DDT Vol. 6, No. 15 (Suppl.), 2001

Novel computational methods in anti-microbial target identification

Christoph Freiberg

The large number of microbial genome sequences currently available facilitates completely new methods of cross-genome analyses, which aid in target identification and prioritization for anti-microbial drug discovery. This article provides a review of novel computational methods in this context. Genome comparison methods are described that are suitable for the generation of a 'periodic table of microbial functions' that will help to detect novel cellular functions and metabolic networks. The methods discussed include novel approaches such as ortholog/paralog clustering, phylogenetic profiling, metabolic pathway mapping, gene-neighbor and domain-fusion analyses, differential genome analysis, and co-expression profiling.

Christoph Freiberg
Pharma Research
Anti-infectives 1
Bayer AG
42096 Wuppertal
Germany
tel: +49 202 368461
fax: +49 202 364116
e-mail:
christoph.freiberg.cf@
bayer-ag.de

▼ The genomics era is enabling pharmaceutical companies to apply new research strategies to the discovery of novel anti-microbial drugs. Conventional antibiotics are becoming less effective in treating microbial infections owing to the spread of resistant microbial strains[1], and are known to target only a small number of essential functions in microbes[2]. Thus, identification of novel targets is one way to find new antibiotic chemotypes by screening large chemical libraries for inhibitors.

Target-based drug discovery in the genomics era

The large number of microbial genome sequences already completely deciphered enables cross-genome analyses to an extent which is not yet possible from the limited number of genomes from higher organisms, even with the first draft of the human genome[3,4]. By comparing different microbial genomes, it is possible to select potential targets for the discovery of broad- and narrow-spectrum antibiotics more comprehensively than before.

Microbial genome data and their annotation quality

Fifty-one microbial genomes are already completely annotated and deposited in the Genomes database of the National Center of Biotechnology Information (NCBI, Bethesda, MD, USA; http://www.ncbi.nlm.nih.gov:80/PMGifs/Genomes/micr.html), the main public resource for genomic sequence data. More than 100 additional microbial genomes are currently being sequenced and annotated. Unfortunately, there are no standards for 'complete annotation' of a genomic sequence. Typically, the location of genes (which mainly represent the protein coding regions) and the description (showing which coding region resembles which protein in the databases) are annotated. However, after submission to the public databases, the annotations of genomic sequences are rarely updated. That is why outdated information from one genome is often transferred to another by automated similarity searches, thereby contaminating the public databases.

Towards a periodic table of microbial functions

Finding homologs

Bacterial genomes are between 0.6 and 8.0 megabases (Mb) in size and generally encode 600–6,000 proteins. Currently, even though increasing numbers of genes can be functionally annotated, approx 30% of a microbial genome still consists of so-called hypothetical or orphan genes. Comparison of the complete sets of proteins encoded by pathogens is the initial key step in the

