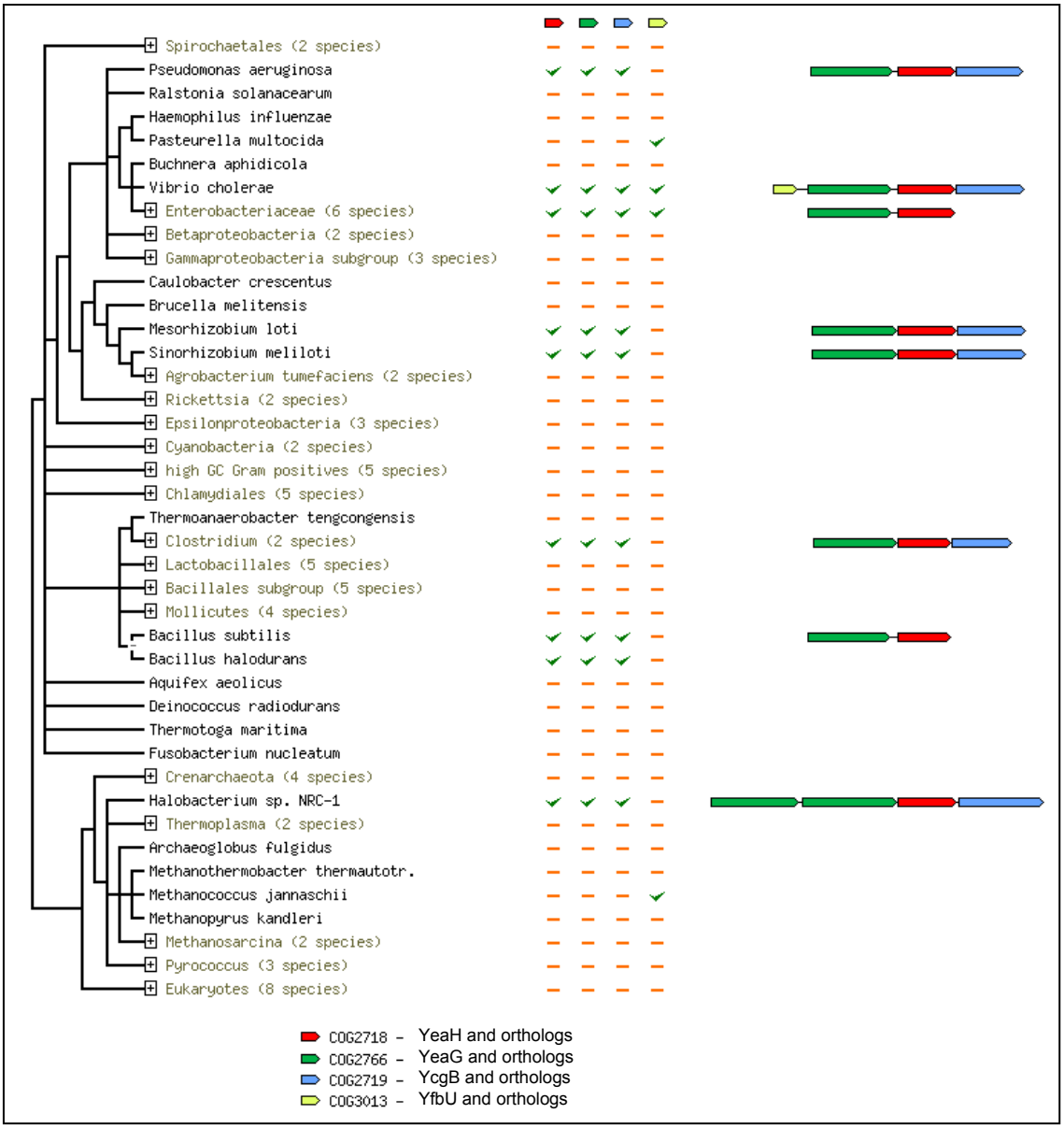

Figure S5: genomic context evidence for a predicted novel functional unit.
The left part of the image represents the 89 species used here. Species which are in exact agreement regarding the evidence shown are collapsed into a single line. The tree topology has been edited manually (to avoid controversial placements) by introducing a number of multiforcating sections). The middle part of the image shows presence or absence of orthologs in the various species, and the right part shows instances where the genes are immediate neighbors on the same strand of the genome, separated by less than 300 bp.

