

## Marrying structure and genomics

### Burkhard Rost

Address: European Molecular Biology Laboratory, 69 012 Heidelberg, Germany.

E-mail: rost@EMBL-Heidelberg.DE

**Structure** 15 March 1998, 6:259–263  
<http://biomednet.com/elecref/0969212600600259>

© Current Biology Ltd ISSN 0969-2126

#### Introduction

##### Today

Large-scale genome sequencing is filling up the catalogue of natural proteins at a breathtaking speed. Today, we have available not just a large number of sequences, but also glimpses of the inventory of entire organisms. This information will soon improve our understanding of cells and of life in general. Three means will contribute to this expanding body of knowledge: sequencing genomes (genomics); the determination of protein structures; and the determination of protein function. Protein structure is interwoven with function (e.g. [1–3]). Sequence analysis and determination of function are also routinely combined (e.g. [4]). What about the relationship between structure determination and genomics, however?

##### Tomorrow

Structural genomics, the marriage between protein structure determination and genomics, is already beginning.

Attempts are made here to illustrate the likely direction this marriage will take. Structure determination will be pushed by, and profit from, genomics. Furthermore, basing research and technical developments, such as drug design, on all three pillars (sequence, structure and function) will provide a large step towards the understanding of life.

#### Objectives

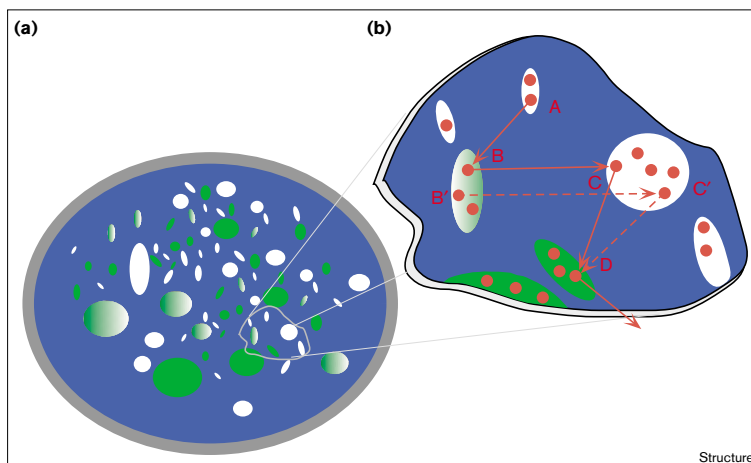
Structure determination will benefit from genomics in two ways (Figure 1). Firstly, the mass of available sequences will facilitate the quick determination of structure for most existing folds. Secondly, the availability of sequences for the entire genome of an organism will not only help us to unravel missing links in functional pathways, but also to explore alternative pathways and to widen our understanding of principle mechanisms and evolutionary cross-links.

#### Genomics and structure: two flourishing fields

The first sequence of an entire genome of an organism was published in 1995. Two years on, another ten complete genome sequences have been published (Table 1). Nucleotide databases have increased two times more over the last two years, than in the previous 20 years (Figure 2). The growth of these databases now outpaces even the development of computers (Figure 2). This is merely the beginning.

**Figure 1**

Objectives for structure determination in the era of genomics. **(a)** The first objective is to utilise the mass of known protein sequences to determine all natural folds. The ellipsoid symbolises the universe of protein structures and the islands symbolise existing structures (or folds). Some of the islands are larger as some folds occur more often (e.g. TIM barrel, immunoglobulin-like, NTP hydrolase, ferredoxin-like, Rossmann fold, globin-like, flavodoxin-like and Ribonuclease H fold [15]). The colour coding distinguishes three situations: most structures are known (green); some structures (i.e. the principle folds are known; half green, half white); no structure is known (white). **(b)** The second objective is to utilise the entire genome sequences of organisms to fill in the missing links in pathways and mechanisms. The red circles indicate sequence families (sequences within a family have significant levels of pairwise sequence identity), the solid arrows symbolise a well known pathway in organism X ( $A \rightarrow B \rightarrow C \rightarrow D$ ) and the dashed arrows symbolise the analogous pathway in organism Y ( $A \rightarrow B' \rightarrow C' \rightarrow D$ ). If proteins B and C do not exist in organism Y, this pathway cannot be



