Finding drug targets in microbial genomes

Timothy D. Read, Steven R. Gill, Hervé Tettelin and Brian A. Dougherty

In this era of genomic science, knowledge about biological function is integrated increasingly with DNA sequence data. One area that has been significantly impacted by this accumulation of information is the discovery of drugs to treat microbial infections. Genome sequencing and bioinformatics is driving the discovery and development of novel classes of broad-spectrum antimicrobial compounds, and could enable medical science to keep pace with the increasing resistance of bacteria, fungi and parasites to current antimicrobials. This review discusses the use of genomic information in the rapid identification of target genes for antimicrobial drug discovery.

***Timothy D. Read**
Steven R. Gill
Hervé Tettelin
Microbial Genomics Group
The Institute for Genomic Research
9712 Medical Center Drive
Rockville, MD 20851, USA
*tel: +1 301 838 3554
fax: +1 301 838 0208
e-mail: tread@tigr.org
Brian A. Dougherty
The Bristol-Myers Squibb Pharmaceutical Research Institute
Wallingford, CT, USA

▼ In the new pathway for the discovery of drugs to combat the increasing menace of drug-resistant microorganisms[1], the crucial first step is to generate a set of target genes, from whole-genome data, that offer the potential for effective therapeutic intervention[2–5] (Fig. 1). The raw material for scientists interested in *in silico* drug-target discovery is sequence data, and it is important to consider the source and quality of this information. To date, >40 complete microbial genomes sequences have been published and are publicly available[6]; for an updated list of published microbal genomes, see http://www.tigr.org/tdb/mdb/mdbcomplete.html. There are also many incompletely sequenced genomes, which are generally random shotgun data with a three- to eightfold average base coverage, available on non-commercial websites as files of assembled contiguous sequences (contigs; Box 1). The generation of incomplete sequences is much faster and cheaper than the laborious task of linking gaps between contigs to produce finished genomes. However, the pay-off for the time saved generating the data is reduced informational content: not all genes in the organism are represented, the sequence quality is generally lower (especially at the ends of contigs and at repeats) and contaminating DNA sequences are more probable. Additionally, important relationships among genes that comprise the entire genome organization are more difficult to understand when the underlying sequence data are fragmented and unlinked.

The next step is to find genes encoded within the raw sequence data: these are identified in microbial genome sequences using software such as GLIMMER (Ref. 7) or GENEMARK (Ref. 8), which use methods of probability to distinguish coding from non-coding DNA. Both of these algorithms require a training set of DNA sequences from the organism to enable the computer to learn the parameters of a coding sequence. Different gene-finding programs produce different results depending on the size of the training set, the genomic DNA and the program settings[7]. If the gene-finding program is not sufficiently sensitive – undercalling – the missing of open reading frames (ORFs) results (although the sensitivity of the program GLIMMER 2.0 is usually >99%). Overcalling produces a large number of small (<300 nucleotides), overlapping ORFs and the potential fragmentation of larger genes. To find genes from incomplete genomic data, contigs are usually concatenated to form 'pseudomolecules'. Unless care is taken to include translational stop codons between contig breaks this can produce ORFs that are unnatural fusion products.

With a list of predicted genes in hand, the next step is to determine the similarity of the identified genes to databases of previously discovered genes. Historically, the basis for functional assignments of gene products has been sequence similarity with characterized proteins based on BLAST (Ref. 9) or FASTA (Ref. 10) sequence alignments. Often, however, these similarity-based methods fail to