A detailed inspection of the proteins using profile searches and structure prediction [18] reveals that the YeaG protein contains an AAA-ATPase domain, and the YeaH protein contains an Integrin I domain. This combination of domains is known to occur in metal-chelatases [28], and indeed both proteins show remote sequence similarity to magnesium chelatase. Orthologs of YeaG and YcgB in *Bacillus subtilis* are known to cause defects in endospore formation when mutated [29, 30]. This would point to a function of the module in spore-formation. However, the phylogenetic distribution of the module clearly extends beyond spore-forming organisms, indicating that its function in spore-formation must be rather general. One hypothesis is that the module is a metal chelatase of widespread importance (for an unknown metal). This is supported by the presence of the module in metallo-tolerant and salt-loving organisms (*Geobacter metallicreducens, Ralstonia metallidurans, Magnetospirillum magnetotacticum, Bacillus halodurans, Halobacterium sp. NRC-1*).

As an example for an independently confirmed novel functional module, consider the ethanolamine utilization pathway, which is present in a limited number of prokaryotes, including enteric bacteria such as *E. coli* or *Salmonella typhimurium*. This pathway is not yet annotated in EcoCyc, but has recently been described genetically and functionally [31]. Of the 17 described genes, 13 are successfully predicted to be a single functional modules, with only two additional proteins predicted (these appear to be false positives).

2.6. Why three different genomic context methods? Partial complementarity & overlap

Many of the functional modules reported here are supported by evidence from more than one genomic context method. This provides increased confidence in the predictions, and is probably one of the reasons for the observed good pathway specificity. We sought to quantify the overlap between the three methods by assessing how many pathways were covered by what type of genomic context evidence. The results are summarized in Table S3. We observed that conserved gene neighborhood is the method contributing most of the signal, but also that each of the three methods covers some pathways exclusively. This latter observation means that combining the different methods not only increases specificity, but also the achievable total coverage.

	N	F	P
1CMET2-PWY	x	x	-
ACETOACETATE-DEG-PWY	x	x	x
AERESPDON-PWY	x	x	x
ALADEG-PWY	x	x	x
ALANINE-VALINESYN-PWY	x	-	-
ANARESPACC-PWY	x	x	x
ANARESPDON-PWY	x	x	x
ARABCAT-PWY	x	x	x
ARGSYN-PWY	x	x	x
ARO-PWY	x	x	x
ASPASN-PWY	x	x	x
AST-PWY	x	x	x
BIOTIN-SYNTHESIS-PWY	x	x	x
COBALSYN-PWY	x	-	x
COLANSYN-PWY	x	-	-
CYSTSYN-PWY	x	x	x
DAPLYSINESYN-PWY	x	-	-
DARABCAT-PWY	x	-	-
DEOXYRIBONUCMET-PWY	x	x	x
DETOX1-PWY	x	x	x
DRIBOPMET-PWY	x	-	-
DTDPRHAMSYN-PWY	x	x	x
ECASYN-PWY	x	x	x
ENTBACSYN-PWY	x	-	-
FAO-PWY	x	x	x
FASYN-INITIAL-PWY	x	x	x
FOLSYN-PWY	x	x	x
GALACTCAT-PWY	x	x	x
GALACTITOLCAT-PWY	x	x	x
GALACTMETAB-PWY	x	-	-
GALACTUROCAT-PWY	x	x	x
GLCMANNANAUT-PWY	x	x	-
GLUCARDEG-PWY	x	x	x
GLUCARGALACTSUPER-PWY	x	x	x
GLUCONEO-PWY	x	-	x
GLUCOSE1PMETAB-PWY	x	-	-
GLUTATHIONESYN-PWY	-	x	-
GLUTSYN-PWY	x	x	-
GLYCEROLMETAB-PWY	x	x	x
GLYCLEAV-PWY	x	-	x
GLYCOCAT-PWY	x	x	x
GLYCOGENSYNTH-PWY	x	-	-
GLYCOLYSIS	x	x	x
GLYOXYLATE-BYPASS	x	x	-
HCAMHPDEG-PWY	x	x	x
HEMESYN2-PWY	x	-	x
HISTSYN-PWY	x	x	x
HOMOSER-METSYN-PWY	-	x	x
HOMOSER-THRESYN-PWY	x	-	-
HOMOSERSYN-PWY	x	x	x

