Marrying structure and genomics

Burkhard Rost

European Molecular Biology Laboratory

69 012 Heidelberg, Germany; rost@embl-heidelberg.de; http://www.embl-heidelberg.de/~rost/

Introduction

Today. Large-scale genome sequencing is filling up the catalogue of natural proteins at a breath-taking speed. Today, we have available not just a large number of sequences, but also glimpses of the inventory of entire organisms. This will soon improve our understanding of cells, in particular, and of life, in general. Three means will contribute: (1) sequencing genomes (genomics), (2) determining protein structures, and (3) determining protein function. Protein structure is interwoven with function (see *Structure*, in general, [1, 2, 3] , in particular). Sequencing and determining function are also routinely combined (e.g. [4]). However, what about the relation between structure determination and genomics?

Tomorrow. Structural genomics, the marriage between protein structure determination and genomics, is already beginning. Here, I attempted to illustrate the likely direction this marriage will take. Structure determination will be pushed by, and profit from genomics. Basing research and technical developments (such as drug design) on all three pillars (sequence, structure, function) will be a big step toward understanding of life.

Objectives. Structure determination will benefit from genomics in two ways (Fig. 1). (1) The mass of available sequences will facilitate quick determination of structure for most existing folds. (2) Sequences for entire organisms will help to unravel missing links in functional pathways, to explore alternative pathways, and to widen our understanding of principle mechanisms and of evolutionary cross-links.

Genomics and structures: two flourishing fields

Tons of bricks ... The first entirely sequenced organism was published in 1995. Two years later, another ten have been published (Tab. 1). Nucleotide databases have increased two times more over the last two years, than in the twenty years before (Fig. 2). The growth now outpaces even the development of computers (Fig. 2). This is merely the beginning.
... hit protein structure determination. Structure determination has become almost routine [5] . Currently, as many structures are determined every ten days, as in the first ten years of crystallography (Fig. 2). Hitting on a novel fold, still resembles unearthing a nugget [1, 2] . Each novel fold can contribute to understanding functional details of entire protein families. How does the rate of new structures compare to that of sequences? We have structural knowledge for every tenth protein in the first genomes (Fig. 3). Three projects have started solving structures systematically for organisms (http://www.mcs.anl.gov/home/gaasterl/sg-review.html).: *Haemophilus influenzae* (John Moult, CARB, Washington in collaboration with TIGR), *Pyrobaculum aerophilum* (Tom Terwillinger, LANL, Livermore; David Eisenberg, and Jeff Miller, both UCLA, Los Angeles), and *Methanococcus jannaschii* (Sung-Hou Kim, LBNL, Berkeley). What are the objectives of structural genomics?

Mass of sequences: populate each island in structure space

Phase 1: fill blank spots in structure space. Already we know about 500 [1] of the estimated 1000 folds [6, 7] .

Thus, only about half the blank spots in structure space remain (Fig. 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: (1) find proteins common to different organisms, (2) exclude those with structural homologues (Å 10%, Fig. 3), (3) exclude integral membrane proteins (Å 20-30% Fig. 4), and (4) exclude all 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 (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., in thermophilic bacteria 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 PDB constitute novel folds [1] . A large-scale structure determination enterprise could easily yield 2,000 (additional) new structures annually. Thus, we shall have one structure per fold in less than a decade (assuming initially 10% yield of novel folds, then exponentially decaying).

Phase 2: adding details to the map. Most pairs of similar structures have < 15% pairwise sequence identity (Fig. 5). Thus, filling all blank spots does not yield all families populating the respective island (Fig. 1). The enormous sequence variation within islands is often associated with functional divergence (or convergence). To benefit from structure determination toward understanding 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, 1145 sequence unique proteins (set used for Fig. 5 ; [10]) cover about 10% of known genomes (Fig. 3). Thus, we need about 10,000 structures (covering one per family). However, phase 2 yields 100% coverage in a large-scale structure determination project (the recipe described above selects candidates representing a single sequence family). Thus, assuming a moderate production of 2,000 structures annually, we should have about a two-fold coverage within a decade.

