Published online December 1, 2004

Nucleic Acids Research, 2004, Vol. 32, No. 21 ***6321–6326***
doi:10.1093/nar/gkh973

Functional clues for hypothetical proteins based on genomic context analysis in prokaryotes

Tobias Doerks*, Christian von Mering and Peer Bork

EMBL, 69012 Heidelberg, Meyerhofstrasse 1 and Max-Delbrueck-Centrum, Robert-Roessle-Strasse 10, D-13092 Berlin, Germany

Received October 6, 2004; Revised and Accepted November 11, 2004

ABSTRACT

Three integrated genomic context methods were used to annotate uncharacterized proteins in 102 bacterial genomes. Of 7853 orthologous groups with unknown function containing 45 110 proteins, 1738 groups could be linked to functionally associated partners. In many cases, those partners are uncharacterized themselves (hinting at newly identified modules) or have been described in general terms only. However, we were able to assign pathways, cellular processes or physical complexes for 273 groups (encompassing 3624 previously functionally uncharacterized proteins).

INTRODUCTION

During the last few years, the speed and cost-effectiveness of genome sequencing has increased enormously, and with >100 bacterial genomes now available, the quality and comprehensiveness of their annotation becomes a demanding problem. Even in the most recent annotations, a large fraction of open reading frames are still labelled as 'conserved hypothetical proteins', sometimes representing more than the half of the potential protein-coding regions of a genome (1).

Many of these 'hypothetical proteins' occur in fact in more than one bacterial species, and can thus be combined into orthologous groups; this subset of proteins contains the majority of biologically relevant sequences (less likely to be artefactual), and it is amenable to analysis by comparative genomics techniques.

For assigning function to novel proteins, homology-based gene annotation has been the standard during the last decades. However, novel methods have been developed recently, which can complement the classical homology search; these methods are designed to detect presumed functional constraints on genome evolution, and are called 'genomic context' approaches.

They predict functional associations between protein-coding genes by analyzing gene fusion events, the conservation of gene neighbourhood, or the significant co-occurrence of genes across different species (2–7). Unlike homology-based annotation, which infers molecular features by information transfer from experimentally characterized proteins, genomic context methods predict functional associations *between* proteins, such as physical interactions, or co-membership in pathways, regulons or other cellular processes (8). Characterizing protein function in this manner (i.e. by predicting associated partners) is intuitive and generally applicable, but it should be noted that it does not provide information about the exact biochemical or enzymatic function of a protein. Genomic context methods have been successfully used to study protein associations, either individually (2–6) or in combination with other methods or data sets (7,9,10). Recently, a combination of genomic context methods has been applied to infer functional associations between proteins in archaea (11), or to identify functional modules in *Escherichia coli* (12). Furthermore, these methods have been used to identify genes that correlate with the hyperthermophilic phenotype (13) or to predict target processes for transcription regulators (14). Yet, despite these efforts >35% of genes in prokaryotes are still annotated as 'function unknown' (15). The urgent need to functionally characterize these proteins and to bridge this gap in our knowledge, is highlighted by a recent call for community action (1).

Here, we aim at reducing this fraction considerably, by a systematic study of uncharacterized proteins in prokaryotes. We exploit the genomic context of this set to predict their functional role, by determining the biological/cellular process in which the proteins participate.

MATERIALS AND METHODS

Starting set of proteins of unknown function

From the manually curated clusters of orthologous groups (COG) database (16) (http://www.ncbi.nlm.nih.gov/COG/), we extracted clusters of orthologous genes, which were annotated as 'hypothetical' or 'uncharacterized'. In the original procedure to create COGs (16), orthologues are identified using an all-against-all sequence comparison of the proteins encoded in completely sequenced genomes. In considering a protein from a given genome, this comparison reveals the one protein from each of the other genomes to which it is most similar. Each of these proteins is in turn considered. If a reciprocal best-hit relationship between these proteins (or a subset) is revealed, then those that are reciprocal best-hits will form a COG. An additional constraint in this procedure is that a COG must be comprised of one protein from at least three phylogenetically distant genomes.