1359-6446/01/$ – see front matter PII: S1471-1931(01)00167-7

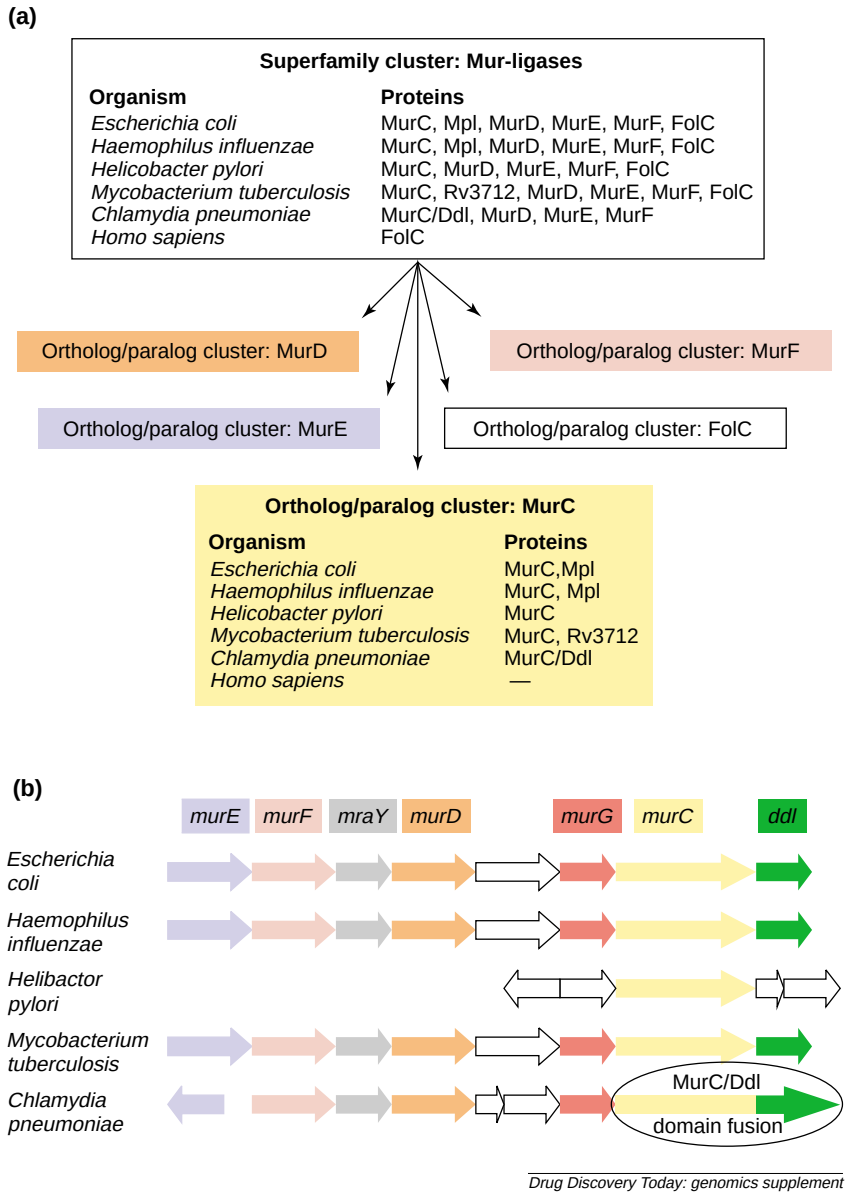


Figure 1. (a) From superfamily clusters to ortholog/paralog clusters. Structural and functional relationships of proteins can be detected on the primary amino acid sequence level. For instance, the listed gene products of the Mur-ligase superfamily, derived from various organisms, share common motifs (Pfam family PF01225; http://www.sanger.ac.uk/Software/Pfam/index.shtml). These proteins ligate amino acids accompanied with ATP hydrolysis [Enzyme Commission (EC) number 6.2.3.-]. The proteins MurC, MurD, MurE, MurF, Mpl are cytoplasmic peptidoglycan synthetases, and FolC is the folylpolyglutamate synthase. Genome comparison enables the classification of these into groups of proteins with identical functions (i.e. orthologs). The MurC, MurD, MurE and MurF orthologs exist in bacteria, while the FolC orthologs occur in both bacteria and humans. Some organisms possess homologs of MurC, called Mpl in *Escherichia coli* and *Haemophilus influenzae* and Rv3712 in *Mycobacterium tuberculosis*. Mpl, which is involved in the recycling of the cell wall component peptidoglycan, is a paralog of the peptidoglycan synthetase MurC. Ortholog/paralog subclustering of protein superfamilies is a crucial step towards a 'periodic table' of microbial functions. In this example, the superfamily and ortholog/paralog clusters were generated by the GD Phylosopher™ (GeneData; Basel, Switzerland). (b) Gene clusters and domain fusion illustrating functional coupling of peptidoglycan biosynthesis genes. The gene maps of the chromosomal regions around *murC* in various bacteria are shown. The genes are represented as arrows indicating their transcriptional orientation. Colored genes encode proteins involved in the biosynthesis of the cell wall constituent peptidogylcan: blue, MurE; pink, MurF; grey, MraY; orange, MurD; red, MurG; yellow, MurC; green, Ddl. Functionally related genes can often be found to be spatially clustered in certain bacteria. In *Chlamydia pneumoniae* the MurC- and Ddl-domain, which belong to the same pathway, are actually fused. The analyses were performed using GD Phylosopher™.

discovery of new anti-microbial targets – although it should not be forgotten that RNA molecules (such as ribosomal RNA) also represent binding sites for antibiotics. A prerequisite for such analyses is the classification of proteins, derived from the different organisms, into groups of proteins which probably fulfil the same or similar function, leading to a periodic table of microbial functions represented by amino acid sequence stretches.

The basic idea behind protein classification systems are sequence similarity search algorithms. The most commonly used programs are BLAST (Ref. 5) and FASTA (Ref. 6). These software tools have been further developed to be more sensitive, for example the iterative search algorithm PSI-BLAST (Ref. 5), which enables the detection of distant relationships among sequences. Several research groups have tried to group protein sequences into families. Some build groups of homologs based on entire protein sequences, while others extract essential features from protein families and build superfamilies that are described by motifs[7]. Many microbial genome databases enable easy identification of homologs in the different organisms and provide additional information about motif and superfamily assignment. Typical examples of such databases are PEDANT and CMR (Comprehensive Microbial Resource) (Table 1). Nevertheless, simple similarity and motif search-based classification approaches reach their limits when a comprehensive overview of which biological functions are really present in each organism is required.

