Discovering Novel Enzymes and Pathways by Comparative Genomics

Valérie de Crécy-Lagard

Department of Microbiology and Department of Microbiology and Cell Science,
University of Florida, P.O. Box 110700, Gainesville, FL 32611 – 0700, U.S.A.

E-Mail: vcrecy@ufl.edu

Received: 31st July 2006 / Published: 31st August 2007

Abstract

Identifying the function of every gene in all sequenced organisms is one of the major challenges of the post-genomic era and is one of the obligate steps leading to systems biology approaches. This objective is far from being reached. By different estimates, over 30 – 50% of the genes of any given organism are of unknown function, incorrectly annotated or given a broad nonspecific annotation.

Most genome functional annotations programs rely on an homology based approach, using first simple Blast or FASTA scores then more elaborate, sensitive and precise algorithms stemming from the field of protein structure prediction. The inherent limitations of homology based approaches (only similar objects can be identified), has driven the development of non-homology based methods to link gene and function. Integrative genome mining tools that can analyse gene clustering, phylogenetic distribution, or protein fusions on a multi-genome scale have been developed recently. These bioinformatics tools allow the experimental biologist to make predictions on unknown gene function that can be tested experimentally and discover novel enzymes, regulators and transporters that expand our knowledge of metabolism in all species.

Introduction

The availability of nearly four hundred complete genomes (http://www.genomesonline.org/) has changed the way the experimental scientist generates hypothesis and identifies novel enzymes. The computer programming illiterate bench scientist has the unique possibility to link genes and function by combining comparative genomic tools that are freely available, with the experimental tools of physiology, genetics and enzymology. These approaches are leading to the discovery of novel enzymes and pathways of both fundamental and applied interest and also improve the general quality of genome annotations.

Towards a Complete Functional Analysis of Genomes: The Post-Genome Challenge

Identifying the function of every gene in all sequenced organisms is one of the major objectives of the post-genomic era, and one that is driving the development of systems biology [1]. This objective is far from reached as, by different estimates, 30 – 60% of the genes of any given organism have no assigned function [2 – 4]. As more genomes are being sequenced, the number of unknown genes and annotation errors are propagating at an alarming rate, making it increasingly difficult to extract correct functional information. Without a specific functional annotation effort, the genome information generated will become difficult to analyse and greatly underexploited [5].

Mining Genomes for New Enzymes

The availability of genomic sequence from both cultured and non-cultured organisms from diverse environments has had a great impact on the availability of enzymes that are better adapted for biocatalysis (for review see [6]). Also, “biochemical profiling” approaches [7 – 13] have been quite successful in identifying new enzymes [14, 15]. All these methods, however, rely on high throughput protein expression and enzymatic screens, and less labour intensive methods that are also more target specific are clearly needed to fully mine the catalytic potential of genomes. This is especially critical for implementing new biocatalytic activities into industrial processes. As Schmid *et al.* commented, “Future biocatalytic processes generally will not be limited by the available technology or the nature of the substrates and products. Instead, the feasibility of new biocatalytic processes will often be determined by the availability of the biocatalyst...” [16]. An untapped resource of novel catalysts is lying in the thousand of genes of unknown function that are now available at our fingertips if both bioinformatic and experimental methods can be combined to identify them.

Mining Genomes for New Antibacterial Targets

The need for the development of new antibiotics that escape common resistance mechanisms is becoming an acute public health problem. The World Health Organization (WHO) states that "In the race for supremacy, microbes are sprinting ahead" and "Microbial resistance could bring the world to a pre-antibiotic age (http://www.who.int/infectious-disease-report/2000/). The value of using genomics in anti-infective research was recognized early on by both fundamental and applied research enterprises (for review see [17]). Pipelines combining identification of bacterial genes essential for growth or virulence followed by structural efforts have been implemented and leads are starting to trickle out [18]. One major problem in this approach has been that targets identified from genomics approaches are often of unknown function, therefore no assay can be developed to screen for inhibitors.