Entirety of organisms: cover all functional elements

Phase 1: finding missing links in pathways. Knowing all sequences for entire organisms, we can start mapping them to pathways (metabolic, regulatory, signalling, pathogenic), or particular mechanisms (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 is the second objective for structural genomics: find functionally missing links (Fig. 1B). The first phase of this objective will imply to determine missing structures for all major pathways and mechanisms. Step 1 is straightforward: complete structural knowledge for all pathways and mechanisms for which we know the associated proteins. However, step 2 appears hopeless: how can we determine structures for unknown proteins? Firstly, many of the candidates selected to find all blank spots will turn out to be representatives of most major functional protein classes. Secondly, 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 in FHIT and PKCI, and the implications of the structural similarity to GalT [3]).

Phase 2: filling function space. After knowing structures for all major functional elements, we shall have to complete the functional map, i.e., to determine structures for representatives of all pathways and mechanisms. Candidates for structures to determine will be found by structure-based comparative genome analysis, focusing on particular sites (active, binding), 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 bio-chemical 'fingerprints' (structures 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 alignments and structure analysis [14] .

Conclusions

Profiting from mass and entirety. The major objectives of structural genomics I have portrayed here are: (1) to find all natural structures, and (2) to find missing links in all functional pathways and mechanisms (Fig. 1). They correspond to the two aspects of genome sequencing: (i) the mass of sequences produced, and (ii) the entirety of sequencing complete genomes from organisms. We need a large-scale structure determination enterprise.

What will come out? A prerequisite for understanding function is to know structure. Furthermore, large-scale structure determination will enable uncovering most major functional elements. The scaffolds of structures provide the elements for evolution. Most functional motifs we know today are sequence motifs. However, most functional motifs remain hidden without knowing structure. Structural genomics will help to better understand evolution, and by that to carve our understanding into techniques, such as drug discovery, and design. 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? I have 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 studying 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, we already would have knowledge about structure for more than 10,000 of these by combining structure determination and prediction (Fig. 3). However, the human genome won't be completed before 2004. With 2,000 new structures determined by a large-scale enterprise, we shall have structural knowledge for about 70% of all human sequences by 2004; many of the remaining 30,000 will be membrane proteins (Fig. 4).

Reality or dream? The message portrayed is: the mass of sequences produced by genomics should enable most natural folds to be determined within less than a decade. Wishful thinking? Firstly, the - strongly disputed - assumption that there are only 1,000 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 10-15,000 families (1,200 of today, corresponding to 8-18% of all families: Fig. 3) is rather conservative, and to determine one structure for each family is just a matter of a large-scale enterprise. Secondly, 2,000 structures were added to PDB in 1997, and structure determination techniques continue to improve. Thus, the assumption of 2,000 new structures annually is a rather conservative estimate. What remains is the uncertainty about 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, this is no easy answer, yet: we just have to try.

Acknowledgements

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

Figures

Fig. 1

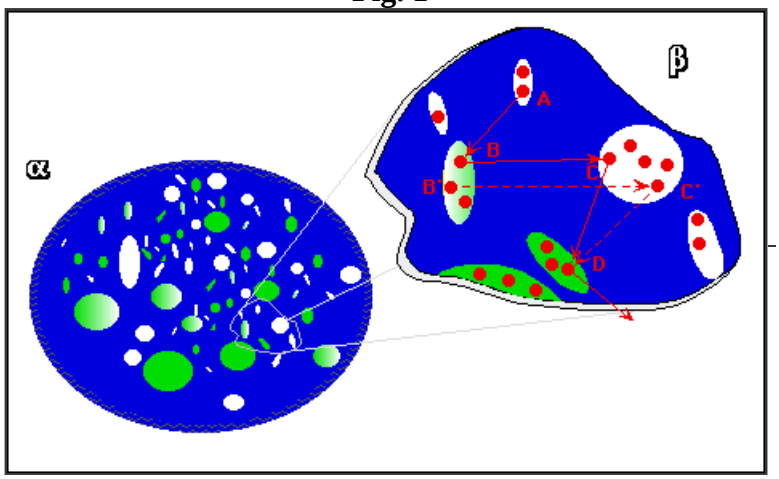

Fig. 1. Objectives for structure determination in the era of genomics.

