From protein sequence to function

Antoine Danchin

As genome sequences and protein structures are deciphered, we wish to predict their corresponding functions. Many functions cannot be told from the sequence, however, although there has been progress in this quest for an impossible Grail. Furthermore, a structure and its corresponding sequence become most interesting when one knows the function. Inductive reasoning, based on the integration of biological and sequence knowledge, should enable sequence and functional data to be combined in a productive way.

Addresses
Régulation de l'Expression Génétique, Institut Pasteur, 28 rue du Docteur Roux, 75724 Paris Cedex 15, France;
e-mail: adanchin@pasteur.fr

Current Opinion in Structural Biology 1999, **9**:363–367

http://biomednet.com/elecref/0959440X00900363

© Elsevier Science Ltd ISSN 0959-440X

Abbreviations

3D	three-dimensional
ABC	ATP-binding cassette
DHQS	dehydroquinate synthase
GlnRS	glutaminyl-tRNA synthetase
PLA2	phospholipase A2

Introduction

This article's title reads like the title of grand research programs, which use huge facilities, a large amount of money and a large number of scientists as well. Is this a reasonable way to view the relationships among sequences, structures and functions in living organisms? The past year has revealed a dozen new genomes [1–10,11••] and a very large number of new structures [12], determined using either what is nowadays the almost standard approach of the X-ray diffraction of crystals (using synchrotron radiation) or a technique that, for a long time, developed steadily, NMR spectroscopy (using all kinds of relaxation of nuclear spins [13,14,15•]), as well as using the still emerging technique of structural electron microscopy [16]. Can we say that this led to the discovery of unexpected functions or, at least, to rules governing the creation of biological functions [17•]? Curiously enough, function prediction from genomes is still mostly performed using approaches that combine Blast or Fasta searches, sometimes with elaborate knowledge-organizing engines, without much investigation into the reality of the underlying biology. Therefore, rather than dwell on automatic or semi-automatic genome annotation software [18–20], I shall review here what I think are crucial pieces of information that show when and why structures and functions are related. I shall further hint that the best way to use genomic text and structural biology is to go from the function to the structure and, sometimes, perhaps, to the sequence.

As a matter of fact, the dialog between structure and function is often reminiscent of the dialog between Lamarckian and Darwinian thinking. Reality is not that heavenly, however, the material objects that evolve to make life are continuously submitted to the Darwinian triplet, variation–selection–amplification, and this is what gives meaning to them. In the course of evolution, interactions between objects create new relationships and this creates functions. In turn, functions capture pre-existing structures and both act and evolve together.

Before reviewing some of the structural data from the past year that have helped us to better understand relationships between gene sequences and biological functions, perhaps we should first try to make this concept clearer. A function is an organized process of actions that cooperate towards a common goal. One has to add a goal to a structure in order to understand its function. Of course, this cannot be a straightforward feature of a structure, even less a feature of a sequence. In addition, one should consider that a function is almost never an isolated process, in that it is often split into a basal function, which is associated with ancillary functions that converge to its actualization. Understanding some ancillary functions is not enough to understand the basal function associated with a structure. As an example, in cytidylate kinase [21], the function of residues Tyr40, Arg110 and Asp132 is to bind specifically to cytosine. This is an ancillary function to making CTP. We also notice, however, that a long insert containing a three-stranded β sheet, which is not present in other nucleotide kinases but is specific to cytidylate kinases, has no recognized function in the enzyme. Remarkably, even ancillary functions are not always revealed by the sequence; conserved residues sometimes fall into different function types [22].

Some structures that tell their function

When known, the goal of a function can help us to connect it to the structure (and sometimes to the sequence, when phylogenetic comparisons with known objects are available). In many cases, the functions of membrane proteins can be understood from the fact that membranes connect the outside to the inside of a cell.

