Searching for drug targets in microbial genomes

Michael Y Galperin* and Eugene V Koonin†

Comparative analysis of the complete genome sequences of 10 bacterial pathogens available in the public databases offers the first insights into the drug discovery approaches of the near future. Genes that are conserved in different genomes often turn out to be essential, which makes them attractive targets for new broad-spectrum antibiotics. Subtractive genome analysis reveals the genes that are conserved in all or most of the pathogenic bacteria but not in eukaryotes; these are the most obvious candidates for drug targets. Species-specific genes, on the other hand, may offer the possibility to design drugs against a particular, narrow group of pathogens.

Addresses
National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA
*e-mail: galperin@ncbi.nlm.nih.gov
†e-mail: koonin@ncbi.nlm.nih.gov

Current Opinion in Biotechnology 1999, **10**:571–578

0958-1669/99/$ – see front matter. Published by Elsevier Science Ltd.

Introduction

Humans have been fighting bacterial pathogens for ages. The first hymns of the Sanskrit text *Atharva-veda Samhita*, dated ~2100 BC, already mention destroying 'invisible worms' in the body and suggest using a dark branched plant growing on the tree (most probably a fungus or a lichen) to cure infections of skin, bones, and whole body [1]. Since then, humans have been utilizing various chemical substances with antibacterial or bacteriostatic properties, which, in some cases, helped to tilt the balance in favor of the clearly disadvantaged side. In the past 50 years, borrowing anti-bacterials from other bacteria and fungi even produced an impression of success in this battle. This relatively peaceful situation could not have lasted long, however, as antibiotics have to be perfected in many painful stages of negative selection, whereas antibiotic resistance in bacteria is subject to positive selection. Indeed, a drug-resistant bacterium immediately acquires a tremendous evolutionary advantage over its competition, which allows drug resistance to quickly spread in the population. Facing more and more bacteria with multiple drug resistance mechanisms, humans have come to realize that they are still very far from understanding the ways in which bacterial cells function, divide, and behave in a community. In a major step towards achieving an ultimate understanding of bacterial cell physiology, genomes of a number of bacteria, including several pathogens, have been completely sequenced ([2•]; Table 1). Here we discuss the impact of microbial genomics on anti-bacterial drug discovery and review some new approaches to the identification of potential drug targets in complete microbial genomes.

Impact of complete genome sequences

In the past several years, complete genome sequences of several pathogenic bacteria have been determined and many more such projects are currently under way [2•,3–5,6••,7–9,10••,11,12••–14••]. Even though microbial genomics has had little direct impact on antibacterial drug discovery so far, the possibilities of using complete genome sequences for target identification are virtually unlimited [15]. Complete genomes allow us to, firstly, compile a list of all potential gene products produced by a particular organism, secondly, identify the functions (enzymes and pathways) that are missing in a particular organism, and finally, identify genes that are common to all (or most) microorganisms in a chosen group or, vice versa, unique to a particular pathogen. This redefines the problem of searching for potential drug targets into the problem of selecting best targets from the complete list of gene products.

An important advantage of this post-genomic analysis is the possibility of specifically looking for a target that is present in many, or only in several, bacterial genomes — that is, to design an antibiotic that should be active against a wide range of bacteria, or one that should function as a 'magic bullet' against a particular pathogen. In addition, comparing bacterial sequences to the growing database of human genes can eliminate potential drug targets that have close human homologs. This would help to avoid costly dead-ends when a lead target is identified and investigated in great detail only to find out at some later stage that all its inhibitors are invariably toxic for humans (see, however, [16]).

Virulence genes as drug targets

The most natural choice for a drug target would seem to be virulence-related genes, identified by *in vivo* expression technology [17,18], or by increasingly popular DNA microarrays [19]. Although these methods work for any bacteria, the availability of the complete genome sequence of non-pathogenic *Escherichia coli* K-12 strain MG1655 [3] provided a framework for analysis of genomes from other enterobacteria, such as enteropathogenic *E. coli* O157:H7 [20•], *Salmonella typhimurium* [21], *Salmonella typhi* [22], *Yersinia pestis*, and many others. The first comparisons of these genomes provided definite proof to the long-held notion that genes encoding bacterial virulence factors largely reside in well-defined pathogenicity islands that are scattered around the chromosome [23]. These islands often differ substantially from the rest of the genome in such parameters as GC content, codon usage, and gene density, suggesting that they are relatively recent acquisitions that conferred pathogenicity to a relatively benign symbiont [20•]. The proteins encoded by the pathogenicity islands remain attractive targets for drug intervention; however, apart from a few exceptions their functions and

Table 1

Genome sequencing of bacterial pathogens.