1359-6446/01/$ – see front matter PII: S1359-6446(01)01914-6

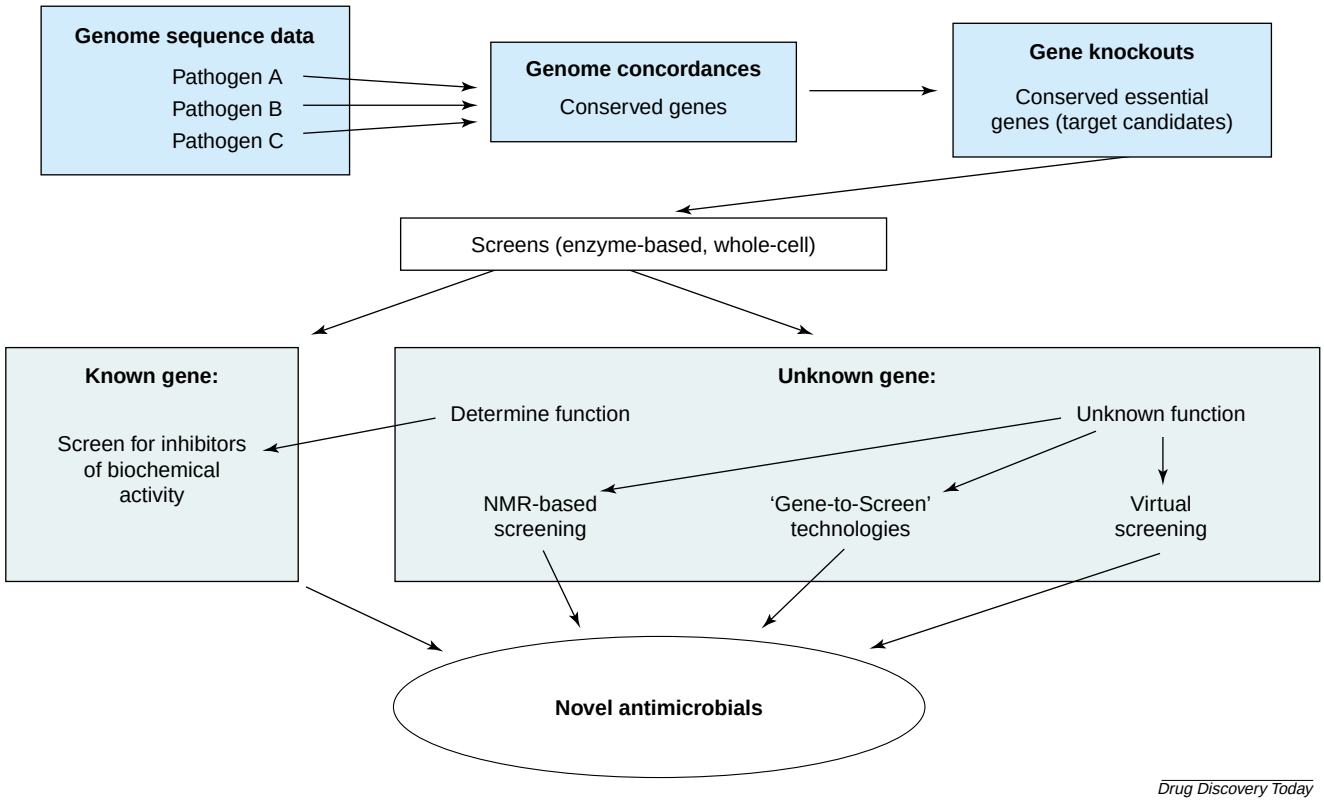

Figure 1. Steps involved in genome-based identification of antimicrobial drug targets. Bioinformatic screening of targets is performed to identify conserved genes, which are then validated by gene knockouts to determine essentiality. Screens are then set up based on function (both known and unknown), and candidate molecules are eventually developed into new antimicrobial drugs.

match proteins that are not closely related. Hidden Markov models [HMMs; e.g. the PFAM (Ref. 11) and TIGRFAM (Ref. 12) sets], PSI-BLAST (Ref. 13) and Clusters of Orthologous Groups (COGs; Ref. 14) are sophisticated alternatives to linear alignments that detect conserved function with greater sensitivity and are contributing increasingly to genome annotation. In addition, there are methods that attempt to take three-dimensional (3D) relationships into account, such as protein-fold recognition by comparison to solved structures[15]. It is important to note that there is no universal standard for genome annotation and that the methodology is still in its developmental phase. This can lead to variation in the annotation of the genome sequences at different centers (e.g. the very similar sequences of *Chlamydia pneumoniae*[16–18]).

Identification and filtering of drug targets

Traditionally, there are two crucial factors that are used to determine potential drug targets among the genes identified in a microbial genome: essentiality and conservation. Essential genes are defined as being essential for growth in rich medium in the laboratory, because this is how the activity of antimicrobial agents is initially tested, but could also include other factors[19], such as survival in a host. Conservation of the drug target among different species of bacteria (antibiotics), fungi (antifungals) or parasites (antiparasitics), and little or no conservation in humans is required for broad-spectrum antimicrobial agents. With the growth of microbial sequence databases, it has now become possible to use *in silico* comparisons among genomes to identify potential targets at the beginning of the drug discovery process (Fig. 1).