	N	F	P
IDNCAT-PWY	x	x	x
KDOSYN-PWY	x	-	-
KETOGLUCONMET-PWY	x	x	x
LEUSYN-PWY	-	x	x
LYXMET-PWY	x	-	-
MENAQUINONESYN-PWY	x	-	x
NADPHOS-DEPHOS-PWY	x	x	x
NAGLIPASYN-PWY	-	-	x
NONMEVIPP-PWY	x	x	x
NONOXIPENT-PWY	x	x	x
NUCMET2-PWY	x	x	x
ORNDEG-PWY	x	x	x
OXIDATIVEPENT-PWY	x	x	-
PANTOSYN-PWY	x	x	x
PEPTIDOGLYCANSYN-PWY	x	-	x
PHOSLIPSYN-PWY	x	x	x
POLYAMSYN-PWY	x	-	-
POLYISOPRENSYN-PWY	x	x	x
PPGPPMET-PWY	x	x	x
PROSYN-PWY	x	x	x
PRPP-PWY	x	-	-
PURSYN-PWY	x	x	x
PWY0-42	-	x	x
PWY0-43	-	x	-
PYR-RIBONUCMET-PWY	-	x	-
PYRIDNUCSYN-PWY	x	-	x
PYRIDOXSYN-PWY	-	x	x
PYRIMSYN-PWY	x	x	x
PYRNUCYC-PWY	x	-	x
PYRUVDEHYD-PWY	-	x	-
PYRUVOX-PWY	x	x	x
RHAMCAT-PWY	x	-	-
RIBOKIN-PWY	-	x	-
RIBOSYN2-PWY	x	x	x
SAM-PWY	x	-	-
SERDEG-PWY	x	x	x
SERSYN-PWY	x	x	x
SO4ASSIM-PWY	x	x	-
TCA	x	x	x
THISYN-PWY	x	x	-
THREOCAT-PWY	x	x	x
TREDEGLOW-PWY	x	x	x
TRESYN-PWY	x	x	x
TRPSYN-PWY	x	x	x
TYRSYN	x	x	x
UBISYN-PWY	x	-	x
UDPNAGSYN-PWY	-	-	x
VALSYN-PWY	x	x	x
XYLCAT-PWY	x	-	-

Table S3: Metabolic pathways – coverage by method

The Table shows the pathways covered by a given genomic context method, and also the overlap among the methods. To compute this, three separate genomic context networks were constructed - one for each method. Unsupervised clustering (UPGMA means clustering, cutoff 400) was then applied to each network, and the resulting clusters were mapped to pathways in the same way as was done for the full network. Pathways were counted as covered if they had at least one predicted cluster mapping to them. Pathway identifiers in the Table are as in the EcoCyc database.

(**N**: Conserved Neighborhood, **F**: Gene Fusions, **P**: Common Phylogenetic Distribution).

