SURVEY AND SUMMARY
‘Conserved hypothetical’ proteins: prioritization of targets for experimental study

Michael Y. Galperin and Eugene V. Koonin*

Nucleic Acids Research Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA

Received August 18, 2004; Revised and Accepted September 22, 2004

ABSTRACT
Comparative genomics shows that a substantial fraction of the genes in sequenced genomes encodes ‘conserved hypothetical’ proteins, i.e. those that are found in organisms from several phylogenetic lineages but have not been functionally characterized. Here, we briefly discuss recent progress in functional characterization of prokaryotic ‘conserved hypothetical’ proteins and the possible criteria for prioritizing targets for experimental study. Based on these criteria, the chief one being wide phylectic spread, we offer two ‘top 10’ lists of highly attractive targets. The first list consists of proteins for which biochemical activity could be predicted with reasonable confidence but the biological function was predicted only in general terms, if at all (‘known unknowns’). The second list includes proteins for which there is no prediction of biochemical activity, even if, for some, general biological clues exist (‘unknown unknowns’). The experimental characterization of these and other ‘conserved hypothetical’ proteins is expected to reveal new, crucial aspects of microbial biology and could also lead to better functional prediction for medically relevant human homologs.

INTRODUCTION
Over the last decade, more than 150 complete genomes of diverse bacteria, archaea and eukaryotes have been sequenced, and many more are currently in the pipeline (1,2). The availability of complete genome sequences has led to radical changes in life sciences, from molecular biology to entomology and plant biology. A popular phrase now refers to the post-genome era as the ‘dark ages’ (3). However, it is well known that, in any newly sequenced bacterial genome, as many as 30–40% of the genes do not have an assigned function (4). This figure is even higher for archaeal and eukaryotic genomes and for the relatively large genomes of bacteria with a complex life style, such as *Anabaena*, *Streptomyces* or *Pirellula* (5–8). Remarkably, species- or genus-specific genes comprise a relatively small fraction of the uncharacterized genes; the majority of such ‘hypothetical’ genes have a wider phylectic distribution and therefore are usually referred to as ‘conserved hypothetical’ (9,10).

Although species-specific ‘ORFans’ should not be neglected (11–14), ‘conserved hypothetical’ proteins pose a challenge not just to functional genomics, but also to biology in general (9). As long as there are hundreds of conserved proteins of unknown function even in model organisms, such as *Escherichia coli*, *Bacillus subtilis* or yeast *Saccharomyces cerevisiae*, any discussion of a ‘complete’ understanding of these organisms as biological systems will remain in the realm of wishful thinking. Indeed, elementary logic suggests that, before attempting to disentangle the riveting complexity of interactions between the parts of biological machines and to develop theoretical and experimental models of these machines, it is necessary to gain the basic understanding of the role of each part. Although it appears likely that the central pathways of information processing and metabolism are already known, crucial elements of these systems could still be lurking among the ‘conserved hypotheticals’, and important mechanisms of signaling and stress response, in all likelihood, remain to be uncovered. The recent discoveries of several nearly universal but not characterized previously tRNA modification enzymes (15–17), of the deoxyxylulose pathway (18,19), and of the central role of cyclic diguanylate in bacterial signaling (20,21) emphasize how much remains to be learnt (for additional examples see Table 1). In addition, it is important to note that, for numerous experimentally characterized enzymes, there is still no available sequence information (22); the ‘conserved hypothetical’ genes are the pool where biologists can fish for these ‘homeless’ activities. Furthermore, there is an important ‘side benefit’: it is often easier to conduct experiments with bacterial than with eukaryotic proteins, and functional characterization of microbial ‘conserved hypothetical’ proteins may facilitate prediction and subsequent experimental study of their human homologs [for a case in point see (23)].

In the epilog to our 2002 book on comparative and evolutionary genomics, we noted that systematic identification of functions of genes, that are conserved in many genomes but remain uncharacterized experimentally, is one of the most tantalizing opportunities offered by the recent progress in genome sequencing and analysis (1) (http://www.ncbi.nlm.nih.gov/books/bv.fcgi?rid=sef). More recently, Richard...
Table 1. Some recently characterized ‘conserved hypothetical’ proteins