Species	Genome size (bp)	Proteins encoded*	Affected tissue	Disease caused	Reference
Proteobacteria					
Escherichia coli	4,639,221	4,289	Intestine and free-living	–	[3]
Haemophilus influenzae	1,830,138	1,709	Lungs	Pneumonia	[4]
Helicobacter pylori	1,667,867	1,566	Gastrum	Ulcer	[5]
Rickettsia prowazekii	1,111,529	834	Intestine	Typhus	[6••]
Gram-positive bacteria					
Bacillus subtilis	4,214,814	4,100	Free-living	–	[7]
Mycoplasma genitalium	580,073	467	Urogenital tract	Urethritis	[8]
Mycoplasma pneumoniae	816,394	677	Lungs	Pneumonia	[9]
Mycobacterium tuberculosis	4,411,529	3,918	Lungs	Tuberculosis	[10••]
Spirochaetae					
Borellia burgdorferi	910,725	850	Skin	Lyme disease	[11]
Treponema pallidum	1,138,011	1,031	Genitals	Syphilis	[12••]
Chlamydiae					
Chlamydia trachomatis	1,042,519	894	Urogenital tract	Trachoma, pelvic inflammation, epididymitis	[13••]
Chlamydia pneumoniae	1,230,230	1,052	Lungs, bronchs	Pneumonia, bronchitis	[14••]
Forthcoming genomes					
Campylobacter jejuni	1,641,480	1,731	Gastrum	Ulcer	(a)
Bordetella pertussis	~3,880,000	NA	Lungs	Whooping cough	(a)
Neisseria gonorrhoeae	~2,200,000	NA	Urogenital tract	Gonorrhea	(a)
Neisseria meningitidis	2,184,406	NA	Brain	Meningitis, septicaemia	(a)
Salmonella typhi	~4,500,000	NA	Intestine	Typhoid fever	(a)
Streptococcus pyogenes	1,877,697	1910	Pharynx, skin, soft tissues	Necrotizing fascitis, toxic shock syndrome	(a)
Vibrio cholerae	~2,500,000	NA	Intestine	Cholera	(a)
Ureaplasma urealyticum	751,719	605	Urogenital tract	Chorioamnionitis, intrauterine infection	(a)
Yersinia pestis	~4,380,000	NA	Intestine	Plague	(a)

*The number of annotated open reading frames in the latest update of GenBank™ Genomes division (ftp://ncbi.nlm.nih.gov/Entrez/genomes), see [2•] for more details. (a) The data on the *C. jejuni*, *B. pertussis*, *N. meningitidis*, *S. typhi*, and *Y. pestis* sequencing projects are from the Sanger Centre website (http://www.sanger.ac.uk/Projects/Microbes); on the *N. gonorrhoea* project from the University of Oklahoma website (http://www.genome@ou.edu); on the *S. pyogenes* project from the presentation by JJ Ferretti at the ASM General Meeting (Chicago, IL, 1999); on the *U. urealyticum* project from the University of Alabama website (http://genome.microbio.uab.edu/uu/uugen.htm); on the *V. cholerae* project from the NIAID website (http://www.niaid.nih.gov/dmid/genome.htm).

roles in pathogenesis remain unknown. It is unclear therefore, what, if any, would be the effect of drugs targeted against such proteins. An important case in point is the recent evidence that production of Shiga toxin 2 by enteropathogenic *E. coli* O157:H7 might be coupled to cell lysis by its lysogenic bacteriophage 933W [24•].

Given the diversity and genetic plasticity of virulence plasmids and pathogenicity islands, selecting a satisfactory drug target from the growing body of sequence information [24•,25–27] remains quite a challenge. To further complicate the case, in addition to the acquisition of pathogenicity islands and/or virulence plasmids, conversion of *E. coli* into a pathogen might involve a loss of certain genome fragments [28•]. To delineate the differences between related pathogens and non-pathogens, the recently developed PCR-based subtractive hybridization method [29•] can be used to specifically amplify DNA sequences that are present in one (e.g. virulent) but not the other (e.g. avirulent) strain, even when complete genomes are not yet available. The most important question in the process of drug target selection thus becomes the requirement that the target be not only expressed in the host organism, but be essential for virulence, or better yet, for the very survival of the microorganism.

Uncharacterized essential genes as drug targets

Identification of new genes as being essential can be accomplished by a number of ways. The experimental approaches usually rely on lethality of gene deletions and/or transposon insertions into the gene in question [30]. Recently, a combination of transposon mutagenesis with PCR-based screening has been used to efficiently identify essential genes in *Haemophilus influenzae* and *Streptococcus pneumoniae* [31••]. The availability of complete genomes, however, allows us to greatly simplify this task by applying computational methods for the initial identification of probable essential genes. The notion that the genes that are conserved across diverse phylogenetic lineages are likely to be essential was first used to delineate a possible 'minimal gene complement' [32]. Since then, well-conserved unidentified genes have indeed been demonstrated to be indispensable for bacterial cell growth in several cases [33••,34•]. Demonstration of essentiality of a particular gene is, of course, only the first step towards using it as a drug target. Significant follow-up

research is required to characterize the cellular functions of these gene products and validate them as targets. Such projects are of major fundamental value, as they help to assign functions to the remaining unidentified genes in microbial genomes, thus filling the 'blank spots' in bacterial proteomes [34•]. Lists of such conserved uncharacterized proteins are available in the Clusters of Orthologous Groups of proteins (COG) database [35,36•] as the S-COGs and in the PROSITE database [37] as Uncharacterized Protein Families (UPFs). Although characterizing a completely new protein family that can be used as a drug target is certainly a worthwhile goal, most research groups lack sufficient resources to systematically implement this approach. It makes sense, therefore, to consider looking for drug targets among previously characterized proteins that are specific and essential for a particular pathogen.

Species-specific genes as drug targets

An interesting approach to the prediction of potential drug targets, designated 'differential genome display', has been proposed by Peer Bork and co-workers [38,39••]. This approach relies on the fact that genomes of parasitic microorganisms are generally much smaller and code for fewer proteins than the genomes of free-living organisms. The genes that are present in the genome of a parasitic bacterium, but absent in a closely related genome of free-living bacterium, are therefore likely to be important for pathogenicity and can be considered candidate drug targets. Exhaustive comparison of *H. influenzae* and *E. coli* gene products identified 40 *H. influenzae* genes that have been exclusively found in pathogens and thus constitute potential drug targets [38]. Recently, this approach was expanded by considering the genome of *H. pylori*, another proteobacterial pathogen [5,40]. By comparing the *H. pylori* genome with those of *E. coli* and *H. influenzae*, 594 *H. pylori*-specific genes were identified, of which 398 had unknown functions, 123 represented known host-interaction factors, and the remaining 73 genes coded for species-specific proteins [39••]. This latter group included metabolic enzymes, restriction enzymes, transposases, and predicted conjugation enzymes. Some of the enzymes identified in this search, such as, for example, pyruvate : ferredoxin oxidoreductase, had relatively narrow phylogenetic distribution and could be used as potential drug targets.