(α) Profiting from the mass of sequences to fill in the white spots, i.e., to determine all natural folds. The ellipsoid symbolises the universe of protein structures; the islands are actually existing structures (or folds), some are larger as some folds occur more often (e.g.: TIM barrel, Immunoglobulin-like, NTP hydrolases, Ferredoxin-like, Rossmann fold, Globin-like, Flavodoxin-like, like Ribonuclease H [15]). The colour coding distinguishes three situations: (1) most structures are known (green), (2) some structures, i.e., the principle folds, are known (half green/half white), (3) no structure is known.

(β) Profiting from the entirety of sequences for complete organisms to fill in missing links in pathways, and mechanisms. The red circles indicate sequence families (sequences within this family have significant levels of pairwise sequence identity), the solid arrows symbolise a well-known pathway in organism X (A -> B -> C -> D), and the dashed arrows symbolise the analogous pathway in organism Y (A -> B -> C’ -> D). If proteins B and C do not exist in organism Y, we cannot map this pathway from knowing all sequences of Y. Imagine we know the structure of B, then threading (fold recognition) may enable to induce that B' adopts the rôle of B. However, without knowing the structures of C or C' we still couldn't guess the interaction partner of B', and thus still couldn't map the pathway. Knowing structures for all families (all red circles) we could easily find the pathway in Y.

Fig. 2

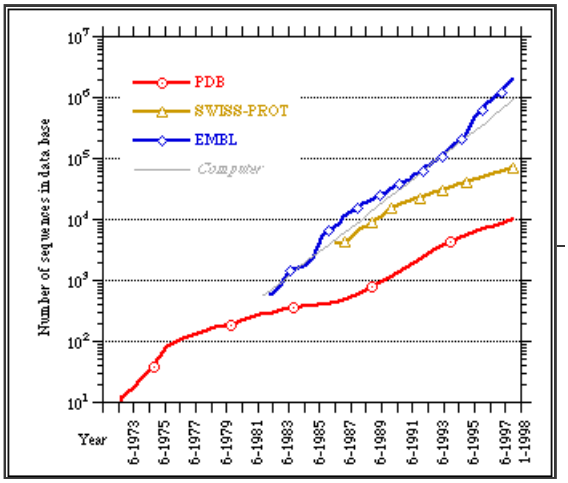

Fig. 2. Rate of growth in datases of bio-molecules.

The explosion of bio-molecular data is illustrated by three representatives: the protein structure database PDB [16] (red line, dashed circles), the protein sequence database SWISS-PROT [17] (brown line, open triangles), and the nucleotide sequence database EMBL [18] (blue line, open 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, assumptions: speed doubles every 18 months).

Fig. 3

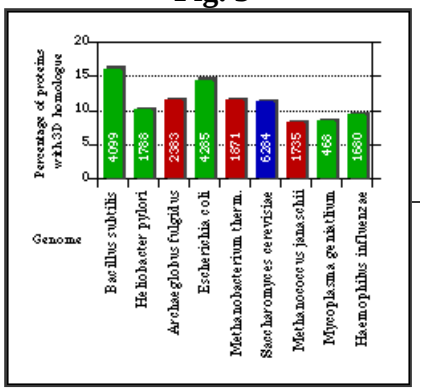

Fig. 3. Percentage of proteins in genomes with homologues of known structure.