Homology-Based Functional Analysis

Functional inferences based on comparative sequence analysis are well-established foundations of genomic annotation. The most significant advancements in this field over the last decade are directly related to the dramatic increase in the amount of sequenced genomes, as well as to the development of the robust and sensitive algorithms, such as FASTA, BLAST and their modifications (for the overview, see [19]). Domain analysis and reduction of the protein space via grouping of putative orthologues (such as COGs [20]) play an important role in the projection of functional assignments between diverse species. A significant contribution is provided by research communities focused on the detailed curation efforts of model organism genomes (e.g., *Escherichia coli* (http://bmb.med.miami.edu/EcoGene/EcoWeb), *Bacillus subtilis* (http://genome.jouy.inra.fr/cgi), *Saccharomyces cerevisiae* (http://www.yeast genome.org/index.html). For well studied gene families, in which the initial annotation has been experimentally verified, these homology based methods are quite accurate in predicting function [21]. However, factors such as poor homologies [21], multi-domain proteins [22], gene duplications [21, 23] and non-orthologous displacements [24] all contribute to incorrect or absent annotations that have accumulated and propagated, leading to the current poor functional annotation status of the genomic data [3, 4, 24, 25]. Furthermore, the inherent limitations of homology based approaches (only similar objects can be identified) require the development of non-homology based methods to link genes to function.

Beyond Homology, from Comparative Genomics to Experimental Verification

From gene to function

Systematic approaches such as structural genomics initiatives, systematic interaction mapping or systematic gene disruption combined with phenotypic screenings has led to the elucidation of some gene functions [26, 27]. Nearly 1000 structures have been deposited to date by structural genomics programs in the Protein Data Bank (http://www.rcsb.org/pdb/). However, examples where cellular functions were inferred directly from structural information are rare – in fact there are only a handful [28, 29]. Large-scale deletion mutant libraries have been completed for *S. cerevisiae* [30], *B. subtilis* [31] and in *E. coli* (http://ecoli.aist-nara.ac.jp/). Broad systematic phenotype screens [32] allowed the prediction of a few functions such as a missing histidine biosynthesis gene [33] or the discovery of anabolic and catabolic phosphorylating glyceraldehyde-3-phosphate dehydrogenises [34]. The real power of these libraries lies in using them in specialized screenings: this strategy has been successfully used in *S. cerevisiae* where novel cell cycle genes have been identified [35]. The road from phenotype to cellular function is often long and requires many downstream characterization steps [35]. Recently, "biochemical profiling" approaches consisting of testing the activity of all the proteins of a given genome (in pools or individually) in specific biochemical assays [7 – 13] or testing hundreds of proteins of unknown function in arrays covering a wide range of enzyme activities have been quite successful in correcting annotations or identifying functions of unknown genes [15, 36]. These large-scale efforts have not been as predictive as anticipated, but have been extremely valuable for the community in producing expression clones, mutants, structural and experimental data that can be used to predict and confirm functions as shown below.

From function to gene; Integration of genomic information

Large-scale cross-genomic integrations (such as NCBI [37], EMBL [38], TIGR [39], Uniprot [40]) provide important environments for extracting information from genomes. A dramatic enhancement of the quality and utility of genomic annotations is achieved by combining genome integration with metabolic reconstruction technology (see below). Among the key public resources supporting this approach are KEGG [41] and MetaCyc [42]. In these methods, genes are not analysed individually or as gene families but in a larger multi-genomic context. Additional information, not related to sequence homology, is gathered to help link gene and function (Fig. 1), and include:

- **Metabolic reconstruction:** by placing the genes in the context of the metabolic pathways found in a given organism, one can evaluate the biological relevance of an annotation [43, 44].
- **Clustering data:** Genes of a given pathway have a high probability of being physically linked on the chromosome [45].
- **Protein fusion events:** Genes of the same pathway can be fused to encode multidomain proteins in some organisms [46].

- **Phylogenetic occurrence profiles or signatures:** Presence/absence patterns of genes (or of set of genes) among genomes can be used to identify candidates for missing genes [47].
- **Shared regulatory sites:** Pathway genes are often regulated by a common protein recognizing a specific DNA sequence [48].
- **Functional and structural genomics:** Efforts provide additional clues to genome interpretation. The rapid increase in the volume and quality of such data, as well as their integration in publicly available repositories is expected to strongly impact gene and pathway analysis. Among the growing number of web-resources are: PDB, the best established collection of protein structures (http://www.rcsb.org/pdb/), SMD for expression data (http://genome-www5.stanford.edu/), DIP for protein–protein interactions (http://dip.doe-mbi.ucla.edu/).