A similar approach, referred to as 'concordance analysis', has been described [41•] that allows one to search for sequences that are present in some genomes (e.g. *E. coli*, *B. subtilis*, *H. influenzae*, *H. pylori*, and *M. tuberculosis*) but not in others (e.g. yeast). This algorithm would search bacterial genomes for conserved bacteria-specific proteins that are absent in eukaryotes. Although the genes identified in such a search are not necessarily essential for bacteria, their conservation across the bacterial domain is indicative of their importance. A significant drawback of this approach, however, is the necessity to rely on certain arbitrary cut-off scores to deduce similarity. This may result in costly mistakes, such as, for example, not recognizing yeast homologs of the *E. coli* genes *fabD*, *galU*, and *glnA* [41•], all of which are not only present in yeast, but have even been experimentally characterized.

Figure 1

'Genome subtraction' using the COG database. The phylogenetic pattern '-hu*****' denotes conserved proteins that are present both in *H. influenzae* and *H. pylori* genomes, but not in the related but much larger genome of *E. coli*. Such unusual phylogenetic patterns often indicate species-specific proteins that might be used as potential drug targets.

Another approach to 'genome subtraction' takes advantage of the COG database, which includes conserved protein families represented in at least three phylogenetically distant organisms with completely sequenced genomes [35,36•]. Clicking on a particular phylogenetic pattern retrieves a list of all COGs with the same pattern (Figure 1); this allows a direct search for rare and unusual phylogenetic patterns. For example, in the pattern '-hu*****', the first slot is reserved for *E. coli*; the dash indicates that the respective protein (COG) is not found in the *E. coli* genome. The letter h stands for *H. influenzae* and the letter u stands for *H. pylori* (ulcer): asterisks in the remaining slots mean that respective proteins (COGs) can be either present or absent in genomes of the other organisms. Entering such a pattern initiates a search for all conserved protein families (COGs) that are present in both *H. influenzae* and *H. pylori*, but absent in *E. coli* genome (Figure 1). In the current release of the COG database (21 genomes), this pattern is found in only 17

Table 2

Examples of species-specific drug targets in bacterial pathogens.

Microorganism	Suggested drug targets*	References
Haemophilus influenzae	SNZ1/pyroA	[42,44]
Helicobacter pylori	Phosphoglycerate mutase, phosphoenolpyruvate synthase	[36•]
Rickettsia prowazekii	Lysyl-tRNA synthetase, ATP/ADP translocase	[55,60]
Mycoplasma genitalium	Phosphoglycerate mutase	[61]
Mycoplasma pneumoniae	Phosphoglycerate mutase	[61]
Mycobacterium tuberculosis	OMP decarboxylase	[51]
Borellia burgdorferi	Lysyl-tRNA synthetase	[60]
Treponema pallidum	Lysyl-tRNA synthetase	[60]
Chlamydia trachomatis	DhnA-type aldolase, ATP/ADP translocase	[13••,55]
Chlamydia pneumoniae	DhnA-type aldolase, ATP/ADP translocase	[13••,55]

*These enzymes are predicted to be indispensable for the indicated microorganisms on the basis of the functional predictions and pathway assignments, described in [35,36•,55].

protein families, five of which include urease and urease-accessory proteins, which are essential for acid tolerance in *H. pylori* and are known virulence factors [36•,40]. In an agreement with the 'differential display' approach [39••], this search also retrieves type II 3-dehydroquinase, the predicted cytochrome c biogenesis protein CcdA, and other proteins (Figure 1). Search for *H. influenzae* genes that are missing both in *E. coli* and *H. pylori* brings up only seven protein families, two of which are remarkably well conserved in very diverse organisms [42,43] and have been recently implicated in pyridoxine biosynthesis [44]. These examples show that 'genome subtraction' approaches are quite useful, particularly when they retrieve species-specific target proteins with well-established activities, or better yet, with known substrate specificities.

Unique enzymes as drug targets

As most currently known antibacterials are essentially inhibitors of certain bacterial enzymes, all bacteria-specific enzymes can be considered potential drug targets. Such enzymes can be identified by the 'genome subtraction' methods described above, followed by a detailed analysis of each of the resulting protein families. Until the complete human genome becomes available, however, this analysis will remain fairly cumbersome, as 'genome subtraction' will have to be done with all available eukaryotic genomes, such as yeast and worm ones, with the human expressed sequence tag (EST) database, and so on. Fortunately, this process can be partially automated by considering only those enzymes that have been already characterized in bacteria but are known to be missing in humans. An attractive variant of this kind is a substitution of 'genome subtraction' by 'pathway subtraction' [45•,46•]. This approach, referred to as PathoLogic [46•], quickly identifies enzyme pathways that are specific for bacteria (e.g. peptodoglycan biosynthesis) and, accordingly, represent convenient drug targets.

A different variant of a similar approach relies on the enzymes that are subject to non-orthologous gene displacement — that is, can be found in nature in two or more different forms [47,48]. In many such cases, one form of the enzyme is found in bacteria, while a different form (distantly related or unrelated) is found in eukaryotes [49••]. Such analogous (as opposed to homologous) enzymes can be identified based on the simple criterion that they have identical Enzyme Commission (EC) numbers, but very low sequence similarity scores (Figure 2). In several cases, detailed analysis of such proteins has led to suggestions that they could be used as drug targets ([49••]; see Table 2).