mapped from knowing all sequences of Y. Imagine we know the structure of B, then threading (fold recognition) may enable us to deduce that B' adopts the role of B. Without knowing the structures of C or C', however,

we still could not guess the interaction partner of B', and thus still could not map the pathway. Knowing structures for all sequence families (all red circles) it would be possible to easily find the pathway in Y.

Table 1

## Completely sequenced genomes\*.

Genome	Date	Reference
<i>Haemophilus influenzae</i>	8/95	[22]
<i>Mycoplasma genitalium</i>	10/95	[23]
<i>Saccharomyces cerevisiae</i>	1/96	[24]
<i>Methanococcus jannaschii</i>	8/96	[25]
<i>Synechocystis</i> sp. PCC6803	9/96	[26]
<i>Mycoplasma pneumoniae</i>	11/96	[27]
<i>Escherichia coli</i>	1/97	[28]
<i>Methanobacterium thermoautotrophicum</i>	5/97	[29]
<i>Archaeoglobus fulgidus</i>	6/97	[30]
<i>Helicobacter pylori</i>	6/97	[31]
<i>Borrelia burgdorferi</i>	7/97	[32]
<i>Treponema pallidum</i>	10/97	†
<i>Bacillus subtilis</i>	11/97	[33]
<i>Pyrococcus horikoshii</i>	1/98	‡
<i>Aquifex aeolicus</i>	2/98	[34]

\*List obtained from T Gaasterland, <http://www.mcs.anl.gov/home/gaasterl/genomes.html>. †Sequences publicly available (CM Fraser *et al.*). ‡Sequences partially publicly available [35].

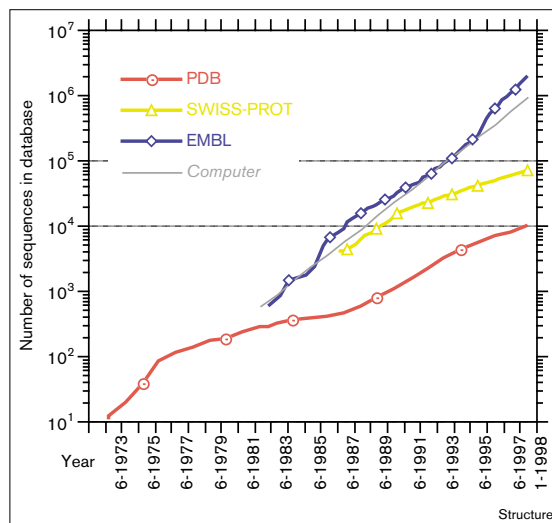
Structure determination has now become almost routine [5]. Currently, as many structures are determined every ten days as in the first ten years of crystallography (Figure 2). Hitting on a novel fold, however, still resembles unearthing a nugget [1,2]. Each novel fold can contribute towards understanding the functional details of entire protein families. How does the rate of determination of new structures compare to the rate of determination of sequence data? Considering the organisms for which the entire genome sequence is known (Figure 3), we have structural knowledge for one in every ten protein sequences. Three projects have recently been initiated to solve structures systematically for all the proteins within an organism (<http://www.mcs.anl.gov/home/gaasterl/sg-review.html>): *Haemophilus influenzae* (J Moulton, Centre for Advanced Research in Biotechnology [CARB], in collaboration with the Institute for Genomic Research [TIGR]); *Pyrobaculum aerophilum* (T Terwillinger, Lawrence Livermore National Laboratory [LANL]; D Eisenberg and J Miller, University of California Los Angeles [UCLA]); and *Methanococcus jannaschii* (S-H Kim, Lawrence Berkeley National Laboratory [LBNL]). With all this wealth of information, what are the objectives of structural genomics?