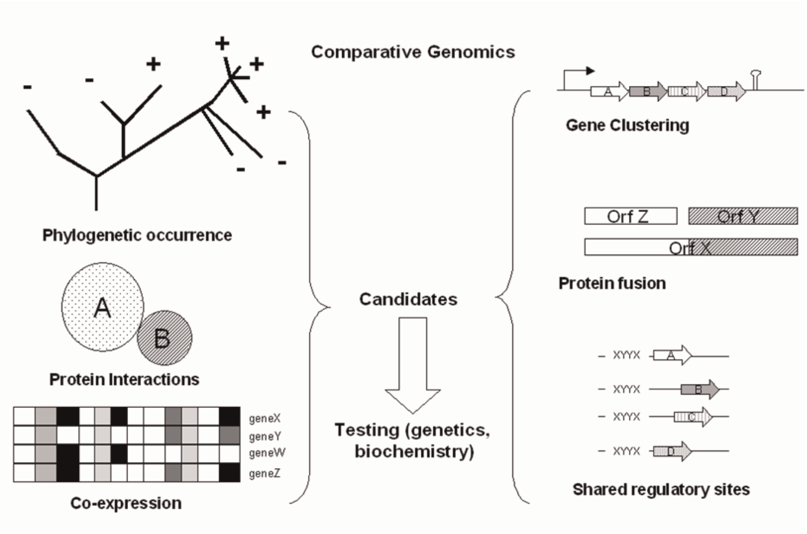

Figure 1. Comparative genomic strategies used to make predictions on gene function.

Early efforts to integrate different types of data to annotate genomes were developed by Koonin and colleagues based on the Cluster of Orthologs Groups (COG) database [49] that lists families of orthologues found in a subset of the sequenced genomes. In the last five years several integrated databases that contain phylogenetic occurrence profiles, clustering or protein fusion data and many combinations of the three have been implemented. These databases are all freely available with web-interfaces and include PhydBac [50], String [51], Microbial Database [52], Genomenet [[41], Plex [53], Cytoscape [54], Metacyc [42] and SEED [55] (Table 1). Genome researchers have built on this multi-tiered approach to help in calling gene function and reducing the number of errors as recently described for the genomes of *Haloarcula marismortui* [56] and *Methylococcus capsulatus* (Bath) [28]. The combination of structural information and comparative genomic methods has led to many robust predictions [1,29].

Table 1. Freely available comparative genomic analysis websites.

Name	Location
Cluster of Orthologous Groups	http://www.ncbi.nlm.nih.gov/COG/
FusionDB and PhydBac	http://igs-server.cnrs-mrs.fr/phydbac/
	http://igs-server.cnrs-mrs.fr/FusionDB/
TIGR-CMR	http://cmr.tigr.org/tigr-scripts/CMR/CmrHomePage.cgi
	http://www.tigr.org/tigr-scripts/CMR2/GenomeSlicer.spl
STRING	http://dag.embl-heidelberg.de/newstring_cgi/show_input_page.pl
IMG	http://img.jgi.doe.gov/cgi-bin/pub/main.cgi
Cytoscape	http://www.cytoscape.org/
GenomeNet and KEGG	http://www.genome.ad.jp/
Protein Link Explorer (Plex)	http://apropos.icmb.utexas.edu/plex/plex.html
MetaCyc	http://metacyc.org/
SEED	http://theseed.uchicago.edu/FIG/