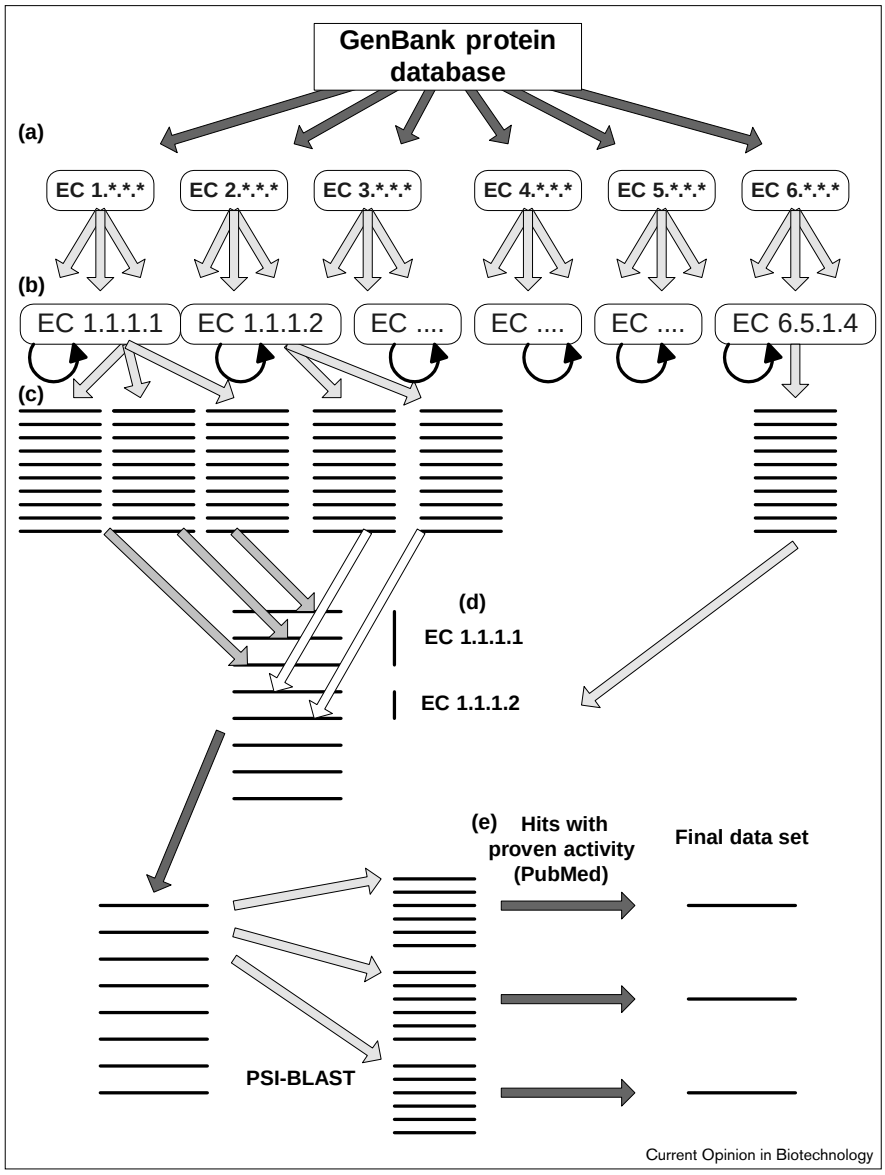
In many cases, metabolic reconstruction fails to identify candidate enzymes for certain critical steps in bacterial metabolism. Non-orthologous gene displacement can be expected for these steps, and such 'unusual' enzymes make very attractive drug targets. Examples of such enzymes include dehydrine-type fructose-1,6-bisphosphate aldolase, originally described in *E. coli* [50] and represents the only aldolase variant in *Chlamydia* [13••,14••], and the unique orotidine-5′-phosphate decarboxylase of *M. tuberculosis* [51], found so far only in other mycobacteria and in *Myxococcus xanthus*, *Thermus thermophilus*, and *Trypanosoma cruzi* (Table 2).

Membrane transporters as drug targets

Comparative analysis of complete genomes revealed that most of the pathogens have drastically diminished biosynthetic capabilities as compared to their free-living relatives [8,56]. Instead, these organisms depend on their hosts to provide essential nutrients such as amino acids, nucleobases, and vitamins. Transport systems for these nutrients are generally well conserved and easily identifiable [52]. Analysis of metabolic pathways allows one to predict which substrates cannot be produced inside the cells and so need to be transported [53]. All this makes bacterial transport proteins attractive drug targets. A number of non-metabolizable analogs of amino acids and nucleobases are already available and can be easily checked for their potential antibacterial activity. The nearly universal distribution of most major classes of transport proteins [54], however, makes it probable that effective inhibitors of bacterial transport turn out to be also toxic for humans. A notable exception is the ATP/ADP translocase of *Chlamydia* and *Rickettsia*, which is related to the plant chloroplast translocase rather than to the mitochondrial one [55]. This translocase, which is clearly indispensable for these pathogens, appears to be a promising drug target (Table 2).