#### Using the mass of sequences to populate each island in structure space

##### Filling the blank spots in structure space

Already we know approximately 500 [1] of the estimated 1000 protein folds [6,7]. Thus, only about half the 'blank spots' in structure space remain to be filled (Figure 1a). Optimistic or not, the first objective for structural genomics will be to determine most water-soluble native folds. Genomics can facilitate finding the blank spots. The recipe is simple: find proteins common to different organisms; exclude those

Figure 2



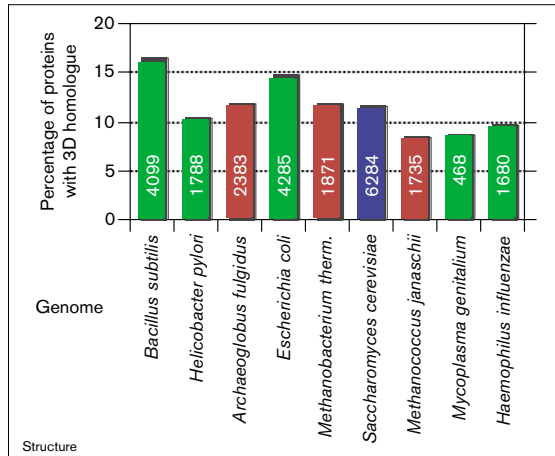
The rate of growth of databases of biomolecules. The explosion of biomolecular data is illustrated by three representative databases: the protein structure database PDB [16] (red line, circles); the protein sequence database SWISS-PROT [17] (yellow line, triangles); and the nucleotide sequence database EMBL [18] (blue line, diamonds). The number of symbols (circles, triangles, diamonds) roughly reflects the number of releases of the respective database. For comparison the growth of computer speed is also shown (grey thin line; speed doubles every 18 months).

with structural homologues (10%; Figure 3); exclude integral membrane proteins (20–30%; Figure 4); and exclude all proteins for which threading detects known folds (<10%) [8,9]. Arriving at the final list requires a large repository of sequences and some skills in bioinformatics. The mass of sequences yielded by genomics will help surmount essential problems in structure determination (e.g. protein expression, purification, and — for crystallography — growth of crystals). For each blank spot candidate, research groups can select the homologue in their favourite organism, (e.g. from thermophilic bacteria, where proteins have the advantage of remaining stable at high temperatures). How likely is a structure, thus selected, to have a novel fold? Today, the specific goal to find novel folds is not driving structure determination. Nevertheless, 10–30% of the structures added to the Protein Data Bank (PDB) constitute novel folds [1]. A large-scale structure determination enterprise could easily yield 2000 (additional) new structures annually. Thus, we shall have at least one representative structure for every fold in less than a decade (assuming an initial 10% yield of novel folds, this yield then decaying exponentially).

##### Adding details to the map

Most pairs of similar structures have <15% pairwise sequence identity (Figure 5). Thus, filling all the blank

Figure 3



Percentage of proteins in genomes with homologues of known structure. For nine whole genomes, the figure shows the number of proteins for which a sequence homologue of known structure exists in the PDB as a percentage of all identified proteins (the total numbers of proteins used are written on the bars). The genomes of eukaryotes are shown in blue, prokaryotes are in green, and archae are in red. For about 80% of the proteins considered here, the structure is inferred by homology modelling [19]. In addition, the estimates are conservative in that high homology thresholds have been applied. Homologues inferred by threading methods were not considered.