Table 1. Computational tools and services in the field of microbial genome comparison

Computational tool/service	Description and comment	URL
Clusters of Orthologous Groups (COGs)	Database providing a phylogenetic classification of proteins encoded in complete genomes. A valuable large-scale implementation of ortholog/paralog clustering without using strict similarity search cut-offs.	http://www.ncbi.nlm.nih.gov/COG/
Comprehensive Microbial Resource (TIGR CMR)	This database of The Institute for Genomic Research (TIGR, Rockville, MD, USA) provides access to all bacterial genome sequences completed to date and includes systematic annotations of proteins. Protein families are generated based on motifs derived from multiple alignments of homologs (TIGRFAMS).	http://www.tigr.org/tigr-scripts/CMR2/CMRHomePage.spl
EcoCyc/MetaCyc	Databases describing pathways, reactions and enzymes of *Escherichia coli* and a variety of other organisms (products of DoubleTwist, Oakland, CA, USA). EcoCyc is frequently used as a reference pathway database for microbial metabolism.	http://ecocyc.pangeasystems.com/
The Enhanced Microbial Genomes Library (EMGLIB)	This database is suitable for searching gene names, annotation keywords and the like, in completely sequenced microbial genomes.	http://pbil.univ-lyon1.fr/emglib/emglib.html
Entrez Genomes/Microbial Genomes – list of projects	These databases contain the most comprehensive lists of microbial genome sequencing projects.	http://www.ncbi.nlm.nih.gov:80/PMGifs/Genomes/bact.html
Genome OnLine Database (GOLD)		http://wit.integratedgenomics.com/GOLD/
TIGR Microbial Database		http://www.tigr.org/tdb/mdb/mdbcomplete.html
ERGO/WIT	The computational system ERGO extensively integrates data generated from various methods in the field of large-scale comparative genome analysis and is actively developed by Integrated Genomics (Chicago, IL, USA). It is an extended version of the WIT database.	http://wit.integratedgenomics.com/ERGO/ http://wit.mcs.anl.gov/WIT2/
GD Phylosopher™	GD Phylosopher™ represents an enterprise solution for extensive integration of data and modern methods in the field of large-scale comparative genome analysis, and is actively developed by GeneData (Basel, Switzerland). The system also provides modules for the analysis of gene expression data and the identification of regulatory genetic elements.	http://www.genedata.com/products/phylosopher/
GeneQuiz	This software is an integrated system for large-scale biological sequence analysis providing complete annotation of genomes, but limited options for cross-genome queries.	http://jura.ebi.ac.uk:8765/ext-genequiz/
genomeSCOUT™	genomeSCOUT™ represents an enterprise solution for integration of genome comparison tools and is actively developed by LION Bioscience (Heidelberg, Germany).	http://www.lionbioscience.com/

Table 1. (cont'd)

Computational tool/service	Description and comment	URL
Genome Information Broker for microbial genomes (GIB)	This database provides a 'comparative genomes information retrieval system' focussing on names and annotation of proteins.	http://gib.genes.nig.ac.jp/
Microbial Genomes Blast Databases	Similarity searches against genome sequences of the most comprehensive set of unfinished sequencing projects can be performed.	http://www.ncbi.nlm.nih.gov/Microb_blast/unfinishedgenome.html
PathDB	This database enables inspection of biochemical pathways and metabolism of various organisms.	http://www.ncgr.org/software/pathdb/
The High quality Automated Microbial Annotation of Proteomes project (HAMAP)	This project aims to automatically annotate microbial protein sets in the SWISS-PROT database. The intention of this project is not directly to provide software or interactive genome comparison tools.	http://www.expasy.ch/sprot/hamap/
HOmologous BACterial GENes database (HOBACGEN)	This database contains all the proteins of completely sequenced bacteria, organized into families of homologs.	http://pbil.univ-lyon1.fr/databases/hobacgen.html
Kyoto Encyclopedia of Genes and Genomes (KEGG)	This database includes commonly used metabolic pathway maps.	http://www.genome.ad.jp/kegg/
Microbial Genome Database for comparative analysis (MBGD)	MBGD is a database workbench system for comparative analysis of completely sequenced microbial genomes. Currently, the function of MBGD is to create homology–orthology clusters of homologs and orthologs.	http://mbgd.genome.ad.jp/
PathoGenome™	This computational system includes public and proprietary sequence information related to significant microbial organisms and fungi. It is a product of Genome Therapeutics (Waltham, MA, USA).	http://www.genomecorp.com/programs/pathogenome.shtml http://www.labonweb.com/
PEDANT/PEDANT-PRO™	This is a computational system for comprehensive analysis and annotation of protein sequence sets with implemented cross-genome categories. This system is actively developed by Biomax Informatics (Munich, Germany).	http://pedant.mips.biochem.mpg.de/index.html http://www.biomax.de/
ProteinPathways Discovery Engine	ProteinPathways (Los Angeles, CA, USA) provides services to elaborate pathways in human pathogens. It uses the ProteinPathways Discovery Engine, a system comprising a combination of genome-comparison methods.	http://www.proteinpathways.com/
Software for the Examination, Exploration and Broad Understanding of Genome Sequences (SEEBUGS)	This database contains the proteins of completely sequenced bacteria organized into families of homologs and orthologs and enables cross-genome analysis.	http://www.congenomics.com/
STD Sequence Databases	This database focuses on genomes derived from sexually transmitted disease pathogens.	http://www.stdgen.lanl.gov/