To select the specificity of the antimicrobial agent, entire genomes are compared using tools such as Concordance analysis[20] or the TIGR Comprehensive Microbial Resource (CMR; see Boxes 1 and 2) using a variety of parameters. For selecting broad-spectrum targets, sets of genes are selected that are found in several Gram-negative and Gram-positive genomes (e.g. at 25% and 30% identity) and from this set human genes are subtracted. It is also possible to identify genes that are conserved in different strains of one specific pathogen, such as *Helicobacter pylori* or *Mycobacterium tuberculosis*; drugs designed for specific pathogens could be desired when long-term therapy for a known pathogen is

Box 1. A selection of public web pages for access to microbial genome data

Whole genome data and analysis tools	
TIGR Comprehensive Microbial Resource	http://www.tigr.org/tigr-scripts/CMR2/CMRHomePage.spl
NCBI Entrez Microbial Genomes	http://www.ncbi.nlm.nih.gov/PMGifs/Genomes/micr.html
SubtiList (*Bacillus subtilis* genome information)	http://genolist.pasteur.fr/SubtiList
Escherichia coli genomes project	http://www.genome.wisc.edu
Sexually transmitted and oral disease pathogens	http://www.stdgen.lanl.gov
TubercuList (*Mycobacterium tuberculosis*)	http://genolist.pasteur.fr/TubercuList
KEGG (Kyoto Encyclopedia of Genes and Genomes)	http://www.genome.ad.jp/kegg
Partial or unfinished genome data	
TIGR unfinished microbial genomes	http://www.tigr.org/cgi-bin/BlastSearch/blast.cgi?
NCBI finished and unfinished genomes	http://www.ncbi.nlm.nih.gov/Microb_blast/unfinishedgenome.html
Microbial genomes at the US Dept of Energy Joint Genome Institute	http://www.jgi.doe.gov
Sanger Centre unfinished microbial genomes	http://www.sanger.ac.uk/Projects/Microbes
University of Oklahoma Genome Center	http://www.genome.ou.edu/

required, to minimize effects on normal flora in the body and the development of drug resistance. Antimicrobials for selected groups of pathogens can also be targeted using comparative genomic techniques. For example, genes that are specifically conserved among the upper-respiratory pathogens *Streptococcus pneumoniae, Haemophilus influenzae* and *Neisseria meningitidis*, could provide an insight into common strategies used to survive in this niche and might represent good targets for targeted therapy. Using a 25% identity threshold, 32 *S. pneumoniae* genes are conserved in either *H. influenzae* or *N. meningitidis* and not the non-pathogenic strains *Escherichia coli* K12 or *Bacillus subtilis*. The list contains some commonly encountered proteins, as well as some gene products whose functions are unknown

Box 2. Demonstration of target screening by comparative genomics using the TIGR Comprehensive Microbial Resource (CMR; Fig. I)

Chlamydia pneumoniae and *Rickettsia prowazekii* are human intracellular pathogens with reduced genomes but are deeply diverged on an evolutionary timescale. The strategy used was to find genes common to *C. pneumoniae* (1.2 Mb; Ref. 16) and *R. prowazekii* Madrid E (1.1 Mb; Ref. 49) but not found in the much larger, free-living non-pathogenic bacteria *Escherichia coli*[50] and *Bacillus subtilis*[51]. The CMR contains information from complete microbial genome sequencing projects. All genomic proteins are searched, using BLASTP, against all other proteins and the results saved in a database allowing convenient on-line comparisons (Fig. I). *C. pneumoniae* proteins were screened for ≥25% identity with proteins from *R. prowazekii*. All proteins with ≥25% identity to *E. coli* and *B. subtilis* were excluded. The list (in the bottom window of the screen shot; Fig. I) includes proteins known to be conserved between the bacteria[52] (e.g. ATP translocases and chloroplast-like ribosomal protein) as well as several conserved hypothetical proteins that could prove promising targets.