*To whom correspondence should be addressed. Tel: +49 6221 387456; Fax: +49 6221 387517; Email: doerks@embl.de
Correspondence may also be addressed to Christian von Mering. Tel: +49 6221 387534; Fax: +49 6221 387517; Email: mering@embl.de

We have recently (7) extended the COG database by considering an additional set of 37 newly sequenced complete genomes. This resulted in both the extension of the original COGs with new proteins, but also in the creation of entirely novel orthologous groups, which we termed 'NOGs' (non-supervised orthologous groups). Similar to the original procedure, assignments for NOGs were made based on triangles of reciprocal best matches between species in all-against-all Smith–Waterman searches, allowing for recent duplications within the genome, and including a clean-up step to join any remaining genes by simple bidirectional hits. NOGs are fully automatically generated, and they do not have any manually curated functional annotation (for a more detailed description see Supplementary Material, p. 12).

Manual genome context analysis using String

Functionally unannotated COGs and NOGs were analysed using the tool STRING (Search Tool for the Retrieval of Interacting Genes/Proteins, http://string.embl.de/) (7), applying a conservative score threshold of 0.4 [for a benchmark see (7)]. STRING calculates this 'confidence score' on the basis of three genomic context methods: conserved gene neighbourhood, gene fusion events and significant co-occurrence of the genes across a specific subset of species. The prediction accuracy of functional links is often higher than the confidence score indicates [e.g. when tested against *E.coli* small molecule metabolism (12)]. Each COG and NOG was queried manually, and the STRING output was inspected—asking, whether it allowed the assignment of a cellular role to the uncharacterized proteins clustered in this group, based on their predicted functional partners.

RESULTS AND DISCUSSION

In a global search for conserved uncharacterized proteins, we retrieved a total of 7853 orthologous groups annotated to be of unknown function. All groups contain orthologous proteins derived from several genomes.

In the set of 102 prokaryotic genomes considered here, 82% of all the genes are included in orthologous groups. Most of the remaining proteins might be without orthology to any other genome and can thus not be included in a comparative analysis; others might even be gene prediction artefacts.

The list of functionally uncharacterized group contains 1162 groups originally assembled by Tatusov *et al*. (16) (COGs), and 6691 groups derived from more recently completed