The most fascinating set of recent structural data is that of the minuscule engine ATP synthase. This enzyme manufactures ATP from ADP and phosphate using a vectorial protonmotive force. From combinations of sequence, biochemical and structural data, it has been suspected for some time that the enzyme comprises a rotor and a stator. A single molecule of F_1-ATPase, the engine portion of ATP synthase, is a rotary device in which a central γ subunit rotates against a surrounding cylinder composed of $\alpha_3\beta_3$ subunits. Driven by ATP hydrolysis in three catalytic β subunits, the γ subunit makes discrete 120° rotations, occasionally moving

backwards. Rotation was confirmed after the isolation of F_1 and its direct visualization by optical fluorescent microscopy of the actual rotation of the central part of the enzyme with respect to the rest of the molecule [23]. The work carried out at each step is constant over a broad range of load. It is close to the free energy of hydrolysis, transmitted by elastic strain, of one ATP molecule [24••,25].

It was, of course, necessary to understand the structural arrangement of the residues that permit the generation of the torque that drives the rotor movement of the enzyme [26]. The corresponding NMR and electron microscopy data have been reviewed recently [27]. Is this all that can be said about ATP synthase? It has just been discovered that, in yeast, the native enzyme is dimerized and that extra subunits are needed for dimerization. This tells us something about the enzyme's function, be it only because the function must be related to the architecture of the cell [28]. These subunits were not discovered by analyzing the yeast genome. One must therefore draw a first important conclusion — is it possible to find out, just by knowing the genomic text, whether a gene product will form a protein complex? This is, of course, even more unlikely than that an amino acid sequence could tell us exactly the fold of a protein without knowing pre-existing folds. Pancreatic RNase should have been known to fold indeed, because selection isolated it with this behavior (it is secreted in bile salts), but this should never have been accepted, as it was, as the paradigm for protein folding. This is still the case, however, and water is now taken into consideration [29]. In the same way, some experimental data are taken into account in order to solve the unsolvable (e.g. see [30]). Another very promising approach, often called 'threading' [31,32], permits one to predict a fold by threading a sequence into a putative predefined fold (e.g. created using multi-alignments [33,34•]). Of course, this implicitly takes into account the selective forces that have, during phylogeny, led to the actual fold found in 3D protein models, as is taken into account in more recent genome annotation systems [35••,36••,37•].

The membrane P-type ATPases, a family that includes the plasma membrane Na^+/K^+-ATPase and the sarcoplasmic reticulum Ca^{2+}-ATPase, are functionally related to ATP synthases. Electron cryocrystallography of two-dimensional surface crystals of both the *Neurospora* H^+-ATPase and the sarcoplasmic reticulum Ca^{2+}-ATPase yielded structure maps for these ion transporters at a resolution of 8 Å. Apart from fitting with the sequence prediction of the membrane helices, this was not enough to visualize the organization of the metal-binding channel [38]. Recently, however, the reality of the metal-binding pores has been explored in depth. The expected pore type was found in the potassium transporter — a hole in a membrane protein. The width of the hole and the bordering residues — identified from multisequence alignments — fit well with the diameter of the potassium ion and explain the gating properties [39]. Many receptor structures fall into this category, although the actual binding sites for the substrates could rarely be predicted from the sequence alone [40–43]. In line with this structure/function identification being well connected to sequence data are the results of multi-alignment algorithms, combined with phylogenetic and biochemical data (e.g. see [43,44•]).

A more involved type of permease is the FhuA iron transport channel in *Escherichia coli*. This protein, which mediates the transport of iron chelated to a siderophore through the outer membrane, exhibits a remarkable structure. The 22 antiparallel β strands of the protein look like the neck of a bottle, closed by a mobile cork. Analysis of the protein complex with its substrate reveals how an allosteric transition, mediated by the energy-transducing TonB protein, permits the iron chelate to translocate into the cell [45••].