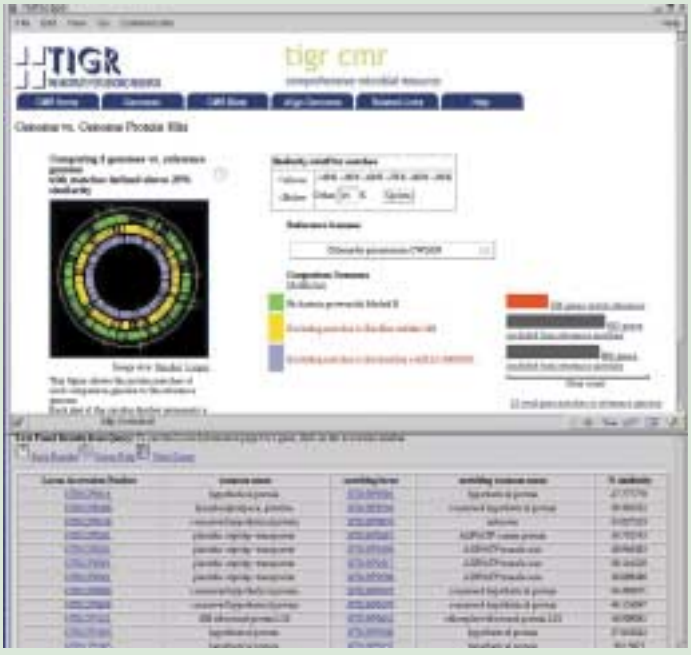

Figure I

or improperly understood but could, nevertheless, have potential uses in therapy. Of the 32 genes, two are methionine sulfoxide reductases, which are potential virulence determinants conserved in other pathogens, such as *H. pylori* and *Vibrio cholerae*. It is increasingly possible to perform this type of straightforward bioinformatic analysis using public websites (e.g. Box 2; Fig I), and comparisons could be performed using more sensitive techniques, such as HMM models or COGs instead of BLAST similarity[21].

Although the focus of this review is drug development, it is clear that genomics is also being used to drive vaccine development. Whereas the treatment of most infections with antimicrobial agents is acute, vaccines are used as a prophylactic tool, often targeting infections first encountered during childhood. In the long term, management of infectious disease by vaccines and other prophylactic approaches would seem preferable, assuming high efficacy rates, lower costs and minimal drug resistance. One example of using a genome sequence to drive vaccine discovery is for *N. meningitidis* serogroup b[22,23], where researchers analyzed the genome for gene products that were predicted to have an outer-membrane subcellular localization, using the TopPred (Ref. 24), SignalP (Ref. 25) and Psort (Ref. 26) programs, which are conserved among *N. meningitidis*. Candidate antigens (~350) were systematically expressed and purified, and used to immunize mice; those that induced a bactericidal antibody response, which correlates with vaccine efficacy in humans, were selected. A similar approach to that of *N. meningitidis* was recently published for *S. pneumoniae*[27] and, as with drug discovery efforts, numerous genome-based vaccine efforts are being undertaken in the pharmaceutical and biotechnology industry. It is important to note that, because of their outer-surface localization, the proteins selected by this type of bioinformatic filtering are also potential targets for antimicrobial therapy because inhibitors would not need to cross bacterial cell membranes[21,28,29].

The development of genome-based discovery approaches opens the way for genes that are not under the '*in vitro* essential' description that have commonly been targeted by the pharmaceutical industry[28]. Genes that are necessary for microbial survival in the human host offer intriguing targets for therapeutic intervention. Besides targeting acute antimicrobial therapy, gene products that promote disease (e.g. colonization factors such as pili) might also be useful targets for prophylaxis. Comparing closely related genomes and finding large pathogenicity islands[30] or lists of genes present specifically in a virulent organism (e.g. the 1300 genes present in the enterohemorrhagic *E. coli* O157:H7 but not found in the laboratory *E. coli* K12 strain[31]) could help to build evidence for newly identified genes that might have an important role in infection. It is possible to imagine that therapies based on non-essential genes might prevent infection without killing the pathogen, thus avoiding problems such as inflammation caused by lysed cellular materials or disruption of the normal host microflora. It will be interesting to see how these approaches, which take microbial pathogenesis and host interactions into account, succeed relative to the 'classical' antimicrobial drug techniques.