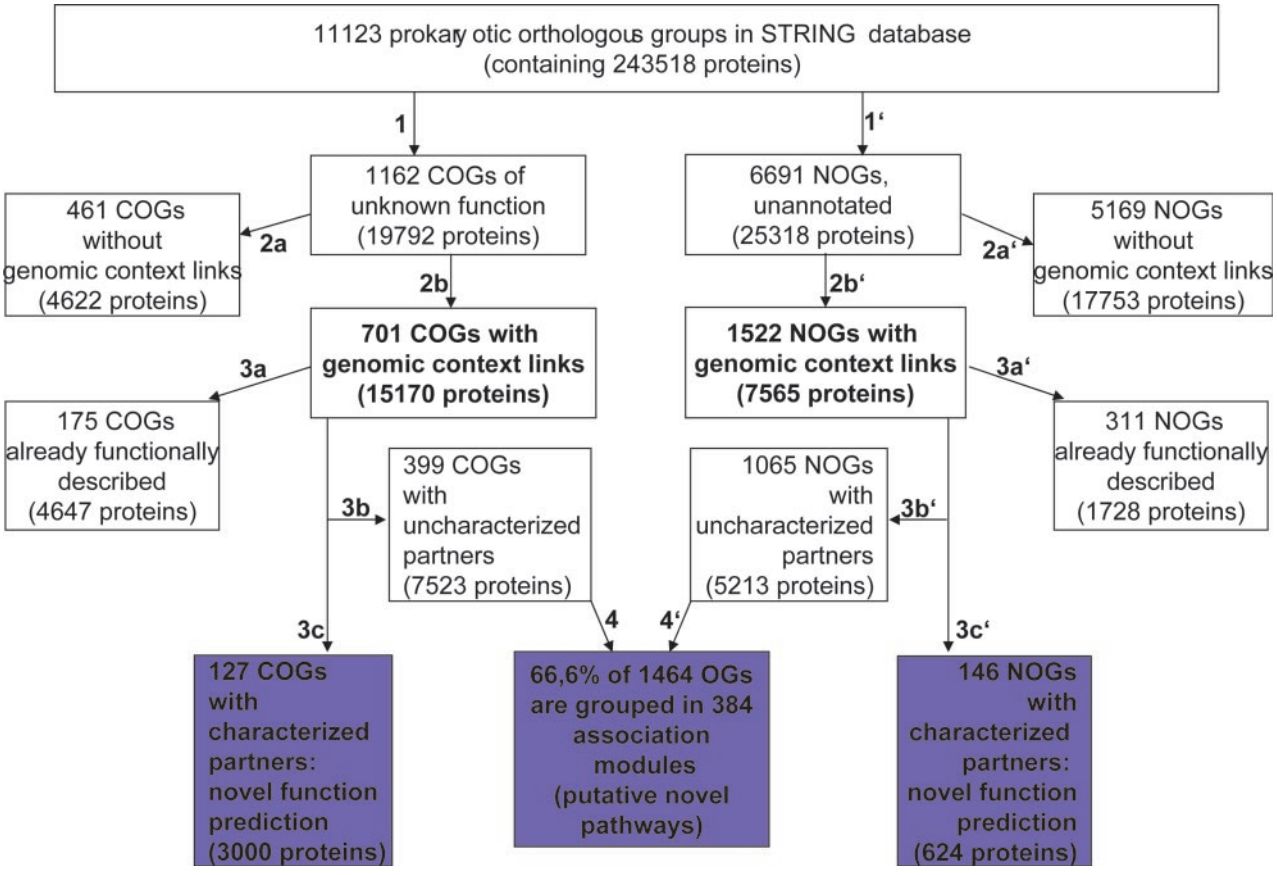

Figure 1. Uncharacterized prokaryotic proteins: prevalence and flow chart of the annotation strategy. (1) Extraction of COGs, annotated as 'hypothetical' or 'uncharacterized', and (1') NOGs without any functional annotation from STRING database. (2a) Exclusion of COGs and (2a') NOGs without any STRING links. (2b) Identification of COGs and (2b') NOGs with STRING links (information about conserved gene neighbourhood, gene fusion events or phylogentic co-occurrence with genes in other COGs/NOGs, respectively). (3a) Identification of COGs and (3a') NOGs already functionally described. (3b) Identification of COGs and (3b') NOGs with STRING links to functionally weak or uncharacterized COGs/NOGs. (3c) Identification of COG and (3c') NOGs with STRING links to functionally informative COGs/NOGs. (4) Grouping of functionally uncharacterized COGs and (4') NOGs into association modules.

genome sequences, by a similar procedure (7) (NOGs). The entire list was submitted to the STRING server for genomic context analysis, using a score threshold of 0.4. We found that 5628 of the 7853 COGs/NOGs did not reveal any genomic context association. The remaining 2225 groups, however (703 COGs and 1522 NOGs, see Figure 1), were found to be functionally related to other orthologous groups based on gene neighbourhood, gene fusion events or phylogenetic co-occurrence.

We manually checked the list to remove COGs/NOGs, for which some molecular function was indeed already known [e.g. in the SWISSPROT database (17)]. This reduced the list to 1737 groups (containing 6367 proteins).

For 399 COGs and 1065 NOGs of these groups (84.1% of 1740), we identified only uncharacterized interaction partners or only proteins with a very unspecific function (see Supplementary Material). We observed that these groups form a larger association network of uncharacterized proteins; we split this network by a standard clustering procedure [for methods see (12)], in order to identify sets of hypothetical proteins that potentially co-operate in separatable functional units. We observe that 975 groups of orthologues (66.6%) are clustered in 384 functional modules (see Figure 1), whereas 492 have only single connections. The modules of unknown function identified here may represent novel pathways, complexes or operons, and should be given high priority for experimental analysis, as they are likely to perform complex functions requiring several gene products, and are important enough for the cell to be selected for as a unit.

For the remaining 127 COGs and 146 NOGs, we were indeed able to predict a cellular role, a connected pathway or a complex (for selected representatives and Supplementary Material for the whole list see Table 1). A novel function was predicted for 24% of COGs, but only for 12% of NOGs, showing a correlation with orthologous group size (average size: ~17 proteins/COG and ~4 protein/NOG) and the likelihood of belonging to a conserved genomic context.