For several whole genomes, the figure shows the number of proteins for which a sequence homologue of known structures exists in PDB as percentage of all identified proteins (the total numbers of proteins used are given in the bars; eukaryote: blue, prokaryotes: green, archer: red). Note that, on average, for about 80% of the proteins considered here, the structure is inferred by homology modelling [19] . Note furthermore that the estimates are conservative in that conservative thresholds have been applied, homology inferred by threading methods (e.g. practised for mycoplasma by [8]) were not considered.

Fig. 4

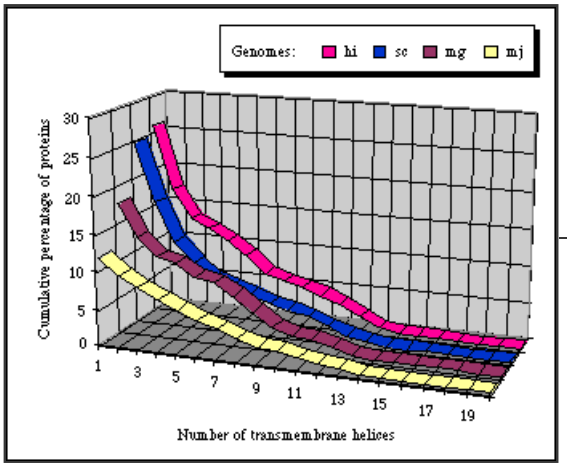

Fig. 4. Percentage of helical transmembrane proteins in organisms.

For the first four entirely sequenced organisms the percentage of proteins from the entire genome predicted by PHDtopology [20] to have helical membrane regions is shown. Colour coding: mj: Methanococcus jannaschii, yellow, first; mg: Mycoplasma genitalium, brown, second; sc: Saccaromyces cerevisiae, blue, third; hi: Haemophilus influenzae, red, fourth. The horizontal axis gives the number of transmembrane helices predicted, the vertical axis the cumulative percentage of proteins in the genome. For example, about 25% of Haemophilus and yeast are predicted to have, at least, one transmembrane helix; 7% of Haemophilus, and 5% in yeast are predicted to contain seven or more transmembrane helices (list of the proteins at: http://www.embl-heidelberg.de/~rost).

Fig. 5

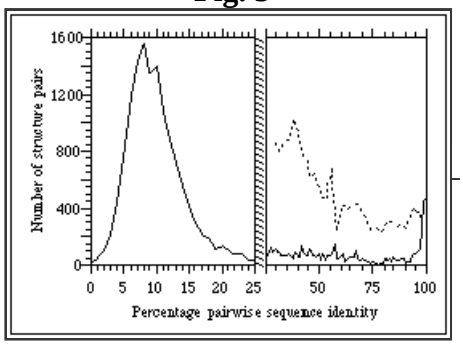

Fig. 5. Evolution of protein structures into the midnight zone of sequence identity.

Proteins evolved into the midnight zone of sequence identity, i.e., into a region where sequence comparisons fail completely to detect structural similarity [21] . Given is the distribution of pairwise sequence identity for structurally aligned protein pairs (full line). The average pairwise sequence identity of all remotely structural similar pairs (< 25% sequence identity, left panel) is below 10% sequence identity. To reduce database bias results displayed on the left panel based on a smaller data set (aligning 1145 sequence-unique structures against themselves) than those displayed on the right panel (set: aligning 1145 against the entire PDB). Consequently, numbers on 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; set: aligning 1145 sequence unique structures against the entire SWISS-PROT).

Fig. 1 Supplement

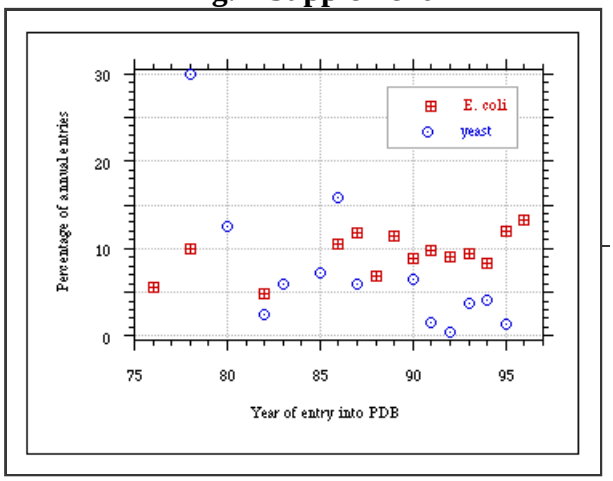

Fig. 1 Supplement. Fraction of particular organisms in annually deposited protein structures.