<table>
<thead>
<tr>
<th>Gene name</th>
<th>Yeast</th>
<th>Human</th>
<th>Protein family</th>
<th>PDB code</th>
<th>Number of species, essentiality</th>
<th>Current functional assignment</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>mesI</td>
<td>NCS6</td>
<td>LOC90353</td>
<td>Pfam</td>
<td>COG</td>
<td>KOG</td>
<td>70; E. coli, B. subtilis</td>
<td>tRNA-Ile lysidine synthetase</td>
</tr>
<tr>
<td>henK</td>
<td>YDR140w</td>
<td>N6AMT1</td>
<td>n/a</td>
<td>2890</td>
<td>3191</td>
<td>1IN5</td>
<td>1IN8</td>
</tr>
<tr>
<td>yaeE</td>
<td>YDR196c</td>
<td>FLJ22955</td>
<td>n/a</td>
<td>10121</td>
<td>0237</td>
<td>3220</td>
<td>1VHL</td>
</tr>
<tr>
<td>yjbN</td>
<td>UTR1</td>
<td>HT029</td>
<td>n/a</td>
<td>01513</td>
<td>0061</td>
<td>2178</td>
<td>1U0R</td>
</tr>
<tr>
<td>yjgQ</td>
<td>TRM4</td>
<td>TRMB</td>
<td>n/a</td>
<td>02590</td>
<td>0220</td>
<td>3115</td>
<td>n/a</td>
</tr>
<tr>
<td>glnG</td>
<td>PCP1</td>
<td>PSARL</td>
<td>n/a</td>
<td>01694</td>
<td>0705</td>
<td>2980</td>
<td>n/a</td>
</tr>
<tr>
<td>isca</td>
<td>IS2A</td>
<td>HESB</td>
<td>n/a</td>
<td>01521</td>
<td>0316</td>
<td>0119</td>
<td>1R94</td>
</tr>
<tr>
<td>ygbO</td>
<td>PUS7</td>
<td>KIAA1897</td>
<td>n/a</td>
<td>01142</td>
<td>0585</td>
<td>2339</td>
<td>1S17</td>
</tr>
</tbody>
</table>

aA hyperlinked version of this table is available in the online version of the paper at http://nar.oupjournals.org
bA dash indicates the absence of an ortholog in the particular organism; n/a, not available.

The number of complete genomes out of the total of 70 (50 bacterial, 13 archaeal and 7 eukaryotic genomes) in the current release of the COG database, http://www.ncbi.nlm.nih.gov/COG, that encodes a member of the given protein family; the organisms (>2) where the respective gene could not be disrupted by transposon mutagenesis or allelic replacement, suggesting that it might be essential for growth. The organisms studied were S. cerevisiae (66), E. coli (67), B. subtilis (66), Staphylococcus aureus (64), Streptococcus pneumoniae (63,133), Haemophilus influenzae (62) and Mycoplasma pneumoniae (134). The data are from the respective references and the Database of Essential Genes (68). ‘None’ indicates apparent non-essentiality in any studied genomes.

Roberts (10) issued a call for community action in identifying functions of ‘conserved hypothetical’ proteins. He stressed the need for a consortium of bioinformaticians to produce a list of all of the conserved hypothetical proteins that are found in multiple genomes, to carry out the best possible bioinformatics analysis, and then to offer those proteins to the biochemical community as potential targets for research (10). Here, we briefly discuss some recent experimental studies on functional characterization of ‘conserved hypothetical’ proteins and outline possible criteria for selecting the priority targets. Based on these criteria, we offer a ‘top 10’ list of proteins for which a general biochemical but not biological function could be confidently predicted and another ‘top 10’ list of proteins that lack any functional assignment at this time. We hope that this effort stimulates discussion on the prioritization criteria and will help attract the attention of biologists worldwide to the challenge of ‘conserved hypothetical’ proteins.

Recent progress

The families of ‘conserved hypothetical’ proteins can be easily identified by similarity-based clustering or using more complex approaches (24,25). Representative collections of such protein families are available in several public databases. The PROSITE database (26) contains an Uncharacterized Protein Families (UPF) list, http://www.expasy.org/cgi-bin/lists? upf-list.txt, which currently includes 295 families of proteins from the SWISS-PROT database (27); some come with a family-specific sequence pattern (family signature). The latest release of the Pfam database (28) includes 1418 Domains of Unknown Function (DUFs), http://www.sanger.ac.uk/ Software/Pfam/browse/DUF.shtml; each entry comes with a multiple sequence alignment, HMM sequence logo, species distribution and a phylogenetic tree of its members. The latest release of theClusters of Orthologous Groups of proteins (COG) database (29) includes 2009 uncharacterized COGs (see Supplementary Material). The abundance of UPFs makes their study a formidable task. There is a clear need for rational criteria that would allow sorting these protein families and selecting the most important ones, i.e. prioritizing the targets for experimental studies; two obvious criteria are the number of proteins in the family and its phyletic spread.