Using Comparative Genomics to Link Genes to Function

Although the field of comparative genomics is still young, these tools have allowed the genetic characterization of a number of critical metabolic pathways that had eluded scientific inquiry for decades and an estimated 100 gene families have been identified successfully using comparative genomic methods to date [57] (Ross Overbeek, personal communication). For example, predictions based exclusively on phylogenetic occurrence profiles resulted in the identification of the last steps of the non-mevalonate isoprenoid pathway [58]. Protein fusion analysis allowed the identification of missing Coenzyme A biosynthesis genes in *Homo sapiens* [59]. Chromosome clustering analysis revealed a missing fatty acid synthesis gene (target of antibacterial compounds) in *Streptococcus pneumoniae* or missing genes in folate biosynthesis [60, 61]. A combination of approaches was used to identify the diverse NAD recycling pathways of cyanobacteria [62]. A search for regulator sites allowed the identification of many missing thiamine biosynthesis genes [63], metal transporters [64] or decipher the N-acetylglucosamine utilization pathway of *Shewanella oneidensis* [65]. The approach has been particularly productive in discovering missing and novel enzymes in Archaea because of the originality of their metabolic solutions and the recent availability of 30 archaeal genomes [66].

Applying these comparative genomic methods to the field of tRNA modification and coenzyme metabolism has allowed us to identify the function of eight enzyme families, unravelling novel enzyme activities, cases of orthologous displacements, novel pathways and potential drug targets (Table 2).

Table 2. Novel genes and pathways identified using comparative genomics techniques.

Functional Role	Novelty	Verified in	Key evidence
Pantetheine-phosphate adenylyltransferase/ dephospho–CoA kinase [67]		B [67]	Fusion
Carbamoyl-threonnyl-Adenosine syntase [a]	Potential target	E [a]	Occurrence profile/Structure
Wyeosine synthase [68]	Novel enzyme	E [68]	Occurrence profile
tRNA dihydrouridine synthase [69]	Novel enzyme	B [70]	Occurrence profile/operon
Queosine/Archeosine biosynthesis YkvJ	Novel enzyme	B [71]	Occurrence profile/operon
Queosine/Archeosine biosynthesis YkvK	Novel Enzyme	B [71]	Occurrence profile/operon
Queosine/Archeosine biosynthesis YkvL	Novel enzyme	B [71]	Occurrence profile/operon
PreQ0 reductase YkvM	Novel enzyme	B [72]	Occurrence profile/operon
GTP Cyclohydrolase I	Potential Target	B [73]	Occurrence profile/operon

B = Bacteria, E = Eukaryotes . [a] de Crécy-Lagard and collaborators (unpublished results)

Conclusion

This body of work opens the problem of how to name enzymes discovered through comparative genomics methods and give them EC numbers, as in general these enzymes have been very poorly described or were totally unknown. The number of enzymes discovered by these methods is steadily increasing and guidelines for the "gene discoverers" who are often not enzymologists need to be defined.

References

[1] Bonneau, R., Baliga, N.S., Deutsch, E.W., Shannon, P.,Hood, L. (2004) Comprehensive de novo structure prediction in a systems-biology context for the archaea *Halobacterium* sp. NRC-1. *Genome Biol.* **5:**R52.

[2] Siew, N., Azaria, Y., Fischer, D. (2004) The ORFanage: an ORFan database *Nucleic Acids Res.* **32 Database issue:** D281–283.

[3] Brenner, S.E. (1999) Errors in genome annotation. *Trends Genet.* **15:**132–133.

[4] Devos**,** D**.,** Valencia**,** A**.** (2001) Intrinsic errors in genome annotation. *Trends Genet.* **17:**429–431.

[5] Roberts, R.J. (2004) Identifying protein function–a call for community action. *PLoS Biol* **2:**E42.

[6] Ferrer, M., Martinez-Abarca, F., Golyshin, P.N. (2005) Mining genomes and 'metagenomes' for novel catalysts. *Curr. Opin. Biotechnol.* **16:**588–593.

[7] Grayhack, E.J., Phizicky, E.M. (2001) Genomic analysis of biochemical function. *Curr. Opin. Chem. Biol.* **5:**34–39.

[8] Gu, W., Jackman, J.E., Lohan, A.J., Gray, M.W., Phizicky, E.M. (2003) tRNA^{His} maturation: an essential yeast protein catalyzes addition of a guanine nucleotide to the 5' end of tRNA^{His}. *Genes Devel.* **17:**2889–2901.

[9] Jackman, J.E., Montange, R.K., Malik, H.S., PPhizicky, E.M. (2003) Identification of the yeast gene encoding the tRNA m^{1}G methyltransferase responsible for modification at position 9. *RNA* **9:**574–585.