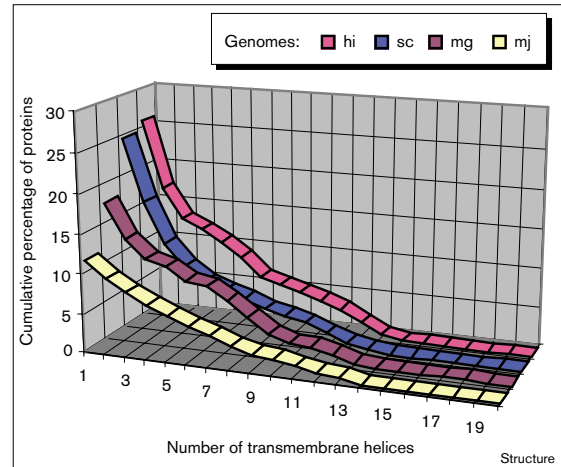
spots does not yield all families populating the respective island (Figure 1). The enormous sequence variation within islands is often associated with functional divergence (or convergence). In order to use structure determination to help further our understanding of protein function, the next goal of structural genomics will be to determine structures for all sequence families (and preferably for more than one representative per family). How many structures would it take to fill the map with such detail? Currently, the structures are known for 1145 proteins of unique sequence (the set used in Figure 5) [10], and these represent about 10% of known genomes (Figure 3). Thus, about 10,000 additional structures are required to provide one structure per sequence family). This second phase, however, yields 100% coverage in a large-scale structure determination project (the recipe described above only selects candidates representing a single sequence family). Thus, assuming a moderate production of 2000 structures annually, approximately twofold coverage should be obtained within a decade.

#### Using the entirety of organisms to cover all functional elements

##### Finding missing links in pathways

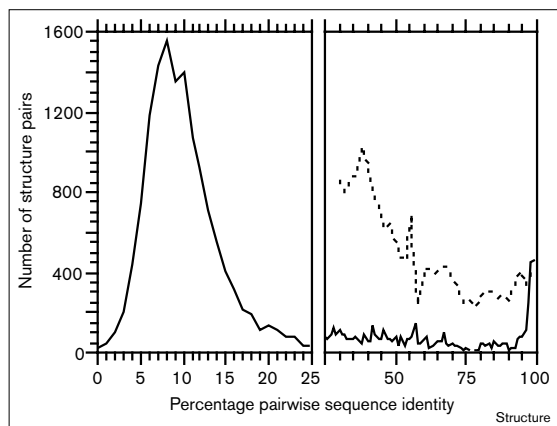
Knowing all the protein sequences for entire organisms, we can start to map them to pathways (e.g. metabolic,

Figure 4



The percentage of helical transmembrane proteins in organisms. For the first four genomes to be entirely sequenced the percentage of proteins from the genome predicted by PHD topology [20] to have helical membrane regions is shown. The four organisms are colour-coded: mj, *Methanococcus jannaschii* (yellow); mg, *Mycoplasma genitalium* (purple); sc, *Saccharomyces cerevisiae* (blue); hi, *Haemophilus influenzae* (magenta). The horizontal axis gives the number of transmembrane helices predicted; the vertical axis gives the cumulative percentage of proteins in the genome. For example, about 25% of the proteins of *H. influenzae* and *S. cerevisiae* are predicted to have at least one transmembrane helix; 7% of the proteins of *H. influenzae* and 5% of the proteins of *S. cerevisiae* are predicted to contain seven or more transmembrane helices. (A list of the proteins is available at: <http://www.embl-heidelberg.de/~rost>).

regulatory, signalling, pathogenic), or particular mechanisms (e.g. expression, transcription, replication, recombination) [11]. Suppose we miss one (or a couple) of the proteins essential for a particular pathway. Can we conclude that this pathway is missing in the organism, or should we try harder to find it? The answer provides the second objective for structural genomics: to find functionally missing links (Figure 1b). Initially this objective will aim to determine the missing structures for all major pathways and mechanisms. The first step, to complete the structural knowledge for all pathways and mechanisms for which we know the associated proteins, is straightforward. The second step, however, appears hopeless: how can we determine structures for unknown proteins? In reality this may not prove to be as difficult as it would at first seem. It is likely that many of the candidates selected to find all the blank spots in structural space will turn out to be representatives of most major functional protein classes. In addition, in the course of large-scale structure determination, cross-links will be uncovered that complete the catalogue of proteins participating in certain functions (e.g. the corresponding mechanisms identified in fragile