The sequence of the *Bacillus subtilis* genome revealed 77 genes encoding putative ATP-binding cassette (ABC) permeases [46]. From eukaryotes to archebacteria, these proteins perform a variety of transport functions (most of which are unknown). They comprise two ABC domains located in the cytoplasm of the cell. These domains are combined with a couple of transmembrane integral proteins. ATP hydrolysis provides the driving force to translocate the transported molecules (to the inside or outside of the cell, depending on the permease type). The structure of the ATP-binding domain is now known [47•]. In the permease, two such ABC domains are associated and cooperate in order to hydrolyze ATP, which is bound at a site predicted from the sequence, presumably by coupling hydrolysis to some conformational change in the integral membrane proteins in the permease. In spite of knowledge of sequences, as well as of the crystal structure, it is not yet possible to predict how this conformational change is performed. Knowing the variety of existing ABC permeases and their importance as multidrug efflux systems [48], it can be surmised that many different systems could be energized by ATP hydrolysis. A common mechanism might be that the energy of hydrolysis exposes a hydrophobic residue at the surface of the ABC domains, thereby inducing a change in the water molecules bound to the surface of the protein, as well as in the proteins connected to the surface. An entropy-driven process would fold the proteins back into their original conformations, as ADP and phosphate are released into the medium. This example illustrates how evolution may create a function by playing variations upon the theme of available structures once a success has been registered.

Some structures that tell something about evolution

Evolution proceeds by the recruitment of pre-existing genes [49]. As a case in point, crystallins are ordinary enzymes that have been recruited for entirely different functions [50•]. This is a strong reason for the lack of intuitive correlation among sequence, structure and function.

A recent example illustrating this general behavior are snake venoms that include phospholipase A2 (PLA2). A number of venoms have been shown to contain a catalytically inactive PLA2 homolog, in which the highly conserved aspartic acid at position 49 is substituted by lysine. These PLA2s disrupt membranes through a Ca^{2+}-independent mechanism of action; however, the structural bases underlying these functional properties do not seem to be related to phospholipase activity [51].

In an entirely different context (in relation to convergent evolution), the crystal structure of dehydroquinate synthase (DHQS), which performs the second step in the shikimate pathway, has revealed entirely unexpected new folds. These folds explain how the enzyme performs several consecutive chemical reactions in one active site [52]. DHQS exhibits a previously unobserved mode of NAD^+ binding and an active site organization that is surprisingly similar to that of alcohol dehydrogenase in a new protein fold. The structure reveals interactions between the active site and a substrate analog inhibitor that suggest an explanation for how DHQS performs multistep catalysis without the formation of unwanted by-products.

Protein–nucleic acid interaction is a much studied area; however, new protein structures pertaining to nucleic-acid-binding domains are continuously discovered. In fact, it is likely that, especially in bacteria, the importance of RNA–protein complexes has been entirely overlooked. RNA-binding protein folds were certainly present very early on in the origin of life and they must have been present at the origin of translation [49]. In this respect, it seems most interesting that ribosomal protein L25 is homologous both to a domain of glutaminyl-tRNA synthetase (GlnRS) and to general stress proteins [53••]. Indeed, GlnRS is an enzyme that is absent in many bacteria, such as Gram positives and archebacteria, which are considered as descending from an important universal ancestor of living organisms [54••]. In the same way, it seems most probable that several stress proteins are ancestral because their function in protein folding must have been present very early. The RNA-binding fold of cold-shock proteins [55] has, for example, been found in the eukaryotic putative translation initiation factor 5A [56•,57]. It can be therefore expected that other RNA-binding protein folds will be discovered [58] and that the progeny of these proteins will have been the source of recruitment for many other types of activities. In this respect, we can expect that, among the 'unknown' genes in genomes ('y' genes), many important genes will be associated with RNA–protein complexes.

Conclusions

At the beginning of 1999, a search in Medline for 'crystal' and 'function' yielded 11,415 references. If one restricted the search to 'cyclase', 43 references were still present (among 18,792 references to 'adenylyl cyclase'). Every new issue of *Science* or *Nature* has its share of crystal structures. It is more than a fashion, it is an explosion of new data. Of course, a structure can only be determined when the corresponding polypeptide sequence is also known; genome sequences match this explosion. What can we tell from the 3D structure of a protein? The long-awaited structure of the catalytic domain of class III [59] adenylyl cyclases was published in 1997. It revealed a set of interesting features, including the binding site for the activator forskolin. In the absence of direct data, however, the catalytic site could not be immediately visualized [60]. Another structure was needed to permit the identification of the catalytic site. This work also led to the proposal of a mechanism for stimulatory G protein α subunit ($G_{s\alpha}$) activation of the enzyme [61] and further work identified the role of metal ions in the enzyme [62].