[10] Phizicky, E.M., Martzen, M.R., McCraith, S.M., Spinelli, S.L., Xing, F., Shull,N.P., Van Slyke, C., Montagne, R.K., Torres, F.M., Fields, S., Grayhack, E.J. (2002) Biochemical genomics approach to map activities to genes. *Methods Enzymol* **350:**546–559.

[11] Polevoda, B., Martzen, M.R., Das, B., Phizicky, E.M., Sherman, F. (2000) Cytochrome c methyltransferase, Ctm1 p, of yeast. *J. Biol. Chem.* **275:**20508–20513.

[12] Steiger, M.A., Kierzek, R., Turner, D.H., Phizicky, E.M. (2001) Substrate recognition by a yeast 2'-phosphotransferase involved in tRNA splicing and by its *Escherichia coli* homolog. *Biochemistry* **40:**14098–14105.

[13] Xing, F., Martzen, M.R., Phizicky, E.M. (2002) A conserved family of *Saccharomyces cerevisiae* synthases effects dihydrouridine modification of tRNA. *RNA* **8:**370–381.

[14] Kutznetosva, E., Proudfoot, M., Sanders, S.A., Reinking, J., Savchenko, A. *et al.* (2005) Enzyme genomics: Application of general enzymatic screens to discover new enzymes. *FEMS Microbiol. Rev.* **29**(2)**:** 263–279.

[15] Martzen, M.R., McCraith, S.M., Spinelli, S.L., Torres, F.M., Fields, S., Grayhack, E.J., Phizicky, E.M. (1999) A biochemical genomics approach for identifying genes by the activity of their products. *Science* **286:**1153–1155.

[16] Schmid, A., Dordick, J.S., Hauer, B., Kiener, A., Wubbolts, M., Witholt, B. (2001) Industrial biocatalysis today and tomorrow. *Nature* **409:**258–268.

[17] Haney, S.A., Alksne, L.E., Dunman, P.M., Murphy, E., Projan, S.J. (2002) Genomics in anti-infective drug discovery–getting to endgame. *Curr. Pharm. Des.* **8:**1099–1118.

[18] Schmid, M.B. (2004) Seeing is believing: the impact of structural genomics on antimicrobial drug discovery. *Na ure Rev. Microbiol.* **2:**739–746.

[19] Koonin, E.V., Galperin, M.Y. (2002) *SEQUENCE – EVOLUTION – FUNCTION. Computational Approaches in Comparative Genomics*. 488 pp. Kluwer Academic Publishers, Boston.

[20] 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., Koonin, E.V. (2001) The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res.* **29:**22–28.

[21] Tian, W., Skolnick, J. (2003) How well is enzyme function conserved as a function of pairwise sequence identity? *J. Mol. Biol.* **333:**863–882.

[22] Hegyi, H., Gerstein, M. (2001) Annotation transfer for genomics: measuring functional divergence in multi-domain proteins. *Genome Res.* **11:**1632–1640.

[23] Gerlt, J.A., Babbitt, P.C. (2000) Can sequence determine function? *Genome Biol.* **1:**REVIEWS 0005.

[24] Galperin, M.Y., Koonin, E.V. (1998) Sources of systematic error in functional annotation of genomes: domain rearrangement, non-orthologous gene displacement and operon disruption. *In Silico Biol.* **1:**55–67.

[25] Attwood, T.K. (2000) Genomics. The Babel of bioinformatics. *Science* **290:**471–473.

[26] Mittl, P.R., Grutter, M.G. (2001) Structural genomics: opportunities and challenges. *Curr. Opin. Chem. Biol.* **5:**402–408.

[27] Huynen, M.A., Snel, B., van Noort, V. (2004) Comparative genomics for reliable protein-function prediction from genomic data. *Trends Genet.* **20:**340–344.

[28] Yakunin, A.F., Yee, A.A., Savchenko, A., Edwards, A.M., Arrowsmith, C.H. (2004) Structural proteomics: a tool for genome annotation. *Curr. Opin. Chem. Biol.* **8:**42–48.