**Figure 5**

The evolution of protein structures into the 'midnight zone' of sequence identity. Proteins have evolved into the midnight zone of sequence identity (i.e. into a region where sequence comparisons fail completely to detect structural similarity [21]). The figure shows the distribution of pairwise sequence identity for structurally aligned protein pairs (full line). The average pairwise sequence identity of all remotely structurally similar pairs (<25% sequence identity; left panel) is below 10%. To reduce database bias, results displayed in the left panel are based on a smaller data set (aligning 1145 structures of unique sequence against themselves) than those displayed in the right panel (aligning 1145 sequences against the entire PDB). Consequently, numbers in the right panel should be scaled down. To obtain a perspective less biased by the choice of proteins for which structures are determined, numbers are also given for a subset of SWISS-PROT for which homology modelling is applicable (dashed line on right panel; aligning 1145 structures of unique sequence against the entire SWISS-PROT database).

histidine triad protein (FHIT) and protein kinase C interacting protein (PKCI) and the implications of their structural similarity to galactose-1-phosphate uridylyl-transferase (GalT) [3]).

#### *Filling function space*

After determining structures for all major functional elements, we shall have to complete the functional map (i.e. it will be necessary to determine structures for representatives of all pathways and mechanisms). Candidates for those structures to be determined will be found by structure-based comparative genome analysis, focusing on particular sites (e.g. active sites and binding sites) or uncovering 'motifs' [1]. For example, the goal could be to find the scaffold containing the common features of all amino hydrolases [12]. Furthermore, alternative pathways will be searched, as well as proteins with particular biochemical 'fingerprints' (the structures of such proteins will be crucial to correctly define the motifs). Finally, unknown functions could be searched for specifically by classifying families of determined and homology-modelled structures into functional groups based on

electrostatic properties [13], or based on simple combinations of sequence alignment and structure analysis [14].

#### **Conclusions**

##### *Profiting from mass and entirety*

The major objectives of structural genomics have been portrayed here: to find all natural structures; and to find missing links in all functional pathways and mechanisms (Figure 1). These objectives correspond to two aspects of genome sequencing: the mass of sequences produced; and the entirety of sequencing complete genomes from organisms. In order to attain the objectives outlined here a large-scale structure determination enterprise is required.

##### *What will come out?*

A prerequisite for understanding the function of a protein is to know its structure. Furthermore, large-scale structure determination will enable us to uncover most major functional elements. The scaffolds of structures provide the elements for evolution. Most functional motifs known today are sequence motifs. In the absence of structural data, however, most functional motifs remain hidden. Structural genomics will help us to further understand evolution, and will also provide the knowledge necessary to improve the techniques used in processes such as drug design and discovery. Finally, entities defined by refined structural [8,15] and functional features [1,2] will permit a more elaborate comparison of organisms than sequence analysis.

##### *What will not come out?*

This review has focused on the description of structural modules, or domains. Clearly, domains are not enough to understand function. Instead, we need to study functional complexes composed of many proteins. Although a large-scale structure determination enterprise may trigger the study of such complexes by uncovering their elements, a comprehensive exploration of functional systems will be the next step.

##### *When will we get there?*

Humans have about 100,000 different proteins. If we knew all these sequences today, through a combination of structure determination and prediction we would already have structural knowledge for more than 10,000 of these (Figure 3). The sequence of the human genome, however, will not be completed before the year 2004. With 2000 new structures determined by a large-scale enterprise, we shall have structural knowledge for about 70% of all human sequences by the year 2004; many of the remaining 30,000 will be membrane proteins (Figure 4).