Finally, an interesting approach that has emerged only recently includes improving activity of the existing antibiotics by inhibiting bacterial multidrug transporters [56]. Genomes of many pathogenic bacteria appear to contain homologs of multidrug pumps, which protect

Figure 2



Search for potential drug targets among analogous enzymes. Analogous enzymes are identified by: **(a)** extracting from the GenBank™ all the sequences with assigned Enzyme Commission (EC) numbers; **(b)** comparing all the sequences with the same EC number using the BLAST program; **(c)** grouping together the sequences that show significant similarity; **(d)** identifying the EC numbers that contain two or more such groups; **(e)** selecting for each such group representative sequences that have links to at least one PubMed citation. It ensures that the EC assignment is correct and the sequences obtained indeed represent analogous enzymes. Identification of potential drug targets is accomplished by, firstly, comparing the list of analogous enzymes against the complete protein sets of the pathogenic bacteria; and secondly, identifying all the cases when a parasitic genome contains enzyme forms that are not found in any eukaryote (or in any metazoa) and thus can potentially be inhibited without major side effects.

bacterial cells by exporting antibiotic molecules. Inhibiting such pumps not only creates convenient model organisms for studying drug effectiveness [57•], but also allows resistance to classical drugs such as tetracycline to be overcome [58•].

Conclusions

The recent efforts in microbial genome sequencing have been largely driven by the necessity to find effective ways to cure and prevent diseases caused by these microorganisms. The profound impact that these new sequence data are going to have on all aspects of life science is only beginning to be felt. In terms of antimicrobial drug discovery, researchers now find themselves in a situation that might remind one of the 'Star Wars' rebels receiving the blueprint of the Imperial battle station: there are numerous possibilities for mounting an attack, the success of which, however, is far from predetermined. The list of potential drug targets encoded in microbial genomes includes outer-membrane proteins, host-interaction factors, permeases, enzymes of intermediary metabolism, systems for DNA replication, transcription, and repair, translation apparatus, and many more. All these possibilities can and should be explored to bring about comprehensive multi-pronged combinations of antibiotics that would effectively fight bacterial pathogens. If recent indications of bacterial involvement in heart disease, atherosclerosis, and stomach cancer [14••] are true, antibacterial drugs may help to cure many of humanity's worst diseases. In this respect, at least, the next century should be very exciting.

Acknowledgements

We appreciate helpful suggestions by L Aravind and K Lewis.

References and recommended reading

Papers of particular interest, published within the annual period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. **Hymns of the Atharva-veda.** Translated by Maurice Bloomfield. *In The Sacred Books of the East*, vol 42. Edited by Muller FM. Oxford: Clarendon Press; 1897: Hymns 1.23, 1.24, 11.31.

2. • Genomes in Progress List on World Wide Web URL: http://www.ncbi.nlm.nih.gov/PMGifs/Genomes/bact.html

This list, which is a part of the GenBank™ Entrez Genome division, contains a constantly updated list of ongoing, non-commercial microbial genome sequencing projects. It has links to similar lists maintained at TIGR, University of Illinois, Argonne National Lab, and INFOBIOGEN, to the list of bacterial genomes being sequenced at the Sanger Centre, and to the two major US funding agencies, The Department of Energy (DOE) and National Institute of Allergy and Infectious Diseases (NIAID). Exploring the latter two sites provides useful information on the status of major projects and the motivation behind them.

3. Blattner FR, Plunkett G III, Bloch CA, Perna NT, Burland V, Riley M, Collado-Vides J, Glasner JD, Rode CK *et al.*: **The complete genome sequence of *Escherichia coli* K-12.** *Science* 1997, **277**:1453-1474.

4. Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, Bult CJ, Tomb J-F, Dougherty BA, Merrick JM *et al.*: **Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd.** *Science* 1995, **269**:496-512.

5. Tomb J-F, White O, Kerlavage AR, Clayton RA, Sutton GG, Fleishmann RF, Ketchum KA, Klenk HP, Gill S, Dougherty BA *et al.*: **The complete genome sequence of the gastric pathogen *Helicobacter pylori.*** *Nature* 1997, **388**:539-547.

6. •• Andersson SG, Zomorodipour A, Andersson JO, Sicheritz-Ponten T, Alsmark UC, Podowski RM, Naslund AK, Eriksson AS, Winkler HH, Kurland CG: **The genome sequence of *Rickettsia prowazekii* and the origin of mitochondria.** *Nature* 1998, **396**:133-140.

Rickettsia prowazekii, besides being an obligate intracellular parasite that causes typhus, is also a surprizingly close relative of the mitochondria. In contrast to the previously sequenced bacterial genomes, *R. prowazekii* lacks genes for glycolytic enzymes but has genes for the tricarboxylic acid cycle enzymes. Its respiratory chain and F_0F_1-type H^+-ATPase are similar to mitochondrial ones, but ATP/ADP translocase is related to that of chlamydia. The close relation of *R. prowazekii* to mitochondria will pose a challenge to devising effective anti-rickettsial drugs.

7. Kunst F, Ogasawara N, Moszer I, Albertini AM, Alloni G, Azevedo V, Bertero MG, Bessieres P, Bolotin A, Borchert S *et al.*: **The complete genome sequence of the gram-positive bacterium *Bacillus subtilis.*** *Nature* 1997, **390**:249-256.

8. Fraser CM, Gocayne JD, White O, Adams MD, Clayton RA, Fleischmann RD, Bult CJ, Kerlavage AR, Sutton G, Kelley JM *et al.*: **The minimal gene complement of *Mycoplasma genitalium.*** *Science* 1995, **270**:397-403.

9. Himmelreich R, Hilbert H, Plagens H, Pirkl E, Li BC, Herrmann R: **Complete sequence analysis of the genome of the bacterium *Mycoplasma pneumoniae.*** *Nucleic Acids Res* 1996, **24**:4420-4449.

10. •• Cole ST, Brosch R, Parkhill J, Garnier T, Churcher C, Harris D, Gordon SV, Eiglmeier K, Gas S, Barry CE III *et al.*: **Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence.** *Nature* 1998, **393**:537-544.