Since the advent of comparative genomics, wide (better yet, universal) phylogenetic distribution and indispensability for cell growth have been taken into consideration by some researchers when choosing uncharacterized genes for experimental study (30).

Significant positive correlation between the phyletic spread of a gene and the likelihood that it is essential for cell growth has been demonstrated (31). On a number of occasions, experiments with proteins that met one or both of these criteria led to major discoveries. For example, a structural genomics project yielded the three-dimensional (3D) structure and led to subsequent functional characterization of the Methanococcus jannaschii protein MJ0226 (32), a member of the widely distributed HAM1 protein family (Table 1), whose only known function until then had been modulation of sensitivity of 6-N-hydroxyaminopurine mutagenesis in yeast (33). Characterization of this protein as a XTP- and ITP-specific pyrophosphatase immediately explained its role in mutagenesis control. Moreover, this protein perfectly fits the description of the ITP pyrophosphatase (ITPase) from human erythrocytes (EC 3.6.1.19) that had been first reported in 1964 (34), purified and extensively characterized five years later (35), but had never been identified with a gene. Based on the sequence of the MJ0226 protein, its human homolog has been characterized and shown to account for the ITPase activity in humans (36). Furthermore, although mutations in the ITPase gene did not seem to have a clear disease phenotype, ITPase deficiency has been associated with adverse reactions to purine analog azathioprine, which is used as immunosuppressant in the treatment of cancer and inflammatory bowel disease (37). This example shows how a supposedly arcane study of an archaeal...
‘conserved hypothetical’ protein can have immediate consequences for understanding human physiology and might be relevant for human health.

The recent identification of NAD kinase (EC 2.7.1.23) follows a similar pattern. Again, the enzyme has been experimentally characterized many years ago, both in avian tissues and in yeast (38,39), but the cognate gene remained unknown. Again, the characterization of the bacterial enzyme allowed assigning this function to a family of previously uncharacterized ‘conserved hypothetical’ proteins (40). Finally, studies on the bacterial enzyme paved the way for the identification of the orthologous enzyme in humans and in yeast (41,42).

Of course, it would be wrong to assume that functional characterization of a ‘conserved hypothetical’ protein would always turn up a previously described enzymatic activity. In many cases, the underlying biology and/or biochemistry could be unknown or at least not properly appreciated. Thus, the now-famous case of the identification of the product of E.coli hemK gene as a glutamine N’-methyltransferase of peptide release factors (43,44) pointed out the importance of this post-translational modification that had been previously largely overlooked (10,45). The orthologs of HemK in humans and other eukaryotes are still annotated (without experimental support) as DNA methyltransferases; glutamine methylation in eukaryotic proteins remains to be investigated.

Similarly, the recent recognition of the roles of the suf genes in the assembly of the iron–sulfur clusters in bacteria (46,47) has important implications for understanding the functioning of chloroplasts, where these processes seem to be basically the same (48). As in the case of HemK, the original annotation of SufC as ‘ABC-type transporter ATPase’ turned out to be less than precise: although SufC is certainly an ATPase of the ABC-type ATPase family, it does not seem to participate in transport (49).

Several recent discoveries made in the course of characterization of conserved hypothetical genes are listed in Table 1.

Criteria for target prioritization

Phyletic distribution. Different researchers will have widely different priorities when it comes to selecting targets for experimental analysis among the conserved hypothetical genes. In general, however, it is hard to deny that ubiquitous or nearly ubiquitous genes are particularly attractive. The number of genes that are conserved in all cellular life forms (as far as can be judged from the current collection of sequenced genomes) is very small, only ~60 (50), and, to our knowledge, only two of these remain experimentally uncharacterized. It seems logical that these genes should be on top of the functional genomics’ ‘deck of cards’ (Table 2). Almost all known ubiquitous genes encode proteins involved in translation and transcription (50). This and additional considerations suggest that the remaining unknowns in this elite group of genes have important roles in the same or closely associated processes (see below).