##### *Reality or dream?*

The mass of sequences produced by genomics should enable most natural folds to be determined within less than a decade. Is this wishful thinking? Firstly, the — strongly disputed — assumption that there are only 1000 folds is not

crucial. Instead, the upper limit for the number is provided by the number of sequence families, and the estimate that there are 10,000–15,000 families is rather conservative (to date 1200 sequence families are known, corresponding to 8–18% of all families; Figure 3). To determine one structure for each family is just a matter of a large-scale structure determination enterprise. Secondly, 2000 structures were added to the PDB in 1997, and structure determination techniques continue to improve. Thus, the assumption that 2000 new structures will be determined annually is a rather conservative estimate. What remains is the uncertainty as to how difficult the unknown folds will be to determine. Here we can only be guided by past experience, which shows that most structure determination problems can be solved — eventually. Of course there is no easy answer, we just have to try.

#### Supplementary material

Supplementary material available with the internet version of this paper contains a diagram showing the percentage of protein structures annually deposited in the PDB from a particular organism.

#### Acknowledgements

Thanks to Alfonso Valencia (CNB Madrid), John Moulton (CARB Washington) and Alexei Murzin (MRC Cambridge) for discussions; to Sean O'Donoghue (EMBL Heidelberg) and Terry Gaasterland (University of Chicago/Argonne) for proof-reading and discussions; to the GeneQuiz consortium (Miguel Andrade, Nigel Brown, Christophe Leroy and Chris Sander, Ebi Hinxton) for permission to use their unpublished data for Figure 3; and to Chris Sander (Millennium Boston), and Matti Saraste (EMBL Heidelberg) for financial support.