We found functional associations for uncharacterized COGs/NOGs to a broad variety of cellular activities. These include chromatin-associated processes such as DNA-repair, transcription and translation, metabolic or signalling pathways and membrane-associated transport and secretion processes.

The specificity and annotation depth of predicted partners for COGs/NOGs of unknown function varies. Associated partners range from well-defined and experimentally characterized COGs, when, for instance, the hypothetical COG3310 phylogenetically co-occurs significantly in several proteobacteria with an experimentally characterized operon of pilus assembly proteins (18), to barely specified pathways such as COG4029, COG4048, COG4050, COG4051, COG4052 and COG4069, which appear only in methanogenic archaea in a conserved operon with genes coding for phosphatases (COG4087) (19,20) (Table 1).

Newly annotated NOGs are often related to very specific cellular processes, present in very few species known so far [e.g. NOG14679 related to Photosytem II (21)].

Recent experimental data confirm and complete our genomic context approach. An example is NOG06495 (see Table 1) consisting of hypothetical proteins from three different species, which our procedure predicts to be associated to proteins of the protein-secretion pathway type III. Recently, experimental evidence was published (22), which is consistent with our genomic context prediction.

In several cases, a number of uncharacterized COGs were found clustering together, but were also linked to at least one better characterized group. We found, for instance, a cluster of 15 groups (including COG3456, see Table 1 and Supplementary Material), revealing a putative novel functional pathway or complex; genes of these orthologous groups are closely associated (scores range >0.7) by conserved gene neighbourhood (data not shown) and phylogenetic co-occurrence (see Figure S1a in Supplementary Material) in several proteobacterial species. However, the proteins of this novel pathway are also connected to homologues of a flagellar motor protein (COG1360) (23) by conserved operon architecture, a fusion event with COG3455 (hypothetical) and phylogenetic

Table 1. Selected uncharacterized orthologous groups, and their putative functions or pathways predicted here

Orthologous group	Represent. gene name	Distribution (genes in species)	Predicted function or associated cellular process
COG0217	YEBC	103/90	Holliday junction process, DNA repair
COG1507	SCO3094	10/10	Septum formation/cell division
COG3496	XAC1376	13/13	Fatty acid biosynthesis
NOG11257	XAC1374	5/5	
COG3055	YJHT	22/17	TRAP-type transport system
COG3310	XAC3675	9/9	Pili assembly (see Supplementary Material, p. 14)
COG3456	Z0258	15/13	Membrane-associated transport processes
COG3501	YPO0507	67/20	
COG3515	Z0252	41/19	
COG3516	Z0264	29/17	
COG3517	Z0262	30/16	
COG3518	Z0261	27/18	
COG3519	Z0260	29/18	
COG3520	Z0259	30/18	
COG3521	Z0257	21/13	
COG3522	Z0256	29/17	
COG3523	Z0250	27/16	
COG3913	PA0076	7/7	
COG4104	PA3904	23/14	
COG4455	XAC4144	10/10	
COG3826	YM27	6/6	DNA-repair
COG4029	MJ0498	5/5	Possibly signalling processes in methane metabolism
COG4048	MJ0405	5/5	
COG4050	MJ0404	5/5	
COG4051	MJ0802	5/5	
COG4052	MJ0094	5/5	
COG4069	MJ0065	8/5	
COG4687	SPS0353	14/13	Phosphotransferase system
NOG06495	hpaB	3/3	Secretion
NOG06496	hrpW	3/3	
NOG08450	hpaP	3/3	
NOG10726	hpaA	3/3	
NOG14679	SSL1690	5/4	Photosystem II

First column: name of orthologous group; second column: representative gene name; third column: gene distribution (number of genes/number of species); and fourth column: function summary (pathway, process or function of predicted partners). The complete list of COGs/NOGs with predicted function is provided in the Supplementary Material. The results were produced by version 4.0 of STRING and can be reproduced here: http://dag.embl-heidelberg.de/newstring_cgi/show_input_page.pl (COG database and STRING as on September 2003).