A considerably greater number of conserved hypothetical genes are either ‘nearly ubiquitous’, i.e. are missing only in some small genomes, primarily those of parasites, or are universal within one or more of the major divisions of life, e.g. in bacteria or archaea. These genes are also likely to perform crucial functions in the organisms that have them. Analysis of phyletic patterns of genes, however, provides research incentives beyond the straightforward ‘the more the merrier’. Genes with relatively patchy distribution that are, however, found in diverse bacteria, archaea and eukaryotes could be as interesting targets for experimental characterization as the (nearly) ubiquitous ones. Owing to the common phenomenon of non-orthologous gene displacement (51), when the same function in different organisms is performed by two or more unrelated proteins, a substantial fraction of such genes, while not ubiquitous themselves, still might be responsible for universal functions. Moreover, genes that show phyletic patterns (partially) complementary to those of known essential genes can be predicted to perform the same function (52). The recent discoveries of an alternative lysyl-tRNA synthetase, thymidylate synthase, shikimate kinase and several enzymes of thiamine biosynthesis are excellent proofs of this principle (53–56). Holes in known metabolic pathways, when a subset of sequenced genomes do not seem to encode familiar enzymes for some of the reactions comprising a pathway whereas the majority of the enzymes are present, remain numerous (1,57–59).

Essentiality. Owing to the recent genome-wide analyses of knockout phenotypes in several organisms, direct data supporting biological importance (in many cases, even essentiality) is available for numerous widespread genes (62–68). Often, members of the same protein family are recognized as essential for survival in more than one organism (Tables 1 and 2), confirming that these proteins, including uncharacterized or poorly characterized ones, play crucial roles in the cell. However, essentiality for growth is a complex phenomenon (67,69). Studies of auxotrophic mutants show that many genes that are essential for survival in a minimal (poor with nutrients) medium become dispensable for growth in a rich medium. Similarly, mechanisms which are essential for maintaining cell integrity in a hypo-osmotic medium become dispensable for survival in an iso-osmotic medium (70). Furthermore, in the general case, it is a function that...
<table>
<thead>
<tr>
<th>Gene name&lt;sup&gt;b&lt;/sup&gt;</th>
<th>Protein family</th>
<th>PDB code&lt;sup&gt;c&lt;/sup&gt;</th>
<th>Number of species, essentiality&lt;sup&gt;d&lt;/sup&gt;</th>
<th>Additional information and predictions</th>
</tr>
</thead>
<tbody>
<tr>
<td>E.coli</td>
<td>B.subtilis</td>
<td>Yeast</td>
<td>Human</td>
<td>Pfam</td>
</tr>
<tr>
<td>yglD (gcP)</td>
<td>ydeE</td>
<td>QRI7</td>
<td>OSGEP</td>
<td>00814</td>
</tr>
<tr>
<td>ychF</td>
<td>yyaF&lt;sup&gt;(engD)&lt;/sup&gt;</td>
<td>YBR025c</td>
<td>PTD004</td>
<td>06071</td>
</tr>
<tr>
<td>yrdC</td>
<td>ywiC</td>
<td>SUA5</td>
<td>YRDC</td>
<td>01300</td>
</tr>
<tr>
<td>ybeM</td>
<td>ykrU</td>
<td>NIT2</td>
<td>NIT1</td>
<td>00795</td>
</tr>
<tr>
<td>yihA</td>
<td>yssC&lt;sup&gt;(engB)&lt;/sup&gt;</td>
<td>YDR336w</td>
<td>HSPC135</td>
<td>01926</td>
</tr>
<tr>
<td>yigB</td>
<td>ysaA&lt;sup&gt;(engB)&lt;/sup&gt;</td>
<td>YMR130c</td>
<td>C0orf158</td>
<td>00702</td>
</tr>
<tr>
<td>yflR</td>
<td>ypkD&lt;sup&gt;(engB)&lt;/sup&gt;</td>
<td>YNL320w</td>
<td>BEM46</td>
<td>00561</td>
</tr>
<tr>
<td>yfeE</td>
<td>ysnB&lt;sup&gt;(engB)&lt;/sup&gt;</td>
<td>VPS29</td>
<td>PEP11</td>
<td>00149</td>
</tr>
<tr>
<td>—</td>
<td>ylfF</td>
<td>NUG1</td>
<td>Nucleostemin</td>
<td>01926</td>
</tr>
<tr>
<td>yheM</td>
<td>—</td>
<td>AFG1</td>
<td>LACE1</td>
<td>03969</td>
</tr>
</tbody>
</table>

<sup>a</sup>A hyperlinked version of this table with appropriate references is available in the online version of the paper at http://nar.oupjournals.org. The genes are sorted by the number of species in which they were detected. Additional selection criteria included existence of human and plant (*Arabidopsis thaliana*) orthologs.

<sup>b</sup>A dash indicates the absence of an ortholog in the particular organism.

<sup>c</sup>PDB codes of family members are shown in normal font, of distant homologs (in italics). Fold prediction is shown where structures are unavailable.