2.7. The comparative genomics of functional modularity

The approach presented here is not limited to well-characterized organisms, but should work for any organism of interest provided its genome has been fully sequenced. As proof of principle we repeated our analysis for the genome of *Haemophilus influenzae* Rd [32], which, like *E. coli*, is a gamma-proteobacterium, and for which a detailed metabolic reconstruction has been carried out [33]. Projection of the predicted functional network to its genome of 1,711 annotated genes [34] results in 1209 proteins connected through 57,987 links. Despite a considerably smaller number of annotated pathways (89) and annotated enzymes (217), the overall performance is essentially the same as that observed in *E. coli* (Tables S2, S4). Although gene order and operon architecture are generally not conserved between *E.coli* and *H. influenzae* [35], functional modules can be reliably detected and appear to be evolutionarily conserved. In fact, considering only the subset of small molecule metabolism which is found in both organisms, we observe that all *H. influenzae* modules recovered after clustering can be matched to a module in *E. coli*. Remarkably, the average overlap between the small–molecule metabolic enzymes in both modules is 90%. This observation suggests that the predicted modules (and the pathways they match) may represent evolutionary units.

UPGMA mean clustering

clustering cutoff	total coverage	average sensitivity	average specificity	average overlap	predicted clusters	singleton clusters	valid clusters	average cluster size
300	0.7327	0.6626	0.8096	0.5257	101	56	45	3.6
350	0.7235	0.6769	0.8055	0.5336	103	58	45	3.5
400	0.7189	0.6658	0.8018	0.5262	104	59	45	3.5
450	0.7097	0.6614	0.8141	0.5329	107	61	46	3.4
500	0.7005	0.6601	0.8074	0.5225	110	63	47	3.3
600	0.6728	0.6513	0.8529	0.5478	115	69	46	3.2

Table S4: Benchmarking of predicted functional modules in *H. influenzae*.
Comparison of predicted functional modules against pathway definitions in small molecule metabolism. The performance is comparable to what was achieved in *E. coli* (Table S2). The main difference lies in average sensitivity (and, consequently, average overlap) which are higher in *H. influenzae* than in *E. coli*. This could be a consequence of differences in pathways definitions - in *H. influenzae,* pathways contain on average less proteins.

2.8. Unexpected pathway connections in H. influenzae

Similarly to what was described for *E. coli*, apparent false positives in modules predicted for *H. influenzae* can often be functionally linked to other cluster members, providing a good basis for biological discovery. One example is cluster no. 734 (Figure S6), which contains two sets of apparently unrelated proteins: the first set contains the anaerobic formate dehydrogenase complex (FdxG, FdxI and FdxH) together with two proteins that are necessary for formate dehydrogenase activity (FdhE and FdhH). The second set of proteins is involved in selenocysteine processing, and would be classified as false positives (SelA, SelB, SelD). The connection between these two sets of proteins emerges, however, if we consider that the formate dehydrogenase complex requires seleno-cysteine for proper activity [36].