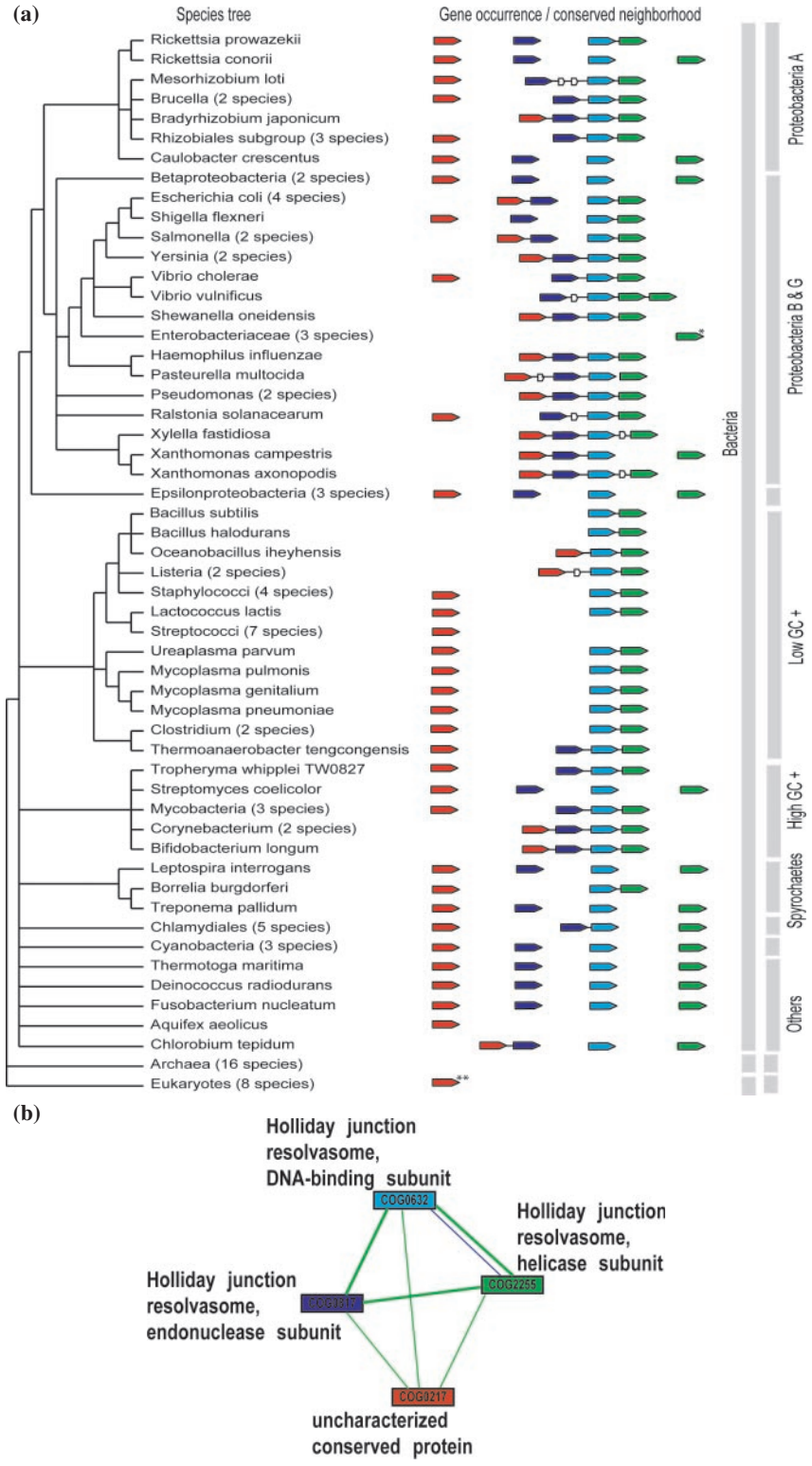

Figure 2. (**a**) Case study: evidence linking COG0217 to well-annotated proteins. Species tree showing conserved operon architecture and co-occurrence of genes coding for subunits of Holliday junction resolvasome (COG0217: uncharacterized conserved protein (red), COG0632: Holliday junction resolvasome, DNA-binding subunit (light blue), COG2255: Holliday junction resolvasome, helicase subunit (dark blue), COG0817: Holliday junction resolvasome, endonuclease subnit (green). Single asterisk (*), not present in Buchnera species and double asterisks(**), not present in Encephalitozoon cuniculi. (**b**) Network representation of evidence related to COG0217 (red). The network edges represent the predicted functional associations. An edge may be drawn with up to three different colour lines—these lines represent the existence of the three types of evidence used in predicting the associations. A red line indicates the presence of fusion evidence; a green line represents the neighbourhood evidence; and a blue line the co-occurrence evidence. Line thickness correlates linearly with STRING scores.