Finding orthologs and paralogs

A promising *in silico* approach to predict the biochemical and cellular function of a protein is the identification of its characterized ortholog from a different species. Sequences derived from two different organisms with the highest level of similarity to each other can be considered to be orthologous[8]. Nevertheless, such a definition sometimes leads to false-positive results and is dependent on certain arbitrary cut-off scores. Additional proteins that represent identical or slightly changed functions have evolved in various organisms. Such homologs within organisms are called paralogs. In order to easily identify orthologs and the number of paralogs in microbes, a valuable large-scale method has been published by Tatusov *et al.* (Ref. 9), which automatically generates clusters of orthologous groups (COGs). These clusters, including orthologs and paralogs from the different microbes, are deposited in the COG database at the NCBI (Table 1). This method fundamentally requires at least three organisms to delineate orthologous relationships.

Some genome comparison tools now include the clustering algorithms for ortholog/paralog classification, or have integrated COG databases. All such classification systems are important refinements over homology-based classifications. These systems reflect the phylogenetic relationships between gene products and even detect weak similarities between proteins of distantly related organisms. Nevertheless, the methods determining orthology/paralogy have to be precisely defined. Therefore, clustering algorithms that do not strongly rely on strict similarity search cut-offs are far better. Although it will be difficult to develop an automatic clustering algorithm that will generate a perfect periodic table of microbial functions, improvements in automatic algorithms are still possible, as shown recently by the approaches of Tatusov *et al.*[10], Overbeek *et al.*[11] and some bioinformatics companies, such as GeneData (Basel, Switzerland). Time-efficient, integrated clustering algorithms that automatically generate hierarchical classification systems on the basis of pair-wise genome comparisons, including superfamilies and homology domain families as well as ortholog/paralog clusters, are worth being optimized in order to obtain a quick, comprehensive overview of the protein sets of a variety of microbes.

Functional assignments to proteins and target selection

Phylogenetic profiling

The initial question for target identification is: in which organisms do potential targets occur? This question is addressed by the phylogenetic profiling approach[12]. The basis of this approach are systems that classify proteins into families as already described. In a second step, for each protein family, the set of organisms is determined that encodes family members. Such a set of organisms is called the phylogenetic profile of a family. Intelligent classification systems can help to obtain comprehensive answers easily (Fig. 1a). For instance, the identification of orthologs of the essential protein MurC (the UDP-*N*-acetylmuramate:alanine ligase)[13] in the majority of pharmaceutically-relevant bacteria makes this protein an attractive broad-spectrum target. Some organisms such as *Escherichia coli*, *Haemophilus influenzae* and *Mycobacterium tuberculosis* possess paralogs called Mpl, which are functionally slightly different and do not complement the MurC orthologs[14]. Nevertheless, the phylogenetic distribution of the complete set of Mur-ligase superfamily members shows that eukaryotes also possess one member of this family, the tetrahydrofolylpolyglutamate synthase FolC (Ref. 15). This information could be helpful when assessing possible side-effects of potential future MurC-targeting antibiotics in the host. Although MurC and Mpl represent paralogs that do not complement each other, there are other cases where paralogs are indeed functionally equivalent. For instance, *Bacillus subtilis* contains two functional deformylases that are 32% identical to each other on the protein-sequence level[16]. These examples clearly show that sequence-based classification systems alone are not enough to predict the exact biological function of proteins.

Nevertheless, such classification systems in combination with concrete biological information help enormously in providing new insights into the physiological characteristics of organisms. For instance, the generally known inorganic pyrophosphatase Ppa is present in eukaryotes and Gram-negative bacteria, but cannot be found in many Gram-positive bacteria[17]. In these bacteria, other proteins non-orthologous to Ppa take over the same function. Such proteins are called analogs[18] (Fig. 2). Analogs represent target candidates with a defined spectrum and selectivity. By searching for protein families with a phylogenetic profile complementing the distribution of Ppa among bacteria, the YybQ family can be identified *in silico*. Indeed, YybQ has been proven experimentally to be the Gram-positive inorganic pyrophosphatase.

Metabolic pathway mapping

Genome comparison tools have to account for the complex relationships that exist between protein families, which, in turn, help to understand the metabolic network of microbes. The phylogenetic profiling approach enables the identification of groups of genes occurring in certain sets of bacteria that indicate that they are functionally coupled[19,20]. Such an approach, termed metabolic pathway mapping, helps to identify members of the recently detected deoxyxylulose 5-phosphate (DOXP) pathway for isopentenyldiphosphate biosynthesis[21]. This pathway produces the precursor of isoprenoids which are essential for cell wall biosynthesis and electron transport processes in

bacteria. While some bacteria possess the 'classical' mevalonate pathway that is also present in humans[22], many harbor the alternative DOXP pathway[23]. Some members of this pathway have been functionally characterized[24], while others remain to be identified. Phylogenetic profiling has helped to suggest candidate genes that could be experimentally shown to be involved in isoprenoid biosynthesis[25,26] (Fig. 3).