<sup>d</sup>See footnote to Table 1 for calculation of the number of species and the sources of essentiality data.
is essential for the survival of an organism rather than a particular gene. Partial redundancy is common even in organisms with relatively small genomes, such that two paralogous or analogous (unrelated) genes can each provide an essential function and, accordingly, are synthetically lethal, even as the knockout of each of them is not (71). Finally, examples of the ITPase and rhomboid-like serine protease (Table 1) clearly show that certain genes might have important functions even if their knockout is non-lethal. However, absence of easily observable phenotype is a major obstacle to deciphering the functions of many conserved hypothetical genes (Table 3).

**Protein structure.** The availability of a 3D structure (X-ray or NMR) can be used as an additional criterion for prioritization. Structural genomics projects keep churning out new structures at an ever-increasing pace (72,73). As a result, many families of conserved hypothetical proteins already have one or more representatives of a known 3D structure (Tables 2 and 3). Although knowledge of the 3D structure rarely allows unequivocal functional prediction, it often provides valuable clues that substantially narrow down the range of possible functions (74). For example, the demonstration that members of the vast superfamily of proteins homologous to the universal stress protein A of *E.coli* have a distinct nucleotide-binding fold and, indeed, bind nucleotides (75,76) suggests a specific range of functions for the numerous uncharacterized members of this superfamily (77) and is likely to channel experimental studies.

**Expression and binding information.** Another criterion for prioritization is the availability of experimental data on the expression of a given gene or binding properties of its protein product. The proliferation of whole-genome expression and protein–protein interaction studies often provides evidence that a particular ‘conserved hypothetical’ protein is over-expressed under certain conditions (nutritional or oxidative stress, ultraviolet irradiation, and so on) (65,78–81). Although this rarely furnishes direct indications of the protein’s function, its likely involvement in the given stress response could make it a priority target for researchers in the respective field. Similarly, involvement of an uncharacterized protein in protein–protein interaction with a particular signaling or regulatory protein or its ability to bind RNA or DNA would attract attention of researchers studying these regulators or responses.

**Other criteria.** Obviously, prediction of biochemical function is an important guide, e.g. for projects aimed at functional characterization of enzymes with a particular activity (such as protein kinases or phosphatases) encoded in a given genome. Last but not least, practical considerations, whether a protein is likely to be soluble, highly expressed, could be easily purified etc., by no means can be dismissed. Only experiment will be the final judge but indications of a protein’s behavior can often be gleaned from its size and properties predicted by sequence analysis.

**‘Known unknowns’ and ‘unknown unknowns’**

As the list of the ‘conserved hypothetical’ proteins keeps growing, comparative genomics can help in identifying (based on the criteria outlined above and, undoubtedly, additional ones) the most interesting proteins in each genome and, in many cases, constructing reasonable and testable hypotheses on their functions. As noted previously (9), when an open reading frame is annotated as a ‘conserved hypothetical’ protein, this does not necessarily mean that the function of its product is completely unknown, let alone that its very existence is questionable. Indeed, certain exceptions notwithstanding, if a gene is conserved in several sufficiently different genomes, it is not really hypothetical anymore. A general prediction of its function often can be made based on a conserved protein sequence motif, subtle sequence similarity to previously characterized proteins or the presence of diagnostic structural features. Many ‘conserved hypothetical’ proteins can be confidently predicted to be ATPases, GTPases, methyltransferases, metalloproteases, DNA- or RNA-binding proteins or membrane transporters (1,29,82–84). Additional hints regarding protein functions come from genome context analysis (known and predicted operons and more complex gene neighborhoods), similar or complementary phylogenetic patterns, domain fusions and protein–protein interaction data (52,85–92). Many databases and tools for gene context analysis are available on the web (93–98). Homology-based methods typically result in the prediction of structural properties and biochemical activity, often only generally defined (e.g. a phosphatase or kinase of unknown specificity). In contrast, context-based methods usually suggest a general biological function (e.g. participation in cell division) but explain little or nothing about the mechanistic role of the given protein in this process. To better define these two groups of proteins, which differ in terms of the nature of the attainable functional prediction, we refer to those that have at least a general prediction of a biochemical function as ‘known unknowns’. In contrast, those proteins that have no assignment of biochemical function are referred to as ‘unknown unknowns’, even when they have been associated with a certain biological process through genome context analysis or some indirect experimental evidence.