co-occurrence (see Figure S1b in Supplementary Material). Proteins of COG1360 contain an OmpA/MotB domain and are thought to function as porin-like integral membrane proteins (24) or lipid-anchored proteins (25); and COG3523 contains an ImcF domain, which has been proposed to be involved in *Vibrio cholerae* cell surface reorganization (26). Furthermore COG3515 contains the 'ImpA-related N-terminal domain' of inner membrane proteins; this domain has been found in extracellular proteins and is associated with colony variations in *Actinobacillus actinomycetemcomitans* (27). These findings support the assumption that the novel pathway/protein complex plays a role in a membrane-associated transport process.

An example for a more specific prediction, via association to a well-characterized pathway, is COG0217, which consists of hypothetical proteins present in a variety of eubacterial species. Crystal structure analyses of a representative of this group (Aq1575 from *Aquifex aeolicus*) give no hint to any putative function (28). As shown in Figure 2a, genes coding for these proteins occur consistently (in 24 species) in a putative operon together with Holliday junction resolvasome genes. The resolvasome in *E.coli* is known to play an important role in the late stages of homologous genetic recombination and in the recombinational repair of damaged DNA (29). In the majority of species, the functional resolvasome depends on three different subunits (DNA-binding subunit RUVA, DNA helicase RUVB and DNA endonuclease RUVC).

Structural studies indicate that two RuvA tetramers sandwich the formation of heteroduplex DNA and hexameric rings of RuvB face the junction. Thus, RuvB promote dual helicase action that 'pumps' DNA through the RuvAB complex by ATP hydrolysis.

The third protein, RuvC endonuclease, resolves the Holliday junction by introducing nicks into two DNA strands (29).

RUVC is absent in most of low GC Gram-positives (30) and in *Borrelia*; further studies in *Mycoplasma pneumoniae* suggest that the Holliday branch migration and resolution is different from *E.coli* and a novel resolvase is being searched for (31).

Proteins of COG0217 are frequently found in conserved gene neighbourhood with the classical 'three-protein-resolvasome', but also in the equivalent operon of the atypical resolvasome of some low GC Gram-positives, missing the endonuclease (Figure 2a). The network view (Figure 2b) illustrates the functional association of COG0217 based on its conserved neighbourhood with well-annotated genes, and the functional associations between the other genes of this operon. Together with other, below-threshold associations (see Supplementary Material, Info-Box 1), an accessory function in DNA repair is predicted, possibly in response to oxidative damage and in conjunction with the resolvasome (32).

It should be noted, that while the predicted association with the resolvasome is quite strong, it does not provide a precise molecular function for the members of COG0217. However, as for many other such cases, the prediction significantly narrows down the putative function of this group. The prediction can serve as a guide for future experimental inquiry, aiming to identify the role of these hypothetical proteins in the Holliday junction resolvasome, for example, by selected mutational analysis in different species.

Taken together, our large-scale analysis of bacterial proteins of unknown function predicts functional associations for 1740 out of 7853 orthologous groups (22.2%), whereby 1466 of these, assign links purely among uncharacterized proteins, thus hinting at potentially novel functional modules. We can assign cellular processes, complexes or operons for 273 so far uncharacterized orthologous groups (i.e. for 3624 proteins). In contrast, homology searches reveal no functional information for these proteins, at best pointing to very poorly characterized domains [e.g. DUF-domains in the PFAM database (33)]. Thus, the results show the significance of context-based methods in function prediction and emphasize the complementarity to homology-based methods. Our study should be a first step in following the call for community action (1), and may help to pave the way for comprehensive protein function annotation.