[29] Zhang, C., Kim, S.H. (2003) Overview of structural genomics: from structure to function. *Curr. Opin. Chem. Biol.* **7:**28–32.

[30] Winzeler, E.A., Shoemaker, D.D., Astromoff, A., Liang, H., Anderson, K., Andre, B., Bangham, R., Benito, R., Boeke, J.D., Bussey, H., Chu, A.M., Connelly, C., Davis, K., Dietrich, F., Dow, S.W., *et al*. (1999) Functional characterization of the *S. cerevisiae* genome by gene deletion and parallel analysis. *Science* **285:**901–906.

[31] Kobayashi, K., Ehrlich, S.D., Albertini, A., Amati, G., Andersen, K.K., Arnaud, M., Asai, K., Ashikaga, S., Aymerich, S., Bessieres, P., Boland, F., Brignell, S.C., Bron, S., Bunai, K., Chapuis, J., *et al*. (2003) Essential *Bacillus subtilis* genes. *Proc. Natl Acad. Sci. U S A* **100:**4678–4683.

[32] Ogasawara, N. (2000) Systematic function analysis of *Bacillus subtilis* genes. *Res. Microbiol.* **151:**129–134.

[33] le Coq, D., Fillinger, S., Aymerich, S. (1999) Histidinol phosphate phosphatase, catalyzing the penultimate step of the histidine biosynthesis pathway, is encoded by *ytvP* (*hisJ*) in *Bacillus subtilis*. *J. Bacteriol.* **181:**3277–3280.

[34] Fillinger, S., Boschi-Muller, S., Azza, S., Dervyn, E., Branlant, G., Aymerich, S. (2000) Two glyceraldehyde-3-phosphate dehydrogenases with opposite physiological roles in a nonphotosynthetic bacterium. *J. Biol. Chem.* **275:**14031–14037.

[35] Carpenter, A.E., Sabatini, D.M. (2004) Systematic genome-wide screens of gene function. *Nature Rev. Genet.* **5:**11–22.

[36] Kuznetsova, E., Proudfoot, M., Sanders, S.A., Reinking, J., Savchenko, A., Arrowsmith, C.H., Edwards, A.M., Yakunin, A.F. (2005) Enzyme genomics: Application of general enzymatic screens to discover new enzymes. *FEMS Microbiol. Rev.* **29:**263–279.

[37] Wheeler, D.L., Barrett, T., Benson, D.A., Bryant, S.H., Canese, K., Chetvernin, V., Church, D.M., DiCuccio, M., Edgar, R., Federhen, S., Geer, L.Y., Helmberg, W., Kapustin, Y., Kenton, D.L., Khovayko, O., *et al*. (2006) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **34:**D 173–180.

[38] Cochrane, G., Aldebert, P., Althorpe, N., Andersson, M., Baker, W., Baldwin, A., Bates, K., Bhattacharyya, S., Browne, P., van den Broek, A., Castro, M., Duggan, K., Eberhardt, R., Faruque, N., Gamble, J., *et al*. (2006) EMBL Nucleotide Sequence Database: developments in 2005. *Nucleic Acids Res.* **34:**D 10–15.

[39] Pertea, G., Huang, X., Liang, F., Antonescu, V., Sultana, R., Karamycheva, S., Lee, Y., White, J., Cheung, F., Parvizi, B., Tsai, J., Quackenbush, J. (2003) TIGR Gene Indices clustering tools (TGICL): a software system for fast clustering of large EST datasets. *Bioinformatics* **19:**651–652.

[40] Wu, C.H., Apweiler, R., Bairoch, A., Natale, D.A., Barker, W.C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., Martin, M.J., Mazumder, R., O'Donovan, C., Redaschi, N., *et al*. (2006) The Universal Protein Resource (UniProt): an expanding universe of protein information. *Nucleic Acids Res.* **34:**D 187–191.

[41] Kanehisa, M., Goto, S., Hattori, M., Aoki-Kinoshita, K.F., Itoh, M., Kawashima, S., Katayama, T., Araki, M., Hirakawa, M. (2006) From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res.* **34:**D 354–357.