This illustrates the need for the integration of biological and biochemical knowledge with sequence data in order to go from a sequence to its function. One must subsequently proceed using inductive reasoning. One way to do this efficiently is to organize the data according to their 'proximity'. This allows one to explore the gene sequence data, which are related to each other by a variety of themes, including proximity in the chromosome, phylogenetic kinship, participation in a common metabolic pathway, their common presence in an article of literature or a similar use of the genetic code [63•], and therefore to make educated guesses about the corresponding function(s).

As biological knowledge is integrated with biochemical and biophysical knowledge, it will become more and more efficient to predict a structure from its polypeptide sequence. Despite continuous progress, the function will remain difficult to predict and will still have to be proved experimentally. This approach will not only allow better exploitation of genome sequence data, but will also permit one to create entirely new proteins *de novo* [64••].

Acknowledgements

Progress in understanding of the concept of function benefitted from the discussion of the group of the 'Causeries du Jeudi soir'. Agnieszka Sekowska pointed out to me the importance of polyamine biosynthesis in understanding RNA–protein interactions. The author's research on genomes was supported by European Union BIOTECH programs and by the Groupement de Recherche et d'Etudes des Génomes.

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. The EMBL nucleotide sequence database on the World Wide Web at URL: http://www.ebi.ac.uk_docs/embl_db/ebi/topembl.html
2. DNA Data Bank of Japan on the World Wide Web at URL: http://www.ddbj.nig.ac.jp/
3. Entrez genomes on the World Wide Web at URL: http://www.ncbi.nlm.nih.gov/Entrez/Genome/org.html
4. GenomeNet WWW server on the World Wide Web at URL: http://www.genome.ad.jp/

5. The Sanger Centre Projects on the World Wide Web at URL: http://www.sanger.ac.uk/Projects/

6. The Institute for Genomic Research on the World Wide Web at URL: http://www.tigr.org/

7. The University of Oklahoma's Advanced Center for Genome Technology on the World Wide Web at URL: http://www.genome.ou.edu/

8. Completed and ongoing genome projects monitor on the World Wide Web at URL: http://geta.life.uiuc.edu/-nikos/genomes.html

9. Genomic databases on the Internet on the World Wide Web at URL: http://bmbsgi12.leeds.ac.uk/bmbknd/DNA/genomic.html

10. Annotated links to Biochemistry and Molecular Biology Internet resources on the World Wide Web at URL: http://www.sgi.bls.umkc.edu/biolinks/

11. ExPASy Molecular Biology Server on the World Wide Web at URL:
•• http://expasy.hcuge.ch/www/expasy-top.html

The SwissProt data library is the best annotated protein sequence library that is used to predict protein functions. Please note that it is not infallible!

12. The Brookhaven National Laboratory Protein Data Bank on the World Wide Web at URL: http://www.pdb.bnl.gov/pdb-docs/index.html

13. Kay LE: **Protein dynamics from NMR.** *Nat Struct Biol* 1998, **5(suppl)**:513-517.

14. Dobson CM, Hore PJ: **Kinetic studies of protein folding using NMR spectroscopy.** *Nat Struct Biol* 1998, **5(suppl)**:504-507.

15. Gardner KH, Kay LE: **The use of 2H, 13C, 15N multidimensional**
• **NMR to study the structure and dynamics of proteins.** *Annu Rev Biophys Biomol Struct* 1998, **27**:357-406.

NMR spectroscopy techniques are developing rapidly. The protein structures that are resolved using this technique are getting bigger and bigger, thanks to the use of heteronuclear coupling, as described in this paper.

16. Chiu W, Avila-Sakar AJ, Schmid MF: **Electron crystallography of macromolecular periodic arrays on phospholipid monolayers.** *Adv Biophys* 1997, **34**:161-172.

17. 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.

A comprehensive view of the potential of *in silico* prediction of gene function, illustrated with specific examples.