SUPPLEMENTARY MATERIAL

Supplementary Material is available at NAR Online.

REFERENCES

1. Roberts,R.J. (2004) Identifying protein function—a call for community action. *PloS Biol*., **2**, E42.
2. Dandekar,T., Snel,B., Huynen,M. and Bork,P. (1998) Conservation of gene order: a fingerprint of proteins that physically interact. *Trends Biochem. Sci*., **23**, 324–328.
3. Enright,A.J., Iliopoulos,I., Kyrpides,N.C. and Ouzounis,C.A. (1999) Protein interaction maps for complete genomes based on gene fusion events. *Nature*, **402**, 86–90.
4. Marcotte,E.M., Pellegrini,M., Ng,H.L., Rice,D.W., Yeates,T.O. and Eisenberg,D. (1999) Detecting protein function and protein–protein interactions. *Science*, **285**, 751–753.
5. Overbeek,R., Fonstein,M., D'Souza,M., Pusch,G.D. and Maltsev,N. (1999) The use of gene clusters to infer functional coupling. *Proc. Natl Acad. Sci. USA*, **96**, 2896–2901.
6. Pellegrini,M., Marcotte,E.M., Thompson,M.J., Eisenberg,D. and Yeates,T.O. (1999) Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc. Natl Acad. Sci. USA*, **96**, 4285–4288.
7. von Mering,C., Huynen,M., Jaeggi,D., Schmidt,S., Bork,P. and Snel,B. (2003) STRING: a database of predicted functional associations between proteins. *Nucleic Acids Res*., **31**, 258–261.
8. Huynen,M., Snel,B., Lathe,W.,III and Bork,P. (2000) Predicting protein function by genomic context: quantitative evaluation and qualitative inferences. *Genome Res*., **10**, 1204–1210.
9. Mellor,J.C., Yanai,I., Clodfelter,K.H., Mintseris,J. and DeLisi,C. (2002) Predictome: a database of putative functional links between proteins. *Nucleic Acids Res*., **30**, 306–309.
10. Marcotte,E.M., Pellegrini,M., Thompson,M.J., Yeates,T.O. and Eisenberg,D. (1999) A combined algorithm for genome-wide prediction of protein function. *Nature*, **402**, 83–86.
11. Makarova,K.S. and Koonin,E.V. (2003) Comparative genomics of Archaea: how much have we learned in six years, and what's next? *Genome Biol*., **4**, 115.
12. von Mering,C., Zdobnov,E.M., Tsoka,S., Ciccarelli,F.D., Pereira-Leal,J.B., Ouzounis,C.A. and Bork,P. (2003) Genome evolution reveals biochemical networks and functional modules. *Proc. Natl Acad. Sci. USA*, **100**, 15428–15433.
13. Makarova,K.S., Wolf,Y.I. and Koonin,E.V. (2003) Potential genomic determinants of hyperthermophily. *Trends Genet*., **19**, 172–176.
14. Doerks,T., Andrade,M.A., Lathe,W.,III, von Mering,C. and Bork,P. (2004) Global analysis of bacterial transcription factors to predict cellular target processes. *Trends Genet*., **20**, 126–131.
15. Karaoz,U., Murali,T.M., Letovsky,S., Zheng,Y., Ding,C., Cantor,C.R. and Kasif,S. (2004) Whole-genome annotation by using evidence