[42] Krieger, C.J., Zhang, P., Mueller, L.A., Wang, A., Paley, S., Arnaud, M., Pick, J., Rhee, S.Y., Karp, P.D. (2004) MetaCyc: a multiorganism database of metabolic pathways and enzymes. *Nucleic Acids Res.* **32 Database issue:** D 438–442.

[43] Selkov, E., Maltsev, N., Olsen, G.J., Overbeek, R., Whitman, W.B. (1997) A reconstruction of the metabolism of Methanococcus jannaschii from sequence data. *Gene* **197:**GC 11–26.

[44] Galperin, M.Y., Brenner, S.E. (1998) Using metabolic pathway databases for functional annotation. *Trends Genet.* **14:**332–333.

[45] Overbeek, R., Fonstein, M., D'Souza, M., Pusch, G.D., Maltsev, N. (1999) The use of gene clusters to infer functional coupling. *Proc. Natl Acad. Sci. U S A* **96:**2896–2901.

[46] Enright, A.J., Iliopoulos, I., Kyrpides, N.C., Ouzounis, C.A. (1999) Protein interaction maps for complete genomes based on gene fusion events. *Nature* **402:**86–90.

[47] Pellegrini, M., Marcotte, E.M., Thompson, M.J., Eisenberg, D., Yeates, T.O. (1999) Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc. Natl Acad. Sci. U S A* **96:**4285–4288.

[48] Gelfand, M.S., Novichkov, P.S., Novichkova, E.S., Mironov, A.A. (2000) Comparative analysis of regulatory patterns in bacterial genomes. *Brief Bioinform.* **1:**357–371.

[49] Natale, D.A., Galperin, M.Y., Tatusov, R.L., Koonin, E.V. (2000) Using the COG database to improve gene recognition in complete genomes. *Genetica* **108:**9–17.

[50] Enault, F., Suhre, K., Poirot, O., Abergel, C., Claverie, J.M. (2004) Phydbac2: improved inference of gene function using interactive phylogenomic profiling and chromosomal location analysis. *Nucleic Acids Res.* **32:**W336–339.

[51] von Mering, C., Jensen, L.J., Snel, B., Hooper, S.D., Krupp, M., Foglierini, M., Jouffre, N., Huynen, M.A., Bork, P. (2005) STRING: known and predicted protein-protein associations, integrated and transferred across organisms. *Nucleic Acids Res.* **33:**D433–437.

[52] Haft, D.H., Selengut, J.D., Brinkac, L.M., Zafar, N., White, O. (2005) Genome Properties: a system for the investigation of prokaryotic genetic content for microbiology, genome annotation and comparative genomics. *Bioinformatics* **21:**293–306.

[53] Date, S.V., Marcotte, E.M. (2005) Protein function prediction using the Protein Link EXplorer (PLEX). *Bioinformatics* **21:**2558–2559.

[54] Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B., Ideker, T. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13:**2498–2504.

[55] Overbeek, R., Begley, T., Butler, R.M., Choudhuri, J.V., Chuang, H.Y., Cohoon, M., de Crécy-Lagard, V., Diaz, N., Disz, T., Edwards, R., Fonstein, M., Frank, E.D., Gerdes, S., Glass, E.M., Goesmann, A., *et al*. (2005) The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. *Nucleic Acids Res.* **33:**5691–5702.

[56] Baliga, N.S., Bonneau, R., Facciotti, M.T., Pan, M., Glusman, G., Deutsch, E.W., Shannon, P., Chiu, Y., Weng, R.S., Gan, R.R., Hung, P., Date, S.V., Marcotte, E., Hood, L., Ng, W.V. (2004) Genome sequence of *Haloarcula marismortui*: a halophilic archaeon from the Dead Sea. *Genome Res.* **14:**2221–2234.

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

[58] Smit, A., Mushegian, A. (2000) Biosynthesis of isoprenoids via mevalonate in Archaea: the lost pathway. *Genome Res.* **10:**1468–1484.

[59] Daugherty, M., Polanuyer, B., Farrell, M., Scholle, M., Lykidis, A., de Crécy-Lagard, V., Osterman, A. (2002) Complete reconstitution of the human coenzyme A biosynthetic pathway via comparative genomics. *J. Biol. Chem.* **277:**21431–21439.