A prerequisite for mapping metabolic pathways on genomic data is the annotation of proteins with metabolic information. The commonly known Enzyme Commission (EC) numbers have been given to characterized proteins in order to classify the enzymatic chemical reactions of proteins. Unfortunately, the EC system only classifies enzymatic proteins and does not describe the cellular role of a gene product. Therefore, various genome comparison databases include additional classification systems that describe the biological process (e.g. transcription, translation), pathway (e.g. fatty acid biosynthesis) or cellular location in addition to the molecular function of the protein. Prominent metabolic role categorizations come, for instance, from the GenProtEC (Ref. 27) and EcoCyc (Ref. 28) databases, which focus on *E. coli* metabolism; the Gene Ontology Consortium, which aims to describe the roles of genes and gene products in any organism with a precisely defined vocabulary[29]; and the COG database[10].

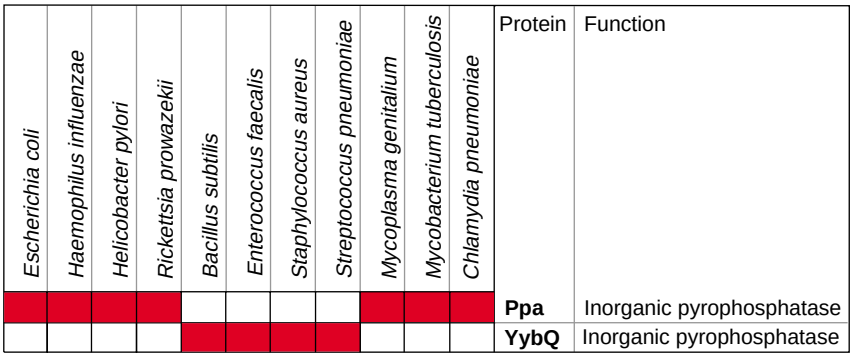


Figure 2. Phylogenetic profiling and identification of analogs. The phylogenetic pattern of inorganic pyrophosphatases among various bacterial species is represented. Red fields indicate the presence of orthologous proteins. Orthologs of the *Escherichia coli* pyrophosphatase Ppa could not be found in the Gram-positive model bacterium *Bacillus subtilis* and in major Gram-positive pathogens such as *Enterococcus faecalis*, *Staphylococcus aureus* and *Streptococcus pneumoniae*. The YybQ proteins represent one of the few ortholog families with a phylogenetic profile that is complementary to the Ppa family. The YybQ proteins have been shown experimentally to be inorganic pyrophosphatases, although they exhibit no sequence similarity to Ppa. YybQ and Ppa are called analogous proteins. The analysis was performed using GD Phylosopher™ (GeneData; Basel, Switzerland).

Functional coupling inferred by domain fusion and gene clusters

Identification of functional protein domains, which are fused to one single polypeptide chain in certain species but separated in others, indicates that the separated domains in other organisms are indeed physically interacting or at least functionally coupled[20,30,31]. One example is the *Chlamydia* gene product which carries out the MurC domain as already described. This polypeptide contains a second domain representing the D-alanine-D-alanine ligase (Ddl), although MurC and Ddl are separate proteins in many other bacteria (Fig. 1b). Both functions are indeed involved in the same pathway – the peptidoglycan biosynthesis.