18. Gaasterland T, Sensen CW: **Fully automated genome analysis that reflects user needs and preferences. A detailed introduction to the MAGPIE system architecture.** *Biochimie* 1996, **78**:302-310.

19. Harris NL: **Genotator: a workbench for sequence annotation.** *Genome Res* 1997, **7**:754-762.

20. Tamames J, Ouzounis C, Casari G, Sander C, Valencia A: **EUCLID: automatic classification of proteins in functional classes by their database annotations.** *Bioinformatics* 1998, **14**:542-543.

21. Briozo P, Golinelli-Pimpaneau B, Gilles A-M, Gaucher J-F, Burlacu-Miron S, Sakamoto H, Janin J, Barzu O: **Structures of *Escherichia coli* CMP kinase alone and in complex with CDP: a new fold of the nucleoside monophosphate binding domain and insights into cytosine nucleotide specificity.** *Structure* 1998, **6**:1517-1527.

22. Schmitt E, Moulinier L, Fujiwara S, Imanaka T, Thierry JC, Moras D: **Crystal structure of aspartyl-tRNA synthetase from *Pyrococcus kodakaraensis* KOD: archaeon specificity and catalytic mechanism of adenylate formation.** *EMBO J* 1998, **17**:5227-5237.

23. Yasuda R, Noji H, Kinosita K Jr, Motojima F, Yoshida M: **Rotation of the gamma subunit in F1-ATPase; evidence that ATP synthase is a rotary motor enzyme.** *J Bioenerg Biomembr* 1997, **29**:207-209.

24. Yasuda R, Noji H, Kinosita K Jr, Yoshida M: **F1-ATPase is a highly**
•• **efficient molecular motor that rotates with discrete 120 degree steps.** *Cell* 1998, **93**:1117-1124.

This beautiful experimental work demonstrates that ATP synthase is the smallest engine in the world.

25. Wang H, Oster G: **Energy transduction in the F1 motor of ATP synthase.** *Nature* 1998, **396**:279-282.

26. Kaim G, Dimroth P: **Voltage-generated torque drives the motor of the ATP synthase.** *EMBO J* 1998, **17**:5887-5895.

27. Wilkens S, Capaldi RA: **Electron microscopic evidence of two stalks linking the F1 and F0 parts of the *Escherichia coli* ATP synthase.** *Biochim Biophys Acta* 1998, **1365**:93-97.

28. Arnold I, Pfeiffer K, Neupert W, Stuart RA, Schägger H: **Yeast mitochondrial F0F1-ATP synthase exists as a dimer: identification of three dimer-specific subunits.** *EMBO J* 1998, **17**:7170-7178.

29. Duan Y, Kollman PA: **Pathways to a protein folding intermediate observed in a 1-microsecond simulation in aqueous solution.** *Science* 1998, **282**:740-744.

30. Skolnick J, Kolinski A, Ortiz AR: **Reduced protein models and their application to the protein folding problem.** *J Biomol Struct Dyn* 1998, **16**:381-396.

31. Yadgari J, Amir A, Unger R: **Genetic algorithms for protein threading.** *ISMB* 1998, **6**:193-202.

32. Parisien M, Major F, Peitsch M: **A protein conformational search space defined by secondary structure contacts.** *Pac Symp Biocomput* 1998:425-436.

33. Holm L, Sander C: **Dictionary of recurrent domains in protein structures.** *Proteins* 1998, **33**:88-96.

34. Holm L, Sander C: **Protein folds and families: sequence and**
• **structure alignments.** *Nucleic Acids Res* 1999, **27**:244-247.

This article, together with [33], provides a dictionary of folds and enables them to be connected with sequence multi-alignments.

35. Overton GC, Bailey C, Crabtree J, Gibson M, Fischer S, Schug J: **The**
•• **GAIA software framework for genome annotation.** *Pac Symp Biocomput* 1998:291-302.

The GAIA software illustrates a recent case of the general tendency to develop multipurpose engines for semi-automatic genome annotation (e.g. Imagene), in contrast to automatic annotation (e.g. Genotator, Genequiz and so on).