[60] Klaus, S.M., Kunji, E.R., Bozzo, G.G., Noiriel, A., de la Garza, R.D., Basset, G.J., Ravanel, S., Rebeille, F., Gregory, J.F., 3rd, Hanson, A.D. (2005) Higher plant plastids and cyanobacteria have folate carriers related to those of trypanosomatids. *J. Biol. Chem.* **280:**38457–38463.

[61] Klaus, S.M., Wegkamp, A., Sybesma, W., Hugenholtz, J., Gregory, J.F., 3rd, Hanson, A.D. (2005) A nudix enzyme removes pyrophosphate from dihydroneopterin triphosphate in the folate synthesis pathway of bacteria and plants. *J. Biol. Chem.* **280:**5274–5280.

[62] Gerdes, S.Y., Kurnasov, O.V., Shatalin, K., Polanuyer, B., Sloutsky, R., Vonstein, V., Overbeek, R., Osterman, A.L. (2006) Comparative genomics of NAD biosynthesis in Cyanobacteria. *J. Bacteriol.* **188:**3012–3023.

[63] Rodionov, D.A., Vitreschak, A.G., Mironov, A.A., Gelfand, M.S. (2002) Comparative genomics of thiamin biosynthesis in procaryotes: new genes and regulatory mechanisms. *J. Biol. Chem.***277:**48949–48959.

[64] Rodionov, D.A., Hebbeln, P., Gelfand, M.S., Eitinger, T. (2006) Comparative and functional genomic analysis of prokaryotic nickel and cobalt uptake transporters: evidence for a novel group of ATP-binding cassette transporters. *J. Bacteriol.* **188:**317–327.

[65] Yang, C., Rodionov, D.A., Li, X., Laikova, O.N., Gelfand, M.S., Zagnitko, O.P., Romine, M.F., Obraztsova, A.Y., Nealson, K.H., Osterman, A.L. (2006) Comparative genomics and experimental characterization of N-acetylglucosamine utilization pathway of *Shewanella oneidensis*. *J. Biol. Chem.* M605052200.

[66] Ettema, T.J., de Vos, W.M., van der Oost, J. (2005) Discovering novel biology by in silico archaeology. *Nature Rev. Microbiol.* **3:**859–869.

[67] Daugherty, M., Polanuyer, B., Farrell, M., Scholle, M., Lykidis, A., de Crecy-Lagard, V., Osterman, A. (2002) Complete reconstitution of the human coenzyme A biosynthetic pathway via comparative genomics. *J. Biol. Chem.* **277:**21431–21439.

[68] Waas, W.F., de Crécy-Lagard, V. Schimmel, P. (2005) Discovery of a gene family critical to wyosine base formation in a subset of phenylalanine-specific transfer RNAs. *J. Biol. Chem.* **280:**37616–37622.

[69] Bishop, A.C., Xu, J., Johnson, R.C., Schimmel, P., de Crécy-Lagard, V. (2002) Identification of the tRNA-dihydrouridine synthase family. *J. Biol. Chem.* **277:**25090–25095.

[70] Bishop, A.C., Xu, J., Johnson, R.C., Schimmel, P., de Crécy-Lagard, V. (2002) Identification of the tRNA-dihydrouridine synthase family. *J. Biol. Chem.* **277:**25090–25095.

[71] Reader, J.S., Metzgar, D., Schimmel, P., de Crécy-Lagard, V. (2004) Identification of four genes necessary for biosynthesis of the modified nucleoside queuosine. *J. Biol. Chem.* **279:**6280–6285.

[72] Van Lanen, S.G., Reader, J.S., Swairjo, M.A., de Crécy-Lagard, V., Lee, B., Iwata-Reuyl, D. (2005) From cyclohydrolase to oxidoreductase: discovery of nitrile reductase activity in a common fold. *Proc. Natl Acad. Sci. U S A* **102:**4264–4269.

[73] El Yacoubi, B., Bonnett, S., Anderson, J.N., Swairjo, M.A., Iwata-Reuyl, D., de .Crécy Lagard, V. (2006) Discovery of a new prokaryotic type I GTP cyclohydrolase family. *J. Biol. Chem.* **281**(49)**:** 37586–37593.