In the latest release of the COG database [(29); http://www.ncbi.nlm.nih.gov/COG/new/release/phyloxCgi], among 4873 COGs, 668 should be considered ‘known unknowns’ (biochemical activity but not biological function predicted) and 1341 have the status of ‘unknown unknowns’ (see Supplementary Material). Some of the more attractive targets for detailed experimental studies in these two categories are listed in Tables 2 and 3. A degree of subjective judgment is inevitable here. We chose not to apply a single criterion but to hand-pick interesting uncharacterized proteins on the basis of a combination of the criteria discussed above; the lists of ‘known unknowns’ and ‘unknown unknowns’ from the COG databases, ranked in the order of decreasing number of species in which they are represented, are available for formal target selection in the Supplementary Material. The criteria employed for the selection of targets listed in Tables 2 and 3 were as follows: (i) wide phyletic distribution, i.e. presence in several distinct phylogenetic lineages (with one exception, these protein families are encoded in representatives of all three domains of life, bacteria, archaea and eukaryotes); (ii) presence in model or otherwise interesting organisms (in particular, we chose to include in Tables 2 and 3 only those genes that have orthologs in humans); (iii) apparent essentiality in at least some organisms; involvement in a fundamental biological function suggested by
<table>
<thead>
<tr>
<th>Gene name&lt;sup&gt;b&lt;/sup&gt;</th>
<th>Protein family</th>
<th>PDB code</th>
<th>Number of species, essentiality&lt;sup&gt;c&lt;/sup&gt;</th>
<th>Additional information and predictions</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>E.coli</strong></td>
<td><strong>B.subtilis</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>yebC</td>
<td>yrbC</td>
<td>PRO0477</td>
<td>01709</td>
<td>0217</td>
</tr>
<tr>
<td>ybgI</td>
<td>yafO</td>
<td>NIF3</td>
<td>NIF3L1</td>
<td>01784</td>
</tr>
<tr>
<td>ybeB</td>
<td>yqeL</td>
<td>C7orf30</td>
<td>02410</td>
<td>0799</td>
</tr>
<tr>
<td>yjeF</td>
<td>—</td>
<td>YNL200c</td>
<td>AIBP</td>
<td>03853</td>
</tr>
<tr>
<td>yigZ</td>
<td>yyiE</td>
<td>YDL177c</td>
<td>IMPACT</td>
<td>01205</td>
</tr>
<tr>
<td>rtcB</td>
<td>—</td>
<td>HSPC17</td>
<td>01139</td>
<td>1690</td>
</tr>
<tr>
<td>ydjX</td>
<td>ytxB</td>
<td>YKR088c</td>
<td>Unnamed</td>
<td>00597</td>
</tr>
<tr>
<td>yygY</td>
<td>—</td>
<td>EMI5</td>
<td>FLJ20487</td>
<td>03937</td>
</tr>
<tr>
<td>—</td>
<td>—</td>
<td>YOR289w</td>
<td>AMMECRI</td>
<td>01871</td>
</tr>
<tr>
<td>ydiU</td>
<td>—</td>
<td>YPL222w</td>
<td>SELO</td>
<td>02696</td>
</tr>
</tbody>
</table>

<sup>a</sup>A hyperlinked version of this table with appropriate references is available in the online version of the paper at http://nar.oupjournals.org. The genes are sorted by the number of species in which they were detected.
<sup>b</sup>A dash indicates the absence of an ortholog in the particular organism; n/a, not available.
<sup>c</sup>See footnote to Table 1 for calculation of the number of species and the sources of essentiality data.
<sup>d</sup>Unconfirmed data.
homology-based and/or context analysis; and (iv) small number of paralogs (which translates into greater confidence in any functional prediction). To illustrate the complexities involved in predicting functions of these proteins, we discuss the top four proteins in each group in greater detail.

‘Known unknowns’. The list of genes in Table 2, which is more informative thanks to the prediction of biochemical activities, includes proteins with predicted diverse fundamental functions. It is no accident that this list is dominated by proteins implicated in translation because this is the function of the majority of ubiquitous or nearly ubiquitous genes (50). Thus, members of the O-sialoglobin protease family (Table 2) are homologous to a single experimentally characterized protein, a neutral metalloprotease from *Pasteurella haemolytica* [a member of the M22 family in the MEROPS database (99)] that has been reported to cleave *O*-glycosylated but not *N*-glycosylated or non-glycosylated proteins (100). Sequence analysis of this protein family revealed a conserved fold and a characteristic ATP-binding site, which are typical of proteins of the actin/HSP70 superfamily, suggesting that these proteins are metal- and ATP-dependent proteases (101). Genomic context data show an association (divergent operon) with the *rpsU* gene encoding ribosomal protein S21 in many enterobacteria and fusion to a Ser/Thr protein kinase domain in some archaea. All this evidence notwithstanding, it is hard to predict the exact function of this protein; gene neighborhood analysis and presence in all completely sequenced genomes suggest association with translation, e.g. co-translational degradation of misfolded proteins (102).