36. Gaasterland T, Ragan MA: **Constructing multigenome views of**
•• **whole microbial genomes.** *Microb Comp Genomics* 1998, **3**:177-192.

A further reflection after MAGPIE, which set the stage for an integrated view of genome annotation. MAGPIE was an important step forward that permitted the community of genome annotators to discover the impossibility, in the absence of entirely new concepts, of managing the huge amount of data generated by genome comparisons.

37. Corpet F, Gouzy J, Kahn D: **Recent improvements of the ProDom**
• **database of protein domain families.** *Nucleic Acids Res* 1999, **27**:263-267.

A standard, but useful, automatic construction of the multi-alignment of protein domains.

38. Kühlbrandt W, Auer M, Scarborough GA: **Structure of the P-type ATPases.** *Curr Opin Struct Biol* 1998, **8**:510-516.

39. Moczydlowski E: **Chemical basis for alkali cation selectivity in potassium-channel proteins.** *Chem Biol* 1998, **5**:R291-R301.

40. Armstrong N, Sun Y, Chen GQ, Gouaux E: **Structure of a glutamate-receptor ligand-binding core in complex with kainate.** *Nature* 1998, **395**:913-917.

41. Chang G, Spencer RH, Lee AT, Barclay MT, Rees DC: **Structure of the MscL homolog from *Mycobacterium tuberculosis*: a gated mechanosensitive ion channel.** *Science* 1998, **282**:2220-2226.

42. Himanen JP, Henkemeyer M, Nikolov DB: **Crystal structure of the ligand-binding domain of the receptor tyrosine kinase EphB2.** *Nature* 1998, **396**:486-491.

43. Thornton JW, Kelley DB: **Evolution of the androgen receptor: structure-function implications.** *Bioessays* 1998, **20**:860-869.

44. Garcia-Vallve S, Palau J: **Nuclear receptors, nuclear-receptor**
• **factors, and nuclear-receptor-like orphans form a large paralog cluster in *Homo sapiens*.** *Mol Biol Evol* 1998, **15**:665-682.

An interesting illustration of the combination of phylogenetic, biochemical and sequence data in order to understand the formation of multidomain proteins.

45. Ferguson AD, Hofmann E, Coulton JW, Diederichs K, Welte W:
•• **Siderophore-mediated iron transport: crystal structure of FhuA with bound lipopolysaccharide.** *Science* 1998, **282**:2215-2226.

A remarkable synthesis of a cork-in-a-bottleneck pore structure. FhuA is located in the outer membrane of *E. coli* and one might have expected a static structure, like that of the porins. This work illustrates how a signal from the inside of the cell, mediated by TonB through the periplasm, might induce an allosteric transition moving the cork and allowing the iron-carrying siderophore to pass through FhuA.

46. Kunst F, Ogasawara N, Moszer I, Albertini AM, Alloni G, Azevedo V, Bertero MG, Bessières P, Bolotin A, Borchert S *et al.*: **The complete genome sequence of the Gram positive model organism *Bacillus subtilis* (strain 168).** *Nature* 1997, **390**:3313-3328.

47. Hung L-W, Wang IX, Nikaido K, Liu P-Q, Ames GF-L, Kim S-H:
• **Crystal structure of the ATP-binding subunit of an ABC transporter.** *Nature* 1998, **396**:703-707.
The long-awaited structure of the ATP-binding subunit of an ABC transporter, raising many new questions.

48. Kolaczkowski M, Kolaczowska A, Luczynski J, Witek S, Goffeau A: ***In vivo* characterization of the drug resistance profile of the major ABC transporters and other components of the yeast pleiotropic drug resistance network.** *Microb Drug Resist* 1998, **4**:143-158.

49. Danchin A: **Homeotopic transformation and the origin of translation.** *Prog Biophys Mol Biol* 1989, **54**:81-86.

50. Piatigorsky J: **Multifunctional lens crystallins and corneal enzymes.**
• **More than meets the eye.** *Ann N Y Acad Sci* 1998, **842**:7-15.
An illustration of the recruitment of new functions by existing structures.