#### References

- Murzin, A.G. (1996). Structural classification of proteins: new superfamilies. *Curr. Opin. Struct. Biol.* **6**, 386-394.
- Holm, L. & Sander, C. (1997). New structure – novel fold? *Structure* **5**, 165-171.
- Lima, C.D., Klein, M.G. & Hendrickson, W.A. (1997). Structure-based analysis of catalysis and substrate definition in the HIT protein family. *Science* **278**, 286-290.
- Warbrick, E. (1997). Two's company, three's a crowd: the yeast two hybrid system for mapping molecular interactions. *Structure* **5**, 13-17.
- Lattman, E.E. (1994). Protein crystallography for all. *Proteins* **18**, 103-106.
- Finkelstein, A.V. & Ptitsyn, O.B. (1987). Why do globular proteins fit the limited set of folding patterns? *Prog. Biophys. Molec. Biol.* **50**, 171-190.
- Chothia, C. (1992). One thousand protein families for the molecular biologist. *Nature* **357**, 543-544.
- Fischer, D. & Eisenberg, D. (1997). Assigning folds to the proteins encoded by the genome of *Mycoplasma genitalium*. *Proc. Natl. Acad. Sci. USA* **94**, 11929-11934.
- Rost, B. & O'Donoghue, S.I. (1997). Sisyphus and prediction of protein structure. *Cabios* **13**, 345-356.
- Holm, L. & Sander, C. (1998). Touring protein fold space with DALI/FSSP. *Nucleic Acids Res.* **26**, 318-321.
- Gaasterland, T. & Sensen, C.W. (1996). Fully automated genome analysis that reflects user needs and preferences – a detailed introduction to the MAGPIE system architecture. *Biochimie* **78**, 302-310.
- Brannigan, J.A., et al., & Murzin, A.G. (1995). A protein catalytic framework with an N-terminal nucleophile is capable of self-activation. *Nature* **378**, 416-419.
- Blomberg, N. & Nilges, M. (1997). Functional diversity of PH domains: an exhaustive modelling study. *Fold. Des.* **2**, 343-355.
- Lichtarge, O., Bourne, H.R. & Cohen, F.E. (1996). An evolutionary trace method defines binding surfaces common to protein families. *J. Mol. Biol.* **257**, 342-358.
- Gerstein, M. & Levitt, M. (1997). A structural census of the current population of protein sequences. *Proc. Natl. Acad. Sci. USA* **94**, 11911-11916.
- Bernstein, F.C., et al., & Tasumi, M. (1977). The protein data bank: a computer based archival file for macromolecular structures. *J. Mol. Biol.* **112**, 535-542.
- Bairoch, A. & Apweiler, R. (1998). The SWISS-PROT protein sequence data bank and its supplement TrEMBL in 1998. *Nucleic Acids Res.* **26**, 38-42.
- Stoesser, G., Sterk, P., Tuli, M.A., Stoehr, P.J. & Cameron, G.N. (1997). The EMBL nucleotide sequence database. *Nucleic Acids Res.* **7**, 1-14.
- Sánchez, R. & Sali, A. (1997). Advances in comparative protein-structure modelling. *Curr. Opin. Struct. Biol.* **7**, 206-214.
- Rost, B., Casadio, R. & Fariselli, P. (1996). Topology prediction for helical transmembrane proteins at 86% accuracy. *Protein Sci.* **5**, 1704-1718.
- Rost, B. (1997). Protein structures sustain evolutionary drift. *Fold. Des.* **2**, S19-S24.
- Fleischmann, R.D., et al., & Venter, J.C. (1995). Whole-genome random sequencing and assembly of *Haemophilus influenzae* rd. *Science* **269**, 496-512.
- Fraser, C.M., et al., & Venter, J.C. (1995). The minimal gene complement of *Mycoplasma genitalium*. *Science* **270**, 397-403.
- Goffeau, A., et al., & Oliver, S.G. (1996). Life with 6000 genes. *Science* **274**, 546-567.
- Bult, C.J., et al., & Geoghagen, N.S.M. (1996). Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii*. *Science* **273**, 1058-1073.
- Kaneko, T., et al., & Tabata, S. (1996). Sequence analysis of the genome of the unicellular cyanobacterium *Synechocystis* sp. Strain PCC6803. II. Sequence determination of the entire genome and assignment of potential protein-coding regions. *DNA Res.* **3**, 109-136.
- Himmelreich, R., Hilbert, H., Plagens, H., Pirkl, E., Li, B.C. & Herrmann, R. (1996). Complete sequence analysis of the genome of the bacterium *Mycoplasma pneumoniae*. *Nucleic Acids Res.* **24**, 4420-4449.
- Blattner, F.R., et al., & Shao, Y. (1997). The complete genome sequence of *Escherichia coli* K-12. *Science* **277**, 1453-1474.
- Smith, D.R., et al., & Reeve, J.N. (1997). Complete genome sequence of *Methanobacterium thermoautotrophicum* delta H: functional analysis and comparative genomics. *J. Bacteriol.* **179**, 7135-7155.
- Klenk, H.-P., et al., & Venter, C. (1997). The complete genome sequence of the hyperthermophilic, sulphate-reducing archaeon *Archaeoglobus fulgidus*. *Nature* **390**, 364-370.
- Tomb, J.-F., et al., & Venter, J.C. (1997). The complete genome sequence of the gastric pathogen *Helicobacter pylori*. *Nature* **388**, 539-547.
- Fraser, C.M., et al., & Venter, J.C. (1997). Genomic sequence of a Lyme disease spirochaete, *Borrelia burgdorferi*. *Nature* **390**, 580-586.
- Kunst, F., et al., & Danchin, A. (1997). The complete genome sequence of the Gram-positive bacterium *Bacillus subtilis*. *Nature* **390**, 249-256.
- Deckert, G., et al., & Swanson, R. (1998). The *Aquifex aeolicus* genome. *Nature*, in press.
- Kawarabayashi, Y., et al., & Kikuchi, H. (1998). Complete genome sequences of *Pyrococcus horikoshii*. *DNA Res.* **5**, in press.