integration in functional-linkage networks. *Proc. Natl Acad. Sci. USA*, **101**, 2888–2893.
16. Tatusov,R.L., Natale,D.A., Garkavtsev,I.V., Tatusova,T.A., Shankavaram,U.T., Rao,B.S., Kiryutin,B., Galperin,M.Y., Fedorova,N.D. and Koonin,E.V. (2001) The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res.*, **29**, 22–28.
17. Boeckmann,B., Bairoch,A., Apweiler,R., Blatter,M.C., Estreicher,A., Gasteiger,E., Martin,M.J., Michoud,K., O'Donovan,C., Phan,I., Pilbout,S. and Schneider,M. (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.*, **31**, 365–370.
18. Martin,P.R., Watson,A.A., McCaul,T.F. and Mattick,J.S. (1995) Characterization of a five-gene cluster required for the biogenesis of type 4 fimbriae in *Pseudomonas aeruginosa*. *Mol. Microbiol.*, **16**, 497–508.
19. Ogawa,H., Haga,T. and Toyoshima,C. (2000) Soluble P-type ATPase from an archaeon, *Methanococcus jannaschii*. *FEBS Lett.*, **471**, 99–102.
20. Bramkamp,M., Gassel,M., Herkenhoff-Hesselmann,B., Bertrand,J. and Altendorf,K. (2003) The *Methanocaldococcus jannaschii* protein Mj0968 is not a P-type ATPase. *FEBS Lett.*, **543**, 31–36.
21. Chitnis,P.R., Reilly,P.A. and Nelson,N. (1989) Insertional inactivation of the gene encoding subunit II of photosystem I from the *Cyanobacterium synechocystis* sp. PCC 6803. *J. Biol. Chem.*, **264**, 18374–18380.
22. Buttner,D., Gurlebeck,D., Noel,L.D. and Bonas,U. (2004) HpaB from *Xanthomonas campestris* pv. *vesicatoria* acts as an exit control protein in type III-dependent protein secretion. *Mol. Microbiol.*, **54**, 755–768.
23. Stader,J., Matsumura,P., Vacante,D., Dean,G.E. and Macnab,R.M. (1986) Nucleotide sequence of the *Escherichia coli* motB gene and site-limited incorporation of its product into the cytoplasmic membrane. *J. Bacteriol.*, **166**, 244–252.
24. Freudl,R., Klose,M. and Henning,U. (1990) Export and sorting of the *Escherichia coli* outer membrane protein OmpA. *J. Bioenerg. Biomembr.*, **22**, 441–449.
25. Bouveret,E., Benedetti,H., Rigal,A., Loret,E. and Lazdunski,C. (1999) *In vitro* characterization of peptidoglycan-associated lipoprotein (PAL)-peptidoglycan and PAL–TolB interactions. *J. Bacteriology*, **181**, 6306–6311.
26. Das,S., Chakrabortty,A., Banerjee,R. and Chaudhuri,K. (1999) Involvement of *in vivo* induced *icmF* gene of *Vibrio cholerae* in motility, adherence to epithelial cells, and conjugation frequency. *Biochem. Biophys. Res. Commun.*, **295**, 922–928.
27. Mintz,K.P. and Fives-Taylor,P.M. (2000) *impA*, a gene coding for an inner membrane protein, influences colonial morphology of *Actinobacillus actinomycetemcomitans*. *Infect. Immun.*, **68**, 6580–6586.
28. Shin,D.H., Yokota,H., Kim,R. and Kim,S.H. (2002) Crystal structure of conserved hypothetical protein Aq1575 from *Aquifex aeolicus*. *Proc. Natl Acad. Sci. USA*, **99**, 7980–7985.
29. West,S.C. (1997) Processing of recombination intermediates by the RuvABC Proteins. *Annu. Rev. Genet.*, **31**, 213–244.
30. Sharples,G.J., Yokota,H., Kim,R. and Kim,S.H. (1999) Holliday junction processing in bacteria: insights from the evolutionary conservation of RuvABC, RecG, and RusA. *J. Bacteriology*, **181**, 5543–5550.
31. Ingleston,S.M., Dickman,M.J., Grasby,J.A., Hornby,D.P., Sharples,G.J. and Lloyd,R.G. (2002) Holliday junction binding and processing by the RuvA protein of *Mycoplasma pneumoniae*. *Eur. J. Biochem.*, **269**, 1525–1533.
32. Gao,L.Y., Groger,R., Cox,J.S., Beverley,S.M., Lawson,E.H. and Brown,E.J. (2003) Transposon mutagenesis of *Mycobacterium marinum* identifies a locus linking pigmentation and intracellular survival. *Infect. Immun.*, **71**, 922–929.
33. Bateman,A., Birney,E., Cerruti,L., Durbin,R., Etwiller,L., Eddy,S.R., Griffiths-Jones,S., Howe,K.L., Marshall,M. and Sonnhammer,E.L. (2002) The Pfam Protein Families Database. *Nucleic Acids Res.*, **30**, 276–280.