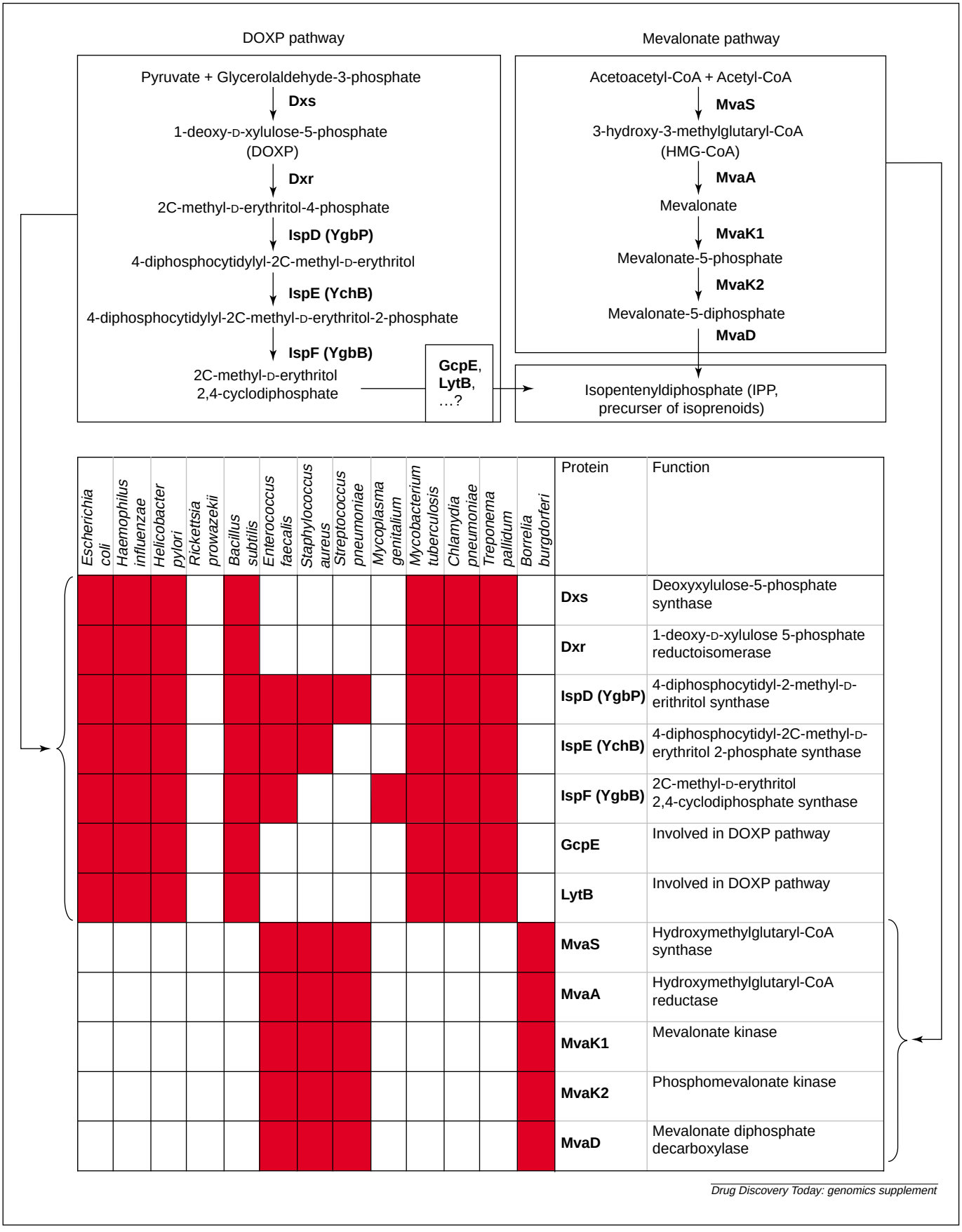
Analysis of the neighboring regions of genes in a large number of bacterial genome sequences enables the identification of conserved gene clusters in phylogenetically diverse organisms, which also pinpoint functional coupling of genes[20,31–33]. Again, the gene encoding MurC is an appropriate example as genes encoding enzymes required for peptidoglycan synthesis are spatially clustered in many bacterial genomes (Fig. 1b).

Strain profiling

Comparing genomes from related pathogenic and non-pathogenic species (differential genome analysis) can help in identifying pathogen-specific genes that could be tested for their pathogenicity[34]. Hybridization experiments using species-specific genome arrays that utilize genomic DNA probes derived from closely related strains of which the complete genomic sequences are unavailable, can enable the differentiation of the core set of genes defining a species from the strain-specific set of genes[35]. Strain-specific genes could be responsible for certain kinds of pathogenicity and resistance mechanisms. Such strain profiling requires computational tools that are appropriate for the analysis of array data. Such an analysis is similar to the analysis of expression profiling data, where genome arrays or chips are probed with RNA and relative amounts of mRNA have to be determined.

Expression profiling

The transcriptional activity of all genes encoded in a microbial genome can be measured using modern expression profiling techniques. Thus, groups of co-induced or co-repressed genes can be identified. The identification of co-regulated genes



Escherichia coli	*Haemophilus influenzae*	*Helicobacter pylori*	*Rickettsia prowazekii*	*Bacillus subtilis*	*Enterococcus faecalis*	*Staphylococcus aureus*	*Streptococcus pneumoniae*	*Mycoplasma genitalium*	*Mycobacterium tuberculosis*	*Chlamydia pneumoniae*	*Treponema pallidum*	*Borrelia burgdorferi*	Protein	Function
■	■	■		■					■	■	■		**Dxs**	Deoxyxylulose-5-phosphate synthase
■	■	■		■					■	■	■		**Dxr**	1-deoxy-D-xylulose 5-phosphate reductoisomerase
■	■	■		■	■	■	■		■	■	■		**IspD (YgbP)**	4-diphosphocytidyl-2-methyl-D-erithritol synthase
■	■	■		■	■	■			■	■	■		**IspE (YchB)**	4-diphosphocytidyl-2C-methyl-D-erythritol 2-phosphate synthase
■	■	■		■	■			■	■	■	■		**IspF (YgbB)**	2C-methyl-D-erythritol 2,4-cyclodiphosphate synthase
■	■	■		■					■	■	■		**GcpE**	Involved in DOXP pathway
■	■	■		■					■	■	■		**LytB**	Involved in DOXP pathway
					■	■	■					■	**MvaS**	Hydroxymethylglutaryl-CoA synthase
					■	■	■					■	**MvaA**	Hydroxymethylglutaryl-CoA reductase
					■	■	■					■	**MvaK1**	Mevalonate kinase
					■	■	■					■	**MvaK2**	Phosphomevalonate kinase
					■	■	■					■	**MvaD**	Mevalonate diphosphate decarboxylase