Genomic approaches to determining gene function

The assignment of gene function is crucial for genome-based drug discovery because many screening strategies make use of the biochemical activity of a gene product (Fig. 1). However, many proteins that are identified, including several that are conserved in numerous organisms, still have no annotated biological function. Of the ORFs in the genomes of the Lyme disease pathogen *Borrelia burgdorferi* and the gastric bacterium *H. pylori*, 65% and 45%, respectively, encode products of unknown function[32]. There are several genome-based strategies that have the potential to chart the *terra incognita* of protein function, including microarray-based expression technology (discussed later), proteomics[32], protein–protein interaction studies and large-scale mutagenesis studies (e.g. the *Mycoplasma* minimal gene project[33] and the yeast functional analysis project[34]). Additionally, there are technologies that establish relationships between genes and pathogenicity that are being applied at the whole-genome level[35]. These include the *in vitro* expression technology, IVET, which aims to identify genes expressed during infection, and signature-tagged mutagenesis (STM), which identifies mutations with reduced fitness when passaged through an animal or cell model. These techniques, when used in combination with previous knowledge of microbial physiology[36,37], should assist in determining the roles of these unknown genes in the microbe. If not, technologies are available for screening potential drug targets of unknown function, such as NMR-based screening[38], 'gene-to-screen' technologies[39] and virtual screening[40] (Fig. 1).

DNA microarray technology

Although there are numerous laboratory 'functional genomics' technologies to help determine gene function, one exciting area in the application of genomics to drug discovery is the development of DNA microarray technology[41]. Microarrays can be used to identify drug candidates that are effective in controlling bacterial growth and infectivity. In one approach, the microarrays are used to monitor bacterial gene expression in response to selected environmental conditions[42–44]. This microarray analysis of gene

expression at the genomic level enables investigators to understand the mechanistic basis of many drugs and design novel drugs based on newly identified regulatory pathways or networks. For example, growth and virulence of many bacterial pathogens is controlled by several regulatory factors, some of which have been identified recently from the genome sequences. Whole-genome analysis of gene expression will enable us to determine what role these regulatory factors play in virulence and to develop new drugs to inhibit their activity. Recently, this approach has been used with *H. influenzae* to explore transcriptional responses triggered by exposure to novobiocin or ciprofloxacin[44] and in *S. pneumoniae* to investigate transcriptional responses of competence[42] and quorum-sensing[43] systems.

DNA microarrays can also be used as a comparative genomics tool for identifying differences in DNA between two related bacterial isolates[45–47]. An excellent example of this approach[45] used DNA microarrays of the *Mycobacterium tuberculosis* genome to identify differences between attenuated (BCG) strains of *Mycobacterium bovis* and *M. tuberculosis*. These live attenuated strains are currently used as a vaccine to prevent tuberculosis. However, continual passage of the BCG isolates has led to a dispersion of substrains with differing levels of virulence and immunogenicity. The *M. tuberculosis* DNA microarrays were used for hybridization experiments with genomic DNA from several BCG isolates to screen for ORFs or genes deleted from BCG strains when compared to *M. tuberculosis*. Sixteen regions, varying in length from 1903 to 12,733 bp, were deleted in BCG. Genes identified within these deleted regions encode likely virulence factors, so they are prime candidate targets for generating a more effective BCG vaccine. Another important potential application for drug discovery is the use of microarrays to screen populations of clinical strains for conserved genes, because it is vital that targets are conserved in the spectrum of relevant natural isolates.

An alternative to investigating gene expression in the pathogen is gene expression of the host in response to challenges with the pathogen[48]. DNA microarrays containing all ORFs from the genomes of several models of infectious disease (in human, rat and mouse) are now, or imminently, available for these experiments. In this case, elucidation of the host response to a pathogen could help determine the most potentially effective drug candidate.

Outlook

In this review we have discussed how the recent surge in genomic sequence data of microbial pathogens can be used to enhance the search for targets for drug and vaccine development. Clinically used antimicrobials inhibit a few dozen targets, whereas microbial genome sequencing has revealed thousands of new potential targets. By using a combination of computational and experimental approaches, the 'black box' of using DNA sequences to design new antimicrobial drugs is being defined.

So, how will genomics impact drug discovery? Since the mid-1990s, genome-sequence information, advances in bioinformatics, and high-throughput biology, have helped drive the discovery of antimicrobial drugs. The first advances were with bacteria, given their small genome size and genetic amenability, followed by fungi and parasites. With the near completion of the human and mouse genomes, the lessons learned from drug discovery in the simpler microbial systems can be applied to human therapeutics. The addition of genomic sequence data from non-pathogens will also aid genome-based drug discovery, improving functional assignment by adding breadth to protein family trees, as well as serving to highlight the unique features of disease-causing microbes. Information on temporal patterns of gene expression, and protein SARs will also increase, with the application of 3D structural modeling opening up exciting avenues for rational drug design.