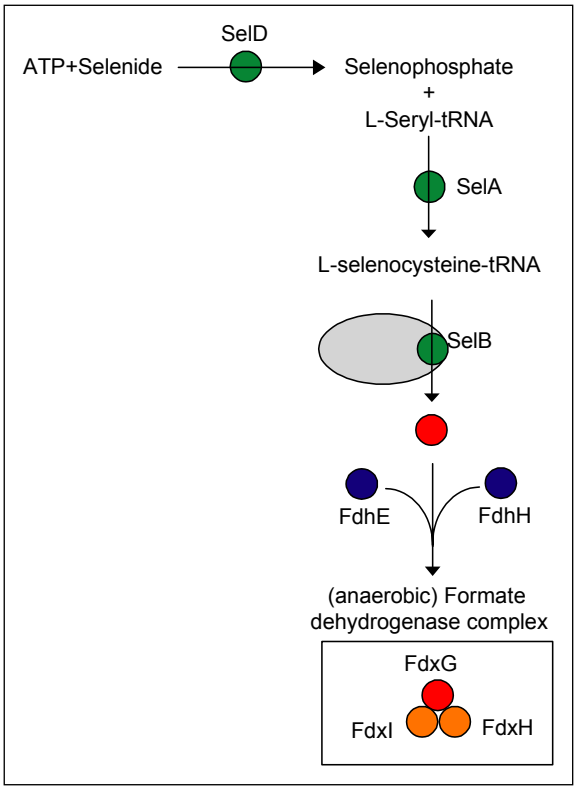

Figure S6: a pathway connection in *H. influenzae*
The figure shows the predicted module 734: this module groups enzymes needed for selenocysteine incorporation, together with enzyme subunits and auxiliary proteins needed for forming the anaerobic formate dehydrogenase complex. SelB is a selenocysteinespecific translation elongation factor which functions in association with the ribosome (the ribosome is symbolized here by a large grey ellipse).

3. Supplementary References

1. von Mering, C., et al., *STRING: a database of predicted functional associations between proteins.* Nucl. Acids Res., 2003. **31**: p. 258-261.
2. *http://www.bork.embl-heidelberg.de/STRING/.*
3. Tatusov, R.L., et al., *The COG database: new developments in phylogenetic classification of proteins from complete genomes.* Nucleic Acids Res, 2001. **29**(1): p. 22-8.
4. Webb, A., *Statistical pattern recognition*. 2nd ed. 2002, Chichester: John Wiley & Sons Ltd.
5. Barton, G., *OC: a cluster analysis program.* http://www.compbio.dundee.ac.uk/manuals/oc/oc_manual.txt.
6. Van Dongen, S., *Graph clustering by flow simulation*. 2000, Center for Mathematics and Computer Science (CWI), Univ. of Amsterdam: Amsterdam.
7. Enright, A.J., S. Van Dongen, and C.A. Ouzounis, *An efficient algorithm for large-scale detection of protein families.* Nucleic Acids Res, 2002. **30**(7): p. 1575-84.
8. Karp, P.D., et al., *The EcoCyc Database.* Nucleic Acids Res, 2002. **30**(1): p. 56-8.
9. Riley, M., *Functions of the gene products of Escherichia coli.* Microbiol. Rev., 1993. **57**: p. 862-952.
10. Ashburner, M., et al., *Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.* Nat Genet, 2000. **25**(1): p. 25-9.
11. Camon, E., et al., *The Gene Ontology Annotation (GOA) Project: Implementation of GO in SWISS-PROT, TrEMBL, and InterPro.* Genome Res, 2003. **13**(4): p. 662-72.
12. *The COG database, functional classification.* http://www.ncbi.nlm.nih.gov/cgi-bin/COG/palox?fun=all.
13. Maggio-Hall, L.A. and J.C. Escalante-Semerena, *In vitro synthesis of the nucleotide loop of cobalamin by Salmonella typhimurium enzymes.* Proc Natl Acad Sci U S A, 1999. **96**(21): p. 11798-803.
14. Cunningham, F.X., Jr., T.P. Lafond, and E. Gantt, *Evidence of a role for LytB in the nonmevalonate pathway of isoprenoid biosynthesis.* J Bacteriol, 2000. **182**(20): p. 5841-8.
15. Rohdich, F., et al., *The deoxyxylulose phosphate pathway of isoprenoid biosynthesis: studies on the mechanisms of the reactions catalyzed by IspG and IspH protein.* Proc Natl Acad Sci U S A, 2003. **100**(4): p. 1586-91.
16. Campos, N., et al., *Identification of gcpE as a novel gene of the 2-C-methyl-D-erythritol 4-phosphate pathway for isoprenoid biosynthesis in Escherichia coli.* FEBS Lett, 2001. **488**(3): p. 170-3.
17. Clegg, S., et al., *The roles of the polytopic membrane proteins NarK, NarU and NirC in Escherichia coli K-12: two nitrate and three nitrite transporters.* Mol Microbiol, 2002. **44**(1): p. 143-55.
18. Bujnicki, J.M., et al., *Structure prediction meta server.* Bioinformatics, 2001. **17**(8): p. 750-1.
19. Fushinobu, S., et al., *Crystal structures of a meta-cleavage product hydrolase from Pseudomonas fluorescens IP01 (CumD) complexed with cleavage products.* Protein Sci, 2002. **11**(9): p. 2184-95.