YchF proteins, members of the second universally represented family of uncharacterized proteins (DUF933 in Pfam), contain a typical GTP-binding N-terminal domain and are confidently predicted to have GTPase activity (Table 2) (103). Structures of two members of this family, *Haemophilus influenzae* HI0393 and *Schizosaccharomyces pombe* protein Spac27e2.03c, have been solved (PDB entries 1jal and 1n13, respectively) and revealed a three-domain organization, with a putative nucleic acid-binding domain and a flexible hinge in addition to the GTPase domain (104). The combination of predicted GTPase activity, universal phylogenetic distribution, coexpression with peptidyl-tRNA hydrolyase and the experimentally demonstrated ability to bind double-stranded nucleic acids (104) strongly suggests that YchF proteins function as GTP-dependent translation factors. Owing to the preference given to genes with a wide phyletic spread, it is not surprising that several proteins in our list of ‘known unknowns’ are implicated in translation and associated functions. Notably, the list includes two more predicted GTPases, which are expected to function as translation factors (Table 2) (103).

The nearly ubiquitous YrdC protein is the ortholog of yeast SUA5 gene product that has been identified as a suppressor of a translational defect in cytochrome c production (105). A subsequent genetic study showed that mutation of the *E.coli* *yrdC* gene caused a defect in 16S rRNA maturation (106). These experiments linked YrdC with translation and ribosomal biogenesis but would place this protein into the ‘unknown unknowns’ category. However, once the structure of the protein had been determined, revealing a novel fold, and its nucleic acid-binding properties, with preference for double-stranded RNA have been demonstrated, YrdC became a ‘known unknown’, an RNA-binding protein associated with ribosomal maturation and function (107).

Two recently characterized families of widespread, essential proteins turned out to be enzymes involved in cofactor biosynthesis, namely, NAD kinase and dephospho-CoA kinase (Table 1). One could expect that at least some of the remaining uncharacterized proteins also will be metabolic enzymes. In particular, the widespread YbeM family belongs to the nitrilase superfamily (108), which includes experimentally characterized amidases, hydrolyzing the C–N bond in various substrates. Some of the members of the YbeM family are fused to the NadE domain, which accounts for the glutamine-dependent NAD synthetase activity in humans and yeast, as opposed to the NH$_3$-dependent NAD synthetase activity of the NadE domain alone (109). Apparently, some YbeM-like nitrilases can provide the glutaminase component of NAD synthetase in *trans* (when they are not fused in a single polypeptide chain) by forming a complex with NadE subunits (K. Shatalin and A. Osterman, personal communication). This opens the intriguing possibility that all or at least the majority of the proteins currently annotated as NH$_3$-dependent NAD synthetases function as glutamine-dependent enzymes *in vivo* and only become NH$_3$-dependent when stripped of their YbeM-like subunits.

‘Unknown unknowns’. Families of ‘conserved hypothetical’ proteins listed in Table 3 are remarkable in that none of them has a recognizable experimentally characterized homolog. Many of them do not seem to be essential for cell growth (Table 3). Hence, phyletic spread had to be the principal criterion for ranking these targets; additional indications of their importance come from the still sparse genome-context data.

Although the function of YebC proteins (UPF0082 and DUF28), the most widespread of the ‘unknown unknowns’, remains unknown, this family has been extensively characterized from the structural perspective. The X-ray structures of YebC proteins from *Aquifex aeolicus*, *E.coli* and *Helicobacter pylori* have been solved (PDB entries 1lfp, 1kon and 1mw7, respectively). Structural analysis revealed a large cavity with a predominance of negatively charged residues on the surface of this protein (110). Given the strong contextual association of the yebC gene in various bacteria with *ruvC* and other *ruv* genes, which encode subunits of the Holliday junction resolvase, the YebC protein might be involved in DNA recombination and repair, perhaps as an auxiliary subunit of the resolvosome. However, eukaryotes, which lack the RuvABC resolvase, have highly conserved orthologs of YebC which are predicted to localize to the mitochondria (M. Y. Galperin and E. V. Koonin; unpublished data). Therefore, the proteins of this COG might have different functions in bacteria and in eukaryotes.