Do we find traces of sequencing entire organisms in the protein structure database PDB? I searched for all structures from proteins of the first entirely sequenced organisms (given are numbers for yeast and E. coli, only). However, so far the fact that we know sequences for entire genomes has not been carved into PDB. (This observation supposedly reflects the differences in complexity: determining all sequences of say *helicobacter* may be faster than determining one structure.)

Table

Table 1 Completely sequenced genomes. *

Genome	Date	Solved by	Quote
Haemophilus influenzae	8/95	TIGR	[22]
Mycoplasma genitalium	10/95	TIGR	[23]
Saccaromyces cerevisiae	1/96	Europe	[24]
Methanococcus jannaschii	8/96	TIGR	[25]
Synechocystis sp. PCC6803	9/96	Japan	[26]
Mycoplasma pneumoniae	11/96	Germany	[27]
Escherichia coli	1/97	Univ. Wisc/Japan	[28]
Methanobacterium thermo.	5/97	GTC	[29]
Archaeoglobus fulgidus	6/97	TIGR	[30]
Helicobacter pylori	6/97	TIGR	[31]
Borrelia burgdorferi	7/97	TIGR	[32]
Treponema pallidum	10/97	TIGR/Univ. Texas	**
Bacillus subtilis	11/97	Europe	[33]
Pyrococcus horikoshii	1/98	Japan	***
Aquifex aeolicus	2/98	RBI	[34]

* List from: Terry Gaasterland, http://www.mcs.anl.gov/home/gaasterl/genomes.html.
** Sequences publicly available, manuscript in preparation (C. M. Fraser, G. M. Weinstock, S. J. Norris et al.).
*** Sequences partially publicly available, manuscript in preparation

---

References

- 1. Murzin, A. G. (1996). Structural classification of proteins: new superfamilies. Curr. Opin. Str. Biol., 6, 386-394.
- 2. Holm, L. & Sander, C. (1997). New structure - novel fold? Structure, 5, 165-171.
- 3. 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.
- 4. Warbrick, E. (1997). Two's company, three's a crowd: the yeast two hybrid system for mapping molecular interactions. Structure, 5, 13-17.
- 5. Lattman, E. E. (1994). Protein crystallography for all. Proteins, 18, 103-106.
- 6. 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.
- 7. Chothia, C. (1992). One thousand protein families for the molecular biologist. Nature, 357, 543-544.
- 8. Fischer, D. & Eisenberg, D. (1997). Assigning folds to the proteins encoded by the genome of Mycoplasma genitalium. Proc. Natl. Acad. Sc. U.S.A., 94, 11929-11934.
- 9. Rost, B. & O'Donoghue, S. I. (1997). Sisyphus and prediction of protein structure. CABIOS, 13, 345-356.
- 10. Holm, L. & Sander, C. (1998). Touring protein fold space with Dali/FSSP. Nucl. Acids Res., 26, 318-321.
- 11. 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.