20. Kasai, S. and T. Sumimoto, *Stimulated biosynthesis of flavins in Photobacterium phosphoreum IFO 13896 and the presence of complete rib operons in two species of luminous bacteria.* Eur J Biochem, 2002. **269**(23): p. 5851-60.
21. Aravind, L., Y.I. Wolf, and E.V. Koonin, *The ATP-cone: an evolutionarily mobile, ATP-binding regulatory domain.* J Mol Microbiol Biotechnol, 2000. **2**(2): p. 191-4.
22. Spitzer, E.D. and B. Weiss, *dfp Gene of Escherichia coli K-12, a locus affecting DNA synthesis, codes for a flavoprotein.* J Bacteriol, 1985. **164**(3): p. 994-1003.
23. Hogrefe, H.H., et al., *Archaeal dUTPase enhances PCR amplifications with archaeal DNA polymerases by preventing dUTP incorporation.* Proc Natl Acad Sci U S A, 2002. **99**(2): p. 596-601.
24. Icho, T., *Membrane-bound phosphatases in Escherichia coli: sequence of the pgpA gene.* J Bacteriol, 1988. **170**(11): p. 5110-6.
25. Funk, C.R., L. Zimniak, and W. Dowhan, *The pgpA and pgpB genes of Escherichia coli are not essential: evidence for a third phosphatidyl-glycerophosphate phosphatase.* J. Bacteriol., 1992. **174**: p. 205-213.
26. Yang, J.W. and M.E. Schweingruber, *The structural gene coding for thiamin-repressible acid phosphatase in Schizosaccharomyces pombe.* Curr Genet, 1990. **18**(3): p. 269-72.
27. Hohmann, S. and P.A. Meacock, *Thiamin metabolism and thiamin diphosphate-dependent enzymes in the yeast Saccharomyces cerevisiae: genetic regulation.* Biochim Biophys Acta, 1998. **1385**(2): p. 201-19.
28. Fodje, M.N., et al., *Interplay between an AAA module and an integrin I domain may regulate the function of magnesium chelatase.* J Mol Biol, 2001. **311**(1): p. 111-22.
29. Beall, B. and C.P. Moran, Jr., *Cloning and characterization of spoVR, a gene from Bacillus subtilis involved in spore cortex formation.* J Bacteriol, 1994. **176**(7): p. 2003-12.
30. Eichenberger, P., et al., *The sigmaE regulon and the identification of additional sporulation genes in Bacillus subtilis.* J Mol Biol, 2003. **327**(5): p. 945-72.
31. Kofoid, E., et al., *The 17-gene ethanolamine (eut) operon of Salmonella typhimurium encodes five homologues of carboxysome shell proteins.* J Bacteriol, 1999. **181**(17): p. 5317-29.
32. Fleischmann, R.D., et al., *Whole-genome random sequencing and assembly of Haemophilus influenzae Rd.* Science, 1995. **269**(5223): p. 496-512.
33. Karp, P.D., C. Ouzounis, and S. Paley, *HinCyc: a knowledge base of the complete genome and metabolic pathways of H. influenzae.* Proc. Int. Conf. Intell. Syst. Mol. Biol., 1996. **4**: p. 116-124.
34. Fleischmann, R.D., et al., *Whole-genome random sequencing and assembly of Haemophilus influenzae Rd.* Science, 1995. **269**(5223): p. 496-512.
35. Mushegian, A.R. and E.V. Koonin, *Gene order is not conserved in bacterial evolution.* Trends Genet., 1996. **12**: p. 289-290.
36. Stadtman, T.C., *Selenocysteine.* Annu Rev Biochem, 1996. **65**: p. 83-100.