The second most common ‘unknown unknown’ protein, NIF3, has been originally described in yeast as an NGG1-interacting protein (111) and later identified in humans and mouse as a protein involved in transcriptional regulation of neural differentiation (112,113). Although its exact function remains unknown, this gene has been indirectly implicated in such diseases as juvenile amyotrophic lateral sclerosis (ALS2) and Williams–Beuren syndrome (114,115). The function of this protein in bacteria and archaea remains enigmatic.
The *iojap* (Iowa-japonica) mutation affecting a representative of the third most common ‘unknown unknown’ family of proteins (YbeB/DUF143) has been described back in 1924 as causing a characteristic pattern of white (plastid-less) and green (normal) stripes in the leaves of maize (116). Studies of the past two decades showed that *iojap* mutants lack functional plastid ribosomes; neither plastid-encoded proteins nor nuclear-encoded proteins that are associated with thylakoid membranes are detectable in their chloroplasts (117–119). Nevertheless, the actual function of the *iojap* protein remains unknown; the suggestion that it might be involved in RNA editing has been recently investigated and proven incorrect (120). In bacteria, the ortholog of *iojap*, the *ybeB* gene, forms a conserved operon with another uncharacterized gene, *ybeA*, which encodes a predicted methyltransferase, and is often found next to *nadD*, a gene for an enzyme of NAD biosynthesis (nicotinate-mononucleotide adenyllyltransferase). In two bacteria, *NadD* and *YbeB* seem to form a two-domain fusion protein. The significance of this association remains unknown. A recent work has identified the corresponding *E.coli* gene as one of the two genes whose mutations confer lethality on *ΔminCDE* mutants, suggesting involvement of the *Iojap* protein in bacterial cell division (121).

Inspection of the available evidence for the YjeF protein (UPF0029) shows that the distinction between ‘known unknowns’ and ‘unknown unknowns’ sometimes becomes murky. Specific prediction of the biochemical activity of this protein does not seem feasible, which is why we kept it among ‘unknown unknowns’ (Table 3). However, combination of detailed structural and sequence analyses, with examination of the domain fusion context strongly suggests that this protein is a Rossmann-fold enzyme involved in RNA metabolism (122,123).

Conclusions and perspective

Sequencing of multiple genomes from all walks of life and the concomitant development of computational approaches of comparative genomics create an opportunity for biology that was hardly imaginable even 10 years ago: a directed, systematic effort aimed at producing a complete catalog of biochemical activities, biological functions and the responsible genes, at least for simpler, prokaryotic life forms. A coordinated program on elucidation of the functions of conserved hypothetical proteins has the potential of taking us (as a community) a long way on the road to this lofty goal. It is worth emphasizing that the number of conserved hypotheticals that are widely represented among diverse life forms is not huge, a few thousand on the outside. However incomplete the current collection of genomes turns out to be, genes from new genomes increasingly fall within already established orthologous gene sets (124). Thus, although a truly comprehensive gene catalog might belong in the distant future, a concise dictionary of the ‘main’ functions and the corresponding genes is likely to be well within reach of the current generation of researchers, provided a reasonable degree of coordination of research projects is achieved.

Certainly, science does not live by plans conceived by any group of researchers. Nevertheless, community research programs can be viable as illustrated, most relevantly for our subject here, by the brief (so far) but notable history of the structural genomics initiative. Although the development of structural genomics might not have been as rapid as initially hoped, the overall success is undeniable (73,125–127). The number of protein structures solved within this initiative, including those of many conserved hypothetical proteins that might have never been tackled, if not for the structural genomics paradigm, is already in the hundreds, and the movement is just picking up speed. There seems to be no reason why systematic, genome-wide identification of gene functions could not proceed along the same lines, perhaps with the benefit of even better coordination between researchers.

SUPPLEMENTARY MATERIAL

Supplementary Material is available at NAR Online.

NOTE ADDED IN PROOF

While this article was in production, an in-depth discussion of the uncharacterized genes of yeast *Saccharomyces cerevisiae* and strategies for completing the ‘encyclopedia of the yeast cell’ has been published [Hughes,T.R., Robinson,M.D., Mitsakakis,N. and Johnston,M. (2004) The promise of functional genomics: completing the encyclopedia of a cell. *Curr. Opin. Microbiol.*, 7, 546–554].

ACKNOWLEDGEMENTS

We thank Rich Roberts, Eugene Kolker, Andrei Osterman and other participants of the American Academy of Microbiology colloquium ‘An Experimental Approach to Genome Annotation’ (July 19–20, Washington, DC) for inspiring discussions and Daniel Rigden for helpful comments.

REFERENCES