- 12. Brannigan, J. A., Dodson, G., Duggleby, H. J., Moody, P. C. E., Smith, J. L. et al. (1995). A protein catalytic framework with an N-terminal nucleophile is capable of self-activation. Nature, 378, 416-419.
- 13. Blomberg, N. & Nilges, M. (1997). Functional diversity of PH domains: an exhaustive modelling study. Folding & Design, 2, 343-355.
- 14. 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.
- 15. Gerstein, M. & Levitt, M. (1997). A structural census of the current population of protein sequences. Proc. Natl. Acad. Sc. U.S.A., 94, 11911-11916.
- 16. Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F., Brice, M. D. et al. (1977). The Protein Data Bank: a computer based archival file for macromolecular structures. J. Mol. Biol., 112, 535-542.
- 17. Bairoch, A. & Apweiler, R. (1998). The SWISS-PROT protein sequence data bank and its supplement TrEMBL in 1998. Nucl. Acids Res., 26, 38-42.
- 18. Stoesser, G., Sterk, P., Tuli, M. A., Stoehr, P. J. & Cameron, G. N. (1997). The EMBL nucleotide sequence database. Nucl. Acids Res., 7-14.
- 19. Sánchez, R. & Sali, A. (1997). Advances in comparative protein-structure modelling. Curr. Opin. Str. Biol., 7, 206-214.
- 20. Rost, B., Casadio, R. & Fariselli, P. (1996). Topology prediction for helical transmembrane proteins at 86% accuracy. Prot. Sci., 5, 1704-1718.
- 21. Rost, B. (1997). Protein structures sustain evolutionary drift. Folding & Design, 2, S19-S24.
- 22. Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F. et al. (1995). Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. Science, 269, 496-512.
- 23. Fraser, C. M., Gocayne, J. D., White, O., Adams, M. D., Clayton, R. A. et al. (1995). The minimal gene complement of Mycoplasma genitalium. Science, 270, 397-403.
- 24. Goffeau, A., Barrell, B. G., Bussey, H., Davis, R. W., Dujon, B. et al. (1996). Life with 6000 genes. Science, 274, 546-567.
- 25. Bult, C. J., White, O. W., Olsen, G. J., Zhou, L. Z., Fleischmann, R. D. et al. (1996). Complete genome sequence of the Methanogenic Archaeon, Methanococcus jannaschii. Science, 273, 1058-1073.
- 26. Kaneko, T., Sato, S., Kotani, H., Tanaka, A., Asamizu, E. et al. (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.
- 27. Himmelreich, R., Hilbert, H., Plagens, H., Pirkl, E., Li, B. C. et al. (1996). Complete sequence analysis of the genome of the bacterium Mycoplasma pneumoniae. Nucl. Acids Res., 24, 4420-4449.
- 28. Blattner, F. R., Plunkett, G., Bloch, C. A., Perna, N. T., Burland, V. et al. (1997). The complete genome sequence of Escherichia coli K-12. Science, 277, 1453-1474.
- 29. Smith, D. R., Doucette-Stamm, L. A., Deloughery, C., Lee, H., Dubois, J. et al. (1997). Complete genome sequence of Methanobacterium thermoautotrophicum delta H: functional analysis and comparative genomics. J. Bacteriol., 179, 7135-7155.
- 30. Klenk, H.-P., Clayton, R. A., Tomb, J.-F., White, O., Nelson, K. E. et al. (1997). The complete genome sequence of the hyperthermopholic, sulphate-reducing archaeon Archaeoglobus fulgidus. Nature, 390, 364-370.
- 31. Tomb, J.-F., White, O., Kerlavage, A. r., Clayton, R. A., Sutton, G. G. et al. (1997). The complete genome sequence of the gastric pathogen Helicobacter pylori. Nature, 388, 539-547.
- 32. Fraser, C. M., Casjens, S., Huang, W. M., Sutton, G. G., Clayton, R. et al. (1997). Genomic sequence of a Lyme disease spirochaete, Borrelia burgdorferi. Nature, 390, 580-586.
- 33. Kunst, F., Ogasawara, N., Moszer, I., Albertini, A. M., Alloni, G. et al. (1997). The complete genome sequence of the Gram-positive bacterium Bacillus subtilis. Nature, 390, 249-256.
- 34. Deckert, G., Warren, P., Gaasterland, T., Young, W. G., Lenox, A. L. et al. (1998). The Aquifex aeolicus genome. Nature, in press.


HOME
Predict
Protein
AQUA