Drug Discovery Today: genomics supplement

◀ Figure 3. Phylogenetic profiling and pathway mapping. The two metabolic pathways leading to isopentenyldiphosphate (IPP), which are represented in the upper part of the figure, are the deoxyxylulose 5-phosphate (DOXP) pathway and the alternative mevalonate pathway. IPP is the precursor of isoprenoids essential for cell wall biosynthesis in bacteria. Red colored fields indicate presence of orthologs in the respective organisms. The ortholog clusters of the proteins involved in the respective pathways exhibit similar phylogenetic profiles, as shown in the lower part of the figure. Only the phylogenetic profiles of IspD, IspE and IspF do not correspond exactly to the phylogenetic profile of the respective pathway. The role of these proteins in *Mycoplasma genitalium*, *Enterococcus faecalis*, *Staphylococcus aureus* and *Streptococcus pneumoniae* is unclear. Such a phylogenetic profiling approach led to the identification of two novel target candidate proteins, GcpE and LytB, which are involved in the not-yet completely discovered branch of the DOXP pathway. The analysis was performed using the phylogenetic profiling module of GD Phylosopher™ (GeneData; Basel, Switzerland).

introduces new types of data to enable the prediction of functional coupling of genes[20]. The detection of co-regulated genes requires the application of adequate statistical tests, clustering algorithms and visualization tools[36]. In this field, the majority of analysis tools are still in early development with only a few at a more advanced level.

Recently, profiling experiments have been published based on arrays that contain, on average, one 25-mer oligonucleotide probe per 30 base pairs (bp) over the entire *E. coli* genome[37]. By using such high-resolution genome arrays, start and stop sites of transcripts can be mapped onto a genome more comprehensively and precisely than before. Such data will enable more precise deduction of promoter regions and the definition of common sequence-based features of co-regulated promoters.

Additionally, expression analyses provide the opportunity to study the global expression profile in response to antibiotic stress. Using this method, the mode of action of antibiotic compounds and some types of resistance mechanisms can be identified using appropriate marker gene identification algorithms[38].

Other genomics technologies

Proteome analyses (where the expression of proteins in 2D-gel electrophoresis is studied)[36,38] and metabolome analyses (where the amount of the cell's metabolites is measured)[39], together with other genomics-based technologies, represent challenges for scientists and algorithm developers to deduce valid information about the functional network in the cell. We are gradually beginning to be able to deduce the connections between gene products in order to build complex models that describe cellular physiology.

Integrated genome comparison and target prioritization

Integrated genome comparison systems, including protein-function classification systems and prediction tools as described here, are of high value for industrial research as they enable researchers to gain an efficient overview of the phylogenetic diversity and, consequently, the metabolic flexibility of microbes. A list of computational tools and services actually available in the field of microbial genome comparison is given in Table 1. We are now able to select antibiotic targets derived from completely deciphered genomes and investigate their phylogenetic distribution. The problem of searching for potential drug targets is now becoming increasingly shifted towards the question: which is the best target among the selected ones? Prioritizing targets will therefore be an essential process in industrial drug discovery. Information about knockout phenotypes, biochemical activities and metabolic roles, substrates, inhibitors and 3D structures will need to be integrated into the target prioritization tools. These integrated systems will enable estimation of the likely success of finding novel, promising chemical leads for drug discovery from target-based high-throughput screens.

References

1 Hand, W.L. (2000) Current challenges in antibiotic resistance. *Adolesc. Med.* 11, 427–438

2 Moir, D.T. *et al.* (1999) Genomics and antimicrobial drug discovery. *Antimicrob. Agents Chemother.* 43, 439–446

3 Venter, J.C. *et al.* (2001) The sequence of the human genome. *Science* 291, 1304–1351

4 International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. *Nature* 409, 860–921

5 Altschul, S.F. *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25, 3389–3402

6 Pearson, W.R. and Lipman, D.J. (1988) Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci. U. S. A.* 85, 2444–2448

7 Persson, B. (2000) Bioinformatics in protein analysis. In *Proteomics in Functional Genomics. Protein Structure Analysis* (Jolles, P. and Jornvall, H., eds), pp. 215–231, Birkhäuser

8 Huynen, M.A. and Bork, P. (1998) Measuring genome evolution. *Proc. Natl. Acad. Sci. U. S. A.* 95, 5849–5856

9 Tatusov, R.L. *et al.* (1997) A genomic perspective on protein families. *Science* 278, 631–637

10 Tatusov, R.L. *et al.* (2001) The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res.* 29, 22–28

11 Overbeek, R. *et al.* (2000) WIT: integrated system for high-throughput genome sequence analysis and metabolic reconstruction. *Nucleic Acids Res.* 28, 123–125

12 Marcotte, E.M. *et al.* (2000) Localizing proteins in the cell from their phylogenetic profiles. *Proc. Natl. Acad. Sci. U. S. A.* 97, 12115–12120