51. Ward RJ, de Azevedo WF Jr, Arni RK: **At the interface: crystal structures of phospholipases A2.** *Toxicon* 1998, **36**:1623-1633.

52. Carpenter EP, Hawkins AR, Frost JW, Brown KA: **Structure of dehydroquinate synthase reveals an active site capable of multistep catalysis.** *Nature* 1998, **394**:299-302.

53. Stoldt M, Wohnert J, Gorlach M, Brown LR: **The NMR structure of**
•• ***Escherichia coli* ribosomal protein L25 shows homology to general stress proteins and glutaminyl-tRNA synthetases.** *EMBO J* 1998, **17**:6377-6384.
Resolving the ribosome protein L25 structure illustrates the power of the new NMR techniques. It also raises major questions about the origin and fate of RNA-binding proteins. What are the similarities among sequences with a structure in common and do they display common functions?

54. Gupta RS: **Protein phylogenies and signature sequences: a**
•• **reappraisal of evolutionary relationships among archaebacteria, eubacteria, and eukaryotes.** *Microbiol Mol Biol Rev* 1998, **62**:1435-1491.
A very controversial view on the first cells found at the origin of life. Using sequence data correlated to the corresponding functions, Gupta reverts to the prokaryote/eukaryote dichotomy, proposing that Gram-positive eubacteria and archaebacteria form the original class (monoderms) and that Gram-negative eubacteria (diderms) engulfed an archaebacterial type to form the eukaryotic lineage.

55. Schindelin H, Marahiel MA, Heinemann U: **Universal nucleic acid-binding domain revealed by crystal structure of the *B. subtilis* major cold-shock protein.** *Nature* 1993, **364**:164-168.

56. Kim KK, Hung LW, Yokota H, Kim R, Kim SH: **Crystal structures of**
• **eukaryotic translation initiation factor 5A from *Methanococcus jannaschii* at 1.8 Å resolution.** *Proc Natl Acad Sci USA* 1998, **95**:10419-10424.
An important protein containing hypusine, an unusual amino acid, is described (see, also, [57]).

57. Peat TS, Newman J, Waldo GS, Berendzen J, Terwilliger TC: **Structure of translation initiation factor 5A from *Pyrobaculum aerophilum* at 1.75 Å resolution.** *Structure* 1998, **6**:1207-1214.

58. Draper DE, Reynaldo LP: **RNA binding strategies of ribosomal proteins.** *Nucleic Acids Res* 1999, **27**:381-388.

59. Barzu O, Danchin A: **Adenylyl cyclases: a heterogeneous class of ATP-utilizing enzymes.** *Prog Nucleic Acid Res Mol Biol* 1994, **49**:241-283.

60. Zhang G, Liu Y, Ruoho AE, Hurley JH: **Structure of the adenylyl cyclase catalytic core.** *Nature* 1997, **386**:247-253. [Published erratum appears in *Nature* 1997, **388**:204.]

61. Tesmer JJ, Sunahara RK, Gilman AG, Sprang SR: **Crystal structure of the catalytic domains of adenylyl cyclase in a complex with Gsα.GTPγS.** *Science* 1997, **278**:1907-1916.

62. Zimmermann G, Zhou D, Taussig R: **Mutations uncover a role for two magnesium ions in the catalytic mechanism of adenylyl cyclase.** *J Biol Chem* 1998, **273**:19650-19655.

63. Nitschké P, Guerdoux-Jamet P, Chiapello H, Faroux G, Hénaut C,
• Hénaut A, Danchin A: **Indigo: a World-Wide-Web review of genomes and gene functions.** *FEMS Microbiol Rev* 1998, **22**:207-227.
The concept of 'neighborhood' is illustrated in this review, showing how to use an inductive approach to understand, knowing its sequence, the function of a protein.

64. Harbury PB, Plecs JJ, Tidor B, Alber T, Kim PS: **High-resolution**
•• **protein design with backbone freedom.** *Science* 1998, **282**:1462-1467.
This paper looks towards the future of protein design.