Genome sequencing of one of the deadliest human pathogens revealed that it is not a typical parasite, after all [61]. The *M. tuberculosis* genome is comparable in size to those of free-living *B. subtilis* and *Synechocystis sp*. It contains genes for most of the biosynthetic enzymes, which are lacking in most other pathogens, and *M. tuberculosis* does not seem to depend upon the host as much as other pathogens do. The reason for the extremely slow growth of this organism has not become immediately clear. Significant follow-up research will be needed to come up with effective anti-tuberculine drugs.

11. Fraser CM, Casjens S, Huang WM, Sutton GG, Clayton R, Lathigra R, White O, Ketchum KA, Dodson R, Hickey EK *et al.*: **Genomic sequence of a Lyme disease spirochaete, *Borrelia burgdorferi.*** *Nature* 1997, **390**:580-586.

12. •• Fraser CM, Norris SJ, Weinstock GM, White O, Sutton GG, Dodson R, Gwinn M, Hickey EK, Clayton R, Ketchum KA *et al.*: **Complete genome sequence of *Treponema pallidum*, the syphilis spirochete.** *Science* 1998, **281**:375-388.

With syphilis remaining one of the most common human diseases, the complete genome of *T. pallidum* provides much needed information about the metabolic capacities and other cellulor functions of this pathogen. Its similarity to previously sequenced *B. burgdorferi* genome offers interesting possibilities for comparative genomics of these two spirochetes. Remarkably, the *T. pallidus* chromosome is circular, leaving *B. burgdorferi* as the only known bacterium with a linear one.

13. •• Stephens RS, Kalman S, Lammel C, Fan J, Marathe R, Aravind L, Mitchell W, Olinger L, Tatusov RL, Zhao Q *et al.*: **Genome sequence of an obligate intracellular pathogen of humans: *Chlamydia trachomatis.*** *Science* 1998, **282**:754-759.

A comprehensive analysis of the chlamydial genome in an attempt to characterize this organism in as much detail as possible. While some of the motif-based annotations have already been confirmed by direct experimentation (e.g. assignment of fructose-1,6-bisphosphatase activity to CT215 supported by [50]), many of them remain to be verified. The most unexpected result of this genome analysis is probably the observation that certain chlamydial genes may have been recruited from the host eukaryotic cell.

14. •• Kalman S, Mitchell W, Marathe R, Lammel C, Fan J, Hyman RW, Olinger L, Grimwood J, Davis RW, Stephens RS: **Comparative genomes of *Chlamydia pneumoniae* and *C. trachomatis.*** *Nat Genet* 1999, **21**:385-389.

Comparative analysis of the two chlamydial genomes found 214 additional genes in *C. pneumoniae* compared to *C. trachomatis*. The most conspicuous differences are found in outer membrane proteins, the secretion system, and biosynthetic capabilities. *C. pneumoniae* is a suspect accomplice in a variety of supposedly non-infectious diseases, such as atherosclerosis, heart disease, adult-onset asthma and even lung cancer.

15. Saunders NJ, Moxon ER: **Implications of sequencing bacterial genomes for pathogenesis and vaccine development.** *Curr Opin Biotechnol* 1998, **9**:618-623.

16. Moir DT, Shaw KJ, Hare RS, Vovis GF: **Genomics and antimicrobial drug discovery.** *Antimicrob Agents Chemother* 1999, **43**:439-446.

17. Slauch JM, Mahan MJ, Mekalanos JJ: ***In vivo* expression technology for selection of bacterial genes specifically induced in host tissues.** *Methods Enzymol* 1994, **235**:481-492.

18. Mahan MJ, Tobias JW, Slauch JM, Hanna PC, Collier RJ, Mekalanos JJ: **Antibiotic-based selection for bacterial genes that are specifically induced during infection of a host.** *Proc Natl Acad Sci USA* 1995, **92**:669-673.

19. Debouck C, Goodfellow PN: **DNA microarrays in drug discovery and development.** *Nat Genet* 1999, **21**:48-50.

20. • Perna NT, Mayhew GF, Posfai G, Elliott S, Donnenberg MS, Kaper JB, Blattner FR: **Molecular evolution of a pathogenicity island from enterohemorrhagic *Escherichia coli* O157:H7.** *Infect Immun* 1998, **66**:3810-3817.

Comparison of pathogenicity islands from distantly related enterohemorrhagic and enteropathogenic strains of *E. coli* showed that although shared genes of the secretion apparatus were more than 98% identical, the genes coding for host interaction factors were much more divergent. Variability of genes responsible for the colonization of the host appears to be an important mechanism of pathogen evolution.

21. Wong RM, Wong KK, Benson NR, McClelland M: **Sample sequencing of a *Salmonella typhimurium* LT2 lambda library: comparison to the *Escherichia coli* K12 genome.** *FEMS Microbiol Lett* 1999, **173**:411-423.

22. McClelland M, Wilson RK: **Comparison of sample sequences of the *Salmonella typhi* genome to the sequence of the complete *Escherichia coli* K-12 genome.** *Infect Immun* 1998, **66**:4305-4312.

23. Hacker J, Blum-Oehler G, Muhldorfer I, Tschape H: **Pathogenicity islands of virulent bacteria: structure, function and impact on microbial evolution.** *Mol Microbiol* 1997, **23**:1089-1097.

24. • Plunkett G III, Rose DJ, Durfee TJ, Blattner FR: **Sequence of Shiga toxin 2 phage 933W from *Escherichia coli* O157:H7: Shiga toxin as a phage late-gene product.** *J Bacteriol* 1999, **181**:1767-1778.

The evidence suggesting that the Shiga toxin production might be coupled with lytic growth of the lysogenic phage 933W brings up a heretical question: maybe we should target the phage itself and not its unfortunate host?