13 Lowe, A.M. and Deresiewicz, R.L. (1999) Cloning and sequencing of *Staphylococcus aureus murC*, a gene essential for cell wall biosynthesis. *DNA Seq.* 10, 19–23

14 Mengin-Lecreulx, D. *et al.* (1996) Identification of the *mpl* gene encoding UDP-*N*-acetylmuramate:L-alanyl-gamma-D-glutamyl-meso-diaminopimelate ligase in *Escherichia coli* and its role in recycling of cell wall peptidoglycan. *J. Bacteriol.* 178, 5347–5352

15 Bertrand, J.A. *et al.* (2000) 'Open' structures of MurD: domain movements and structural similarities with folylpolyglutamate synthetase. *J. Mol. Biol.* 301, 1257–1266

16 Haas, M. *et al.* (2001) YkrB is the main peptide deformylase in *Bacillus subtilis*, an eubacterium containing two functional peptide deformylases. *Microbiology* 147, 1783–1791

17 Sivula, T. *et al.* (1999) Evolutionary aspects of inorganic pyrophosphatase. *FEBS Lett.* 454, 75–80

18 Galperin, M.Y. *et al.* (1998) Analogous enzymes: independent inventions in enzyme evolution. *Genome Res.* 8, 779–790

19 Pellegrini, M. *et al.* (1999) Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc. Natl. Acad. Sci. U. S. A.* 96, 4285–4288

20 Marcotte, E.M. *et al.* (1999) A combined algorithm for genome-wide prediction of protein function. *Nature* 402, 83–86

21 Eisenreich, W. *et al.* (2001) Deoxyxylulose phosphate pathway to terpenoids. *Trends Plant Sci.* 6, 78–84

22 Wilding, E.I. *et al.* (2000) Identification, evolution, and essentiality of the mevalonate pathway for isopentenyl diphosphate biosynthesis in Gram-positive cocci. *J. Bacteriol.* 182, 4319–4327

23 Boucher, Y. and Doolittle, W.F. (2000) The role of lateral gene transfer in the evolution of isoprenoid biosynthesis pathways. *Mol. Microbiol.* 37, 703–716

24 Rohdich, F. *et al.* (2000) Biosynthesis of terpenoids: 4-diphosphocytidyl-2-C-methyl-D-erythritol kinase from tomato. *Proc. Natl. Acad. Sci. U. S. A.* 97, 8251–8256

25 Campos, N. *et al.* (2001) Identification of *gcpE* as a novel gene of the 2-C-methyl-D-erythritol 4-phosphate pathway for isoprenoid biosynthesis in *Escherichia coli*. *FEBS Lett.* 488, 170–173

26 Cunningham, F.X. *et al.* (2000) Evidence of a role for LytB in the nonmevalonate pathway of isoprenoid biosynthesis. *J. Bacteriol.* 182, 5841–5848

27 Riley, M. (1998) Genes and proteins of *Escherichia coli* K-12. *Nucleic Acids Res.* 26, 54

28 Karp, P.D. *et al.* (1999) EcoCyc: encyclopedia of *Escherichia coli* genes and metabolism. *Nucleic Acids Res.* 27, 55–58

29 Ashburner, M. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* 25, 25–29

30 Enright, A.J. *et al.* (1999) Protein interaction maps for complete genomes based on gene fusion events. *Nature* 402, 86–90

31 Huynen, M. *et al.* (2000) Predicting protein function by genomic context: quantitative evaluation and qualitative inferences. *Genome Res.* 10, 1204–1210

32 Overbeek, R. *et al.* (1999) The use of gene clusters to infer functional coupling. *Proc. Natl. Acad. Sci. U. S. A.* 96, 2896–2901

33 Fujibuchi, W. *et al.* (2000) Automatic detection of conserved gene clusters in multiple genomes by graph comparison and P-quasi grouping. *Nucleic Acids Res.* 28, 4029–4036

34 Huynen, M. *et al.* (1998) Differential genome analysis applied to the species-specific features of *Helicobacter pylori*. *FEBS Lett.* 426, 1–5

35 Salama, N. *et al.* (2000) A whole-genome microarray reveals genetic diversity among *Helicobacter pylori* strains. *Proc. Natl. Acad. Sci. U. S. A.* 97, 14668–14673

36 Celis, J.E. *et al.* (2000) Gene expression profiling: monitoring transcription and translation products using DNA microarrays and proteomics. *FEBS Lett.* 480, 2–16

37 Selinger, D.W. *et al.* (2000) RNA expression analysis using a 30 base pair resolution *Escherichia coli* genome array. *Nat. Biotechnol.* 18, 1262–1268

38 Gmuender, H. *et al.* (2001) Gene expression changes triggered by exposure of *Haemophilus influenzae* to novobiocin or ciprofloxacin: combined transcription and translation analysis. *Genome Res.* 11, 28–42

39 Raamsdonk, L.M. *et al.* (2001) A functional genomics strategy that uses metabolome data to reveal the phenotype of silent mutations. *Nat. Biotechnol.* 19, 45–50