One certainty of the genomic era is that data from sequencing projects, large-scale bioinformatic searches and high-throughput biological experiments will continue to accumulate. This avalanche of information is sometimes regarded as a curse, straining computing resources and human comprehension. However, if the input data are of high quality, are organized and effectively linked, and are thoughtfully filtered and analyzed, genomics will continue to be the driving-force for the future of drug discovery.

Acknowledgements

We would like to thank Tanja Dickinson and Sam Angiuoli for help with the Comprehensive Microbial Resource. Timothy D. Read is supported by grants from the US Department of Energy, Office of Naval Research and National Institutes of Health.

References

1 Cohen, M.L. (2000) Changing patterns of infectious disease. *Nature* 406, 762–767
2 Allsop, A.E. (1998) New antibiotic discovery, novel screens, novel targets and impact of microbial genomics. *Curr. Opin. Microbiol.* 1, 530–534
3 Cockett, M. *et al.* (2000) Applied genomics: integration of the technology within pharmaceutical research and development. *Curr. Opin. Biotechnol.* 11, 602–609
4 Gray, C.P. and Keck, W. (1999) Bacterial targets and antibiotics: genome-based drug discovery. *Cell Mol. Life Sci.* 56, 779–787
5 Rosamond, J. and Allsop, A. (2000) Harnessing the power of the genome in the search for new antibiotics. *Science* 287, 1973–1976

6 Nelson, K.E. *et al.* (2000) Status of genome projects for nonpathogenic bacteria and archaea. *Nat. Biotechnol.* 18, 1049–1054

7 Salzburg, S.L. *et al.* (1998) Microbial gene identification using interpolated Markov models. *Nucleic Acids Res.* 26, 544–548

8 Hayes, W.S. and Borodovsky, M. (1998) How to interpret an anonymous bacterial genome: machine learning approach to gene identification. *Genome Res.* 8, 1154–1171

9 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

10 Pearson, W.R. (2000) Flexible sequence similarity searching with the FASTA3 program package. *Methods Mol. Biol.* 132, 185–219

11 Bateman, A. *et al.* (1999) Pfam3.1: 1313 multiple alignments and profile HMMs match the majority of proteins. *Nucleic Acids Res.* 27, 260–262

12 Haft, D.H. *et al.* (2001) TIGRFAMs: a protein family resource for the functional identification of proteins. *Nucleic Acids Res.* 29, 41–43

13 Altschul, S.F. and Koonin, E.V. (1998) Iterated profile searches with PSI-BLAST – a tool for discovery in protein databases. *Trends Biochem. Sci.* 23, 444–447

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

15 Domingues, F.S. *et al.* (2000) Structure-based evaluation of sequence comparison and fold recognition alignment accuracy. *J. Mol. Biol.* 297, 1003–1013

16 Kalman, S. *et al.* (1999) Comparative genomes of *Chlamydia pneumoniae* and *C. trachomatis*. *Nat. Genet.* 21, 385–389

17 Read, T.D. *et al.* (2000) Genome sequences of *Chlamydia trachomatis* MoPn and *Chlamydia pneumoniae* AR39. *Nucleic Acids Res.* 28, 1397–1406

18 Shirai, M. *et al.* (2000) Comparison of whole genome sequences of *Chlamydia pneumoniae* J138 from Japan and CWL029 from USA. *Nucleic Acids Res.* 28, 2311–2314

19 Peterson, S.N. and Fraser, C.M. (2001) The complexity of simplicity. *Genome Biol.* 2, Comment 2002.1–2002.8

20 Bruccoleri, R.E. *et al.* (1998) Concordance analysis of microbial genomes. *Nucleic Acids Res.* 26, 4482–4486

21 Galperin, M.Y. and Koonin, E.V. (1999) Searching for drug targets in microbial genomes. *Curr. Opin. Biotechnol.* 10, 571–578

22 Pizza, M. *et al.* (2000) Identification of vaccine candidates against serogroup B meningococcus by whole-genome sequencing. *Science* 287, 1816–1820

23 Tettelin, H. *et al.* (2000) Complete genome sequence of *Neisseria meningitidis* serogroup B strain MC58. *Science* 287, 1809–1815

24 Claros, M.G. and von Heijne, G. (1994) TopPred II: an improved software for membrane protein structure predictions. *Comput. Appl. Biosci.* 10, 685–686

25 Nielsen, H. *et al.* (1997) A neural network method for identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Int. J. Neural Syst.* 8, 581–599