25. Lindler LE, Plano GV, Burland V, Mayhew GF, Blattner FR: **Complete DNA sequence and detailed analysis of the *Yersinia pestis* KIM5 plasmid encoding murine toxin and capsular antigen.** *Infect Immun* 1998, **66**:5731-5742.

26. Perry RD, Straley SC, Fetherston JD, Rose DJ, Gregor J, Blattner FR: **DNA sequencing and analysis of the low-Ca^{2+}-response plasmid pCD1 of *Yersinia pestis* KIM5.** *Infect Immun* 1998, **66**:4611-4623.

27. Burland V, Shao Y, Perna NT, Plunkett G, Sofia HJ, Blattner FR: **The complete DNA sequence and analysis of the large virulence plasmid of *Escherichia coli* O157:H7.** *Nucleic Acids Res* 1998, **26**:4196-4204.

28. • Maurelli AT, Fernández RE, Bloch CA, Rode CK, Fasano A: **'Black holes' and bacterial pathogenicity: a large genomic deletion that enhances the virulence of *Shigella spp.* and enteroinvasive *Escherichia coli*.** *Proc Natl Acad Sci USA* 1998, **95**:3943-3948.
A conclusive demonstration that a pathogen is not just a capricious bacterium with a pathogenicity island. Evolution to pathogenicity appears to involve gene loss as well as gene acquisition.

29. • Akopyants NS, Fradkov A, Diatchenko L, Hill JE, Siebert PD, Lukyanov SA, Sverdlov ED, Berg DE: **PCR-based subtractive hybridization and differences in gene content among strains of Helicobacter pylori.** *Proc Natl Acad Sci USA* 1998, **95**:13108-13113.
A long-needed experimental method to identify and amplify DNA fragments that differ in two otherwise closely related organisms. This should be extremely useful in comparing various new isolates against the reference genome sequence.

30. Link AJ, Phillips D, Church GM: **Methods for generating precise deletions and insertions in the genome of wild-type *Escherichia coli*: application to open reading frame characterization.** *J Bacteriol* 1997, **179**:6228-6237.

31. •• Akerley BJ, Rubin EJ, Camilli A, Lampe DJ, Robertson HM, Mekalanos JJ: **Systematic identification of essential genes by *in vitro* mariner mutagenesis.** *Proc Natl Acad Sci USA* 1998, **95**:8927-8932.
A direct method for detecting essential genes in pre-amplified genomic fragments of up to 10 kb in size. As the method identifies transposon insertions that are lethal under given growth conditions, it should be useful for sorting out which genes are essential under each particular condition.

32. Mushegian AR, Koonin EV: **A minimal gene set for cellular life derived by comparison of complete bacterial genomes.** *Proc Natl Acad Sci USA* 1996, **93**:10268-10273.

33. •• Arigoni F, Talabot F, Peitsch M, Edgerton MD, Meldrum E, Allet E, Fish R, Jamotte T, Curchod ML, Loferer H: **A genome-based approach for the identification of essential bacterial genes.** *Nat Biotechnol* 1998, **16**:851-856.
This paper from the now defunct Geneva Biomedical Research Institute contains the only comprehensive investigation of the essentiality of conserved genes to date. Of 26 genes that are conserved in *E. coli* and *M. genitalium*, but have not been experimentally characterized, five were found to be essential for growth of both *E. coli* and *B. subtilis*; one more gene turned out to be essential for *E. coli*, but not *B. subtilis*. As genomes of both *E. coli* and *B. subtilis*, in contrast to much smaller genomes of pathogens, are highly redundant, these results look extremely encouraging.

34. • Apfel CM, Takacs B, Fountoulakis M, Stieger M, Keck W: **Use of genomics to identify bacterial undecaprenyl pyrophosphate synthetase: cloning, expression, and characterization of the essential uppS gene.** *J Bacteriol* 1999, **181**:483-492.
This paper from a Hoffmann-La Roche group is another example of corporate giants going after uncharacterized conserved genes and getting valuable information of general importance. Even if the uppS gene, characterized in this paper, turned out to be a poor drug target, an important missing link in our understanding of lipid biosynthesis has been closed.

35. Tatusov RL, Koonin EV, Lipman DJ: **A genomic perspective on protein families.** *Science* 1997, **278**:631-637.

36. • Koonin EV, Tatusov RL, Galperin MY: **Beyond the complete genomes: from sequences to structure and function.** *Curr Opin Struct Biol* 1998, **8**:355-363.
This paper and [35] give a general description of the Clusters of Orthologous Groups (COG) database (http://www.ncbi.nlm.nih.gov/COG/) which consists of families of conserved proteins from completely sequenced genomes, identified in all-against-all similarity searches. Because orthologous proteins in different genomes tend to carry out the same or very closely related functions, assignment of proteins to the same COG allows one to deduce its function(s), based on the function(s) of the better studied members of the COG.

37. Hofmann K, Bucher P, Falquet L, Bairoch A: **The PROSITE database, its status in 1999.** *Nucleic Acids Res* 1999, **27**:215-219.

38. Huynen MA, Diaz-Lazcoz Y, Bork P: **Differential display of genomes.** *Trends Genet* 1997, **13**:389-390.

39. •• Huynen M, Dandekar T, Bork P: **Differential genome analysis applied to the species-specific features of Helicobacter pylori.** *FEBS Lett* 1998, **426**:1-5.
This reference and the previous one describe an interesting approach to the identification of species-specific genes, which are likely to comprise potential drug targets.

40. Montecucco C, Papini E, de Bernard M, Zoratti M: **Molecular and cellular activities of *Helicobacter pylori* pathogenic factors.** *FEBS Lett* 1999, **452**:16-21.

41. • Bruccoleri RE, Dougherty TJ, Davison DB: **Concordance analysis of microbial genomes.** *Nucleic Acids Res* 1998, **26**:4482-4486.
Another description of a 'genome subtraction' technique that can potentially be used for comparison of any number of genomes, including incomplete and proprietory ones.

42. Galperin MY, Koonin EV: **Sequence analysis of an exceptionally conserved operon suggests enzymes for a new link between histidine and purine biosynthesis.** *Mol Microbiol* 1997, **24**:443-445.

43. Padilla PA, Fuge EK, Crawford ME, Errett A, Werner-Washburne M: **The highly conserved, coregulated SNO and SNZ gene families in *Saccharomyces cerevisiae* respond to nutrient limitation.** *J Bacteriol* 1998, **180**:5718-5726.

44. Osmani AH, May GS, Osmani SA: **The extremely conserved pyroA gene of *Aspergillus nidulans* is required for pyridoxine synthesis and indirectly for resistance to photosensitizers.** *J Biol Chem* 1999, **274**:23565-23569.

45. • Karp PD, Riley M, Paley SM, Pellegrini-Toole A, Krummenacker M: **EcoCyc: encyclopedia of *Escherichia coli* genes and metabolism.** *Nucleic Acids Res* 1999, **27**:55-58.
See annotation to [46•].

46. • Karp PD, Krummenacker M, Paley S, Wagg J: **Integrated pathway genome databases and their role in drug discovery.** *Trends Biotechnol* 1999, **17**:275-281.
This paper and [45•] describe the current state of the EcoCyc database (http://ecocyc.pangeasystems.com) in comparison to the two other major pathway databases, KEGG (http://www.genome.ad.jp/kegg) and WIT (http://wit.mcs.anl.gov/WIT2). The pathway comparison approach to identification of potential drug targets is discussed in detail.

47. Koonin EV, Mushegian AR, Bork P: **Non-orthologous gene displacement.** *Trends Genet* 1996, **12**:334-336.

48. Koonin EV, Galperin MY: **Prokaryotic genomes: the emerging paradigm of genome-based microbiology.** *Curr Opin Genet Dev* 1997, **7**:757-763.

49. •• Galperin MY, Walker DR, Koonin EV: **Analogous enzymes: independent inventions in enzyme evolution.** *Genome Res* 1998, **8**:779-790.
A study of the extreme cases of non-orthologous gene displacement, whereby the enzymes catalyzing the same biochemical reaction have no recognizable sequence similarity. In at least 34 cases out of the total 105 cases identified in this work, the two isoenzymes have or are predicted to have different three-dimensional structures, making them truly analogous (i.e. non-homologous) enzymes. Bacterial enzymes that have no significant sequence similarity to their human (mammalian) counterparts can be considered attractive drug targets.

50. Thomson GJ, Howlett GJ, Ashcroft AE, Berry A: **The *dhnA* gene of *Escherichia coli* encodes a class I fructose bisphosphate aldolase.** *Biochem J* 1998, **331**:437-445.

51. Aldovini A, Husson RN, Young RA: **The *uraA* locus and homologous recombination in *Mycobacterium bovis* BCG.** *J Bacteriol* 1993, **175**:7282-7289.

52. Clayton RA, White O, Ketchum KA, Venter JC: **The first genome from the third domain of life.** *Nature* 1997, **387**:459-462.

53. Galperin MY, Tatusov RL, Koonin EV: **Comparing microbial genomes: how the gene set determines the lifestyle**. In *Organization of the Prokaryotic Genome*. Edited by Charlebois RL. Washington: ASM Press; 1999:91-108.

54. Saier MH Jr: **Eukaryotic transmembrane solute transport systems.** *Int Rev Cytol* 1999, **190**:61-136.

55. Wolf YI, Aravind L, Koonin EV: ***Rickettsiae* and *Chlamydiae*: evidence of horizontal gene transfer and gene exchange.** *Trends Genet* 1999, **15**:173-175.

56. Lewis K: **Multidrug resistance: versatile drug sensors of bacterial cells.** *Curr Biol* 1999, **9**:R403-R407.

57. Hsieh PC, Siegel SA, Rogers B, Davis D, Lewis K: **Bacteria lacking a**
• **multidrug pump: a sensitive tool for drug discovery.** *Proc Natl Acad Sci USA* 1998, **95**:6602-6606.

See annotation to [58•].

58. Nelson ML, Levy SB: **Reversal of tetracycline resistance mediated**
• **by different bacterial tetracycline resistance determinants by an inhibitor of the Tet(B) antiport protein.** *Antimicrob Agents Chemother* 1999, **43**:1719-1724.

These two papers [57•,58•] show that mutidrug pumps play significant role in determining the level of bacterial sensitivity to commonly used antibiotics. Inhibiting these pumps or disrupting the respective genes may restore antibacterial activity of drugs that otherwise would be exported out the cell before they can induce any damage. This makes multidrug pumps natural targets for future antibacterial drugs.

59. Cole ST: **Learning from the genome sequence of *Mycobacterium tuberculosis* H37Rv.** *FEBS Lett* 1999, **452**:7-10.

60. Ibba M, Bono JL, Rosa PA, Soll D: **Archaeal-type lysyl-tRNA synthetase in the Lyme disease spirochete *Borrelia burgdorferi.*** *Proc Natl Acad Sci USA* 1997, **94**:14383-14388.

61. Galperin MY, Bairoch A, Koonin EV: **A superfamily of metalloenzymes unifies phosphopentomutase and cofactor-independent phosphoglycerate mutase with alkaline phosphatases and sulfatases.** *Protein Sci* 1998, **7**:1829-1835.