26 Nakai, K. and Kanehisa, M. (1991) Expert system for predicting protein localization sites in Gram-negative bacteria. *Proteins* 11, 95–110

27 Wizemann, T.M. *et al.* (2001) Use of a whole genome approach to identify vaccine molecules affording protection against *Streptococcus pneumoniae* infection. *Infect. Immun.* 69, 1593–1598

28 Alksne, L.E. and Projan, S.J. (2000) Bacterial virulence as a target for antimicrobial chemotherapy. *Curr. Opin. Biotechnol.* 11, 625–636

29 Alksne, L.E. *et al.* (2000) Identification and analysis of bacterial protein secretion inhibitors utilizing a SecA-LacZ reporter fusion system. *Antimicrob. Agents Chemother.* 44, 1418–1427

30 Hacker, J. *et al.* (1997) Pathogenicity islands of virulent bacteria: structure, function and impact on microbial evolution. *Mol. Microbiol.* 23, 1089–1097

31 Perna, N.T. *et al.* (2001) Genome sequence of enterohaemorrhagic *Escherichia coli* O157:H7. *Nature* 409, 529–533

32 Nierman, W. *et al.* (2000) Microbial genome sequencing 2000: new insights into physiology, evolution and expression analysis. *Res. Microbiol.* 151, 79–84

33 Hutchison, C.A. *et al.* (1999) Global transposon mutagenesis and a minimal Mycoplasma genome. *Science* 286, 2165–2169

34 Dujon, B. (1998) European Functional Analysis Network (EUROFAN) and the functional analysis of the *Saccharomyces cerevisiae* genome. *Electrophoresis* 19, 617–624

35 Chiang, S.L. *et al.* (1999) *In vivo* genetic analysis of bacterial virulence. *Annu. Rev. Microbiol.* 53, 129–154

36 Karp, P.D. *et al.* (1999) Integrated pathway-genome databases and their role in drug discovery. *Trends Biotechnol.* 17, 275–281

37 Selkov, E. *et al.* (2000) Functional analysis of gapped microbial genomes: amino acid metabolism of *Thiobacillus ferrooxidans*. *Proc. Natl. Acad. Sci. U. S. A.* 97, 3509–3514

38 Shuker, S.B. *et al.* (1996) Discovering high-affinity ligands for proteins: SAR by NMR. *Science* 274, 1531–1534

39 Hyde-DeRuyscher, R. *et al.* (2000) Detection of small-molecule enzyme inhibitors with peptides isolated from phage-displayed combinatorial peptide libraries. *Chem. Biol.* 7, 17–25

40 Langer, T. and Hoffman, R.D. (2001) Virtual screening: an effective tool for lead structure discovery? *Curr. Pharm. Des.* 7, 509–527

41 Brown, P.O. and Botstein, D. (1999) Exploring the new world of the genome with DNA microarrays. *Nat. Genet.* 21, 33–37

42 Peterson, S. *et al.* (2000) Gene expression analysis of the *Streptococcus pneumoniae* competence regulons by use of DNA microarrays. *J. Bacteriol.* 182, 6192–6202

43 de Saizieu, A. *et al.* (2000) Microarray-based identification of a novel *Streptococcus pneumoniae* regulon controlled by an autoinduced peptide. *J. Bacteriol.* 182, 4696–4703

44 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

45 Behr, M.A. *et al.* (1999) Comparative genomics of BCG vaccines by whole-genome DNA microarray. *Science* 284, 1520–1523

46 Hakenbeck, R. *et al.* (2001) Mosaic genes and mosaic chromosomes: intra- and interspecies genomic variation of *Streptococcus pneumoniae*. *Infect. Immun.* 69, 2477–2486

47 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

48 Debouck, C. and Goodfellow, P.N. (1999) DNA microarrays in drug discovery and development. *Nat. Genet.* 21, 48–50

49 Andersson, S.G. *et al.* (1998) The genome sequence of *Rickettsia prowazekii* and the origin of mitochondria. *Nature* 396, 133–140

50 Blattner, F.R. *et al.* (1997) The complete genome sequence of *Escherichia coli* K-12. *Science* 277, 1453–1474

51 Kunst, F. *et al.* (1997) The complete genome sequence of the Gram-positive bacterium *Bacillus subtilis*. *Nature* 390, 249–256

52 Wolf, Y.I. *et al.* (1999) Rickettsiae and Chlamydiae: evidence of horizontal gene transfer and gene exchange. *Trends Genet.* 15, 173–175