Systems for categorizing functions of gene products

Monica Riley

As the sequencing of the total DNA of many organisms continues, attention is turning next to the interpretation of the function of all of the genes and gene products, with the aim of learning the full meaning of the entire genetic blueprint of a sequenced organism in concrete terms. To set the stage for accomplishing this, systems need to be constructed for the expression of the functions of gene products in systematic yet rich ways. As much as possible, the same systems should be applicable to all organisms, so that when comparability exists among organisms the connections will become clear.

Addresses
Marine Biological Laboratory, Bay Paul Center, 7 MBL Street, Woods Hole, MA 02543, USA; e-mail: mriley@mbl.edu

Current Opinion in Structural Biology 1998, **8**:388–392

http://biomednet.com/elecref/0959440X00800388

Abbreviation
EC Enzyme Commission

Introduction

The recent determinations of the full genomic sequences of several microorganisms and the promise of many more complete sequences of both unicellular and multicellular organisms in the future requires that we look ahead and consider the best ways of making biological sense of the data and relating findings from the many sequenced organisms. In the case of bacteria, we are now within reach of being able to list all of the individual genes and functions that constitute the totality of information needed to sustain the life of a free-living cell, to give it the ability to adapt to changing circumstances and to perpetuate that life through generations of progeny. In order to list all functions of all genes, the connection must be made between sequence and cellular function.

History

The organism for which we have the most complete knowledge of gene product function is *Escherichia coli*. With thousands of genes and gene products, how can we organize the data so that we can understand all the functions of a unicellular organism (and ultimately a multicellular organism)? Over time, Barbara Bachmann and, more recently Mary Berlyn, have kept a running track of the known genes of *E. coli*, listing them with information on their map positions as well as either the function of the wild-type gene product or the phenotype of a mutant lacking the wild-type gene product. Alphabetical lists of the genes and functions were published at intervals through 1996 [1,2].

Functional classification today

An attempt at organizing this kind of information in terms of cellular function was made as early as 1983 [3]. Seeing that we were approaching a time when we could have full knowledge of the functions of all *E. coli* gene products, I made a further attempt in 1993 to list genes by their cellular function, hoping then to be able to see how far we had come and how much further we still had to go [4]. This was (and is) a primitive approximation for setting out the complex cellular roles of all *E. coli* gene products. A revised and improved version was presented in 1996, as a chapter in the second edition of a two-volume treatise on *E. coli* and *Salmonella* [5•], and occasional updates to the scheme can be found in EcoCyc [6] and GenProtEC [7]. Various versions of this approach to classifying function are in current use by many genomics laboratories.

The classification of the function of *E. coli* gene products has made it possible to report (in terms of cellular function) the distributions of genetic resources in organisms whose genome has been fully sequenced. The assumption is that genes and proteins of a similar sequence in different organisms carry out the same cellular functions. Exceptions exist, but for the time being, this is a reasonable approach. Thus, similaritities between sequences of translated open reading frames relating to sequences of known *E. coli* proteins has helped to make tentative assignments of function to the gene products of newly sequenced organisms whose biology and genetics is not known in detail, for example [8,9]. In future, the multiple functions of multimodular (chimaeric) proteins [10] will also be taken into account in this kind of analysis.

Functional classification systems have also been developed for other organisms for which there is extensive biological information: the Gram positive organism *Bacillus subtilis* [11]; the eukaryotic micro-organism yeast *Saccharomyces cerevisiae* [12•,13•]; and the higher eukaryotic organism *Drosophila* [14••]. One may ask if there is a basis for connecting together the gene/function data for such widely different organisms. Are there enough commonalities to connect their genetics and biology?

Which cell functions are basically the same in different organisms and which are different? Many types of genes and functions are similar in *B. subtilis* and *E. coli*, except that *B. subtilis* has cellular functions that are not seen in *E. coli*, such as sporulation, germination and competence for transformation. There are commonalities among prokaryotes and the eukaryote *S. cerevisiae*. Metabolic reactions and pathways are largely conserved. Macromolecule synthesis and modification are similar in bacteria and yeast, but are more complex in yeast (the addition of RNA splicing in yeast, for instance). Some cellular entities that are

present in yeast are not present in bacteria, such as the cytoskeleton and the endoplasmic reticulum. Some processes seen in yeast are seen in bacteria, such as budding, intracellular transport across organelle and nuclear membranes, many of the signalling systems, and the eukaryotic cell cycle with meiosis and mitosis. Nevertheless, *E. coli* and yeast do share many types and functions of gene products. The yeast genome is only half as large again as the *E. coli* genome, and a substantial fraction is devoted to a shared process — metabolism.

As one expects, the classification of *Drosophila* functions is far more complex. Michael Ashburner at Cambridge University in England has worked out a highly detailed hierarchical organization of the cellular processes and functions of gene products from *Drosophila* [14••]. Some of these processes have no counterpart in yeast or bacteria and are unique to the more complex multicellular organism. Examples are endo- and exocytosis, pinocytosis, gametogenesis, fertilization, patterning, the developmental stages of the whole organism in embryogenesis and morphogenesis, sex determination, behavior and defense systems. Other *Drosophila* processes do have counterparts in yeast and bacteria but are more elaborate in *Drosophila*. Examples include RNA processing and protein modification, and several kinds of extracellular and intracellular signalling. Finally, all elements of the metabolic processes in *Drosophila* have counterparts in unicellullar organisms. In all of these organisms, the functions of intermediary metabolism are similar. Over half of Ashburner's scheme is concerned with one or another metabolic process that has counterparts in lower organisms.

Other commonalities exist across kingdom boundaries, but only come to light through experimentation and close inspection. It has been known for some time that the principal protein in the lens of the eye is similar in sequence to an enzyme from *E. coli* [15]. Recently, the close connections between a mouse and an *E. coli* protein were shown dramatically when a mouse protein, Btcd (binds to curved DNA), transferred to *E. coli*, was able to substitute for the missing H-NS protein of mutant *E. coli* [16•]. The gene for H-NS is often described as a broad regulatory protein and is present in the nucleoid. In order to recognize the connection between genes of *E. coli* and mouse through a desciption of their function in any kind of automatic system, the description of the functions of the gene products would have had to include the DNA-binding property.

Improvements needed

Attempts to describe the cellular and organismal function of any one protein in a few words presents a challenge. Biological complexity is often difficult to summarize briefly. How can current schemes for classifying functions of gene products be improved, with an eye towards future connectivity? There are two main ways — one is to use the same words to describe the same function in different organisms, the other is to more correctly represent complexity by describing the multiple functions of some gene products and the multiple aspects of the biology of organisms.

The first scheme will require cooperation in establishing a commonly used vocabulary for similar functions in different organisms. The wide choice of words available in English to describe the same thing, and the existence of multiple aspects of function for one gene product complicates efforts to discover commonality. Stan Letovsky [17••] discusses the desirability of a common vocabulary on the World Wide Web. The second objective, to capture multiplicity, will require the expansion of functional schemes to allow multiple aspects of the function of a given gene product to be reflected in the classification scheme.

Microbial physiologists, biochemists, structural biologists and others outside the field of genomics are justified in viewing the current flurry of naming gene product function in a few words as naive and primitive. A relatively simple one-dimensional classification system is seriously lacking in full information concerning the physiological role of many gene products, especially those that are required to play more than one role in the cell. A one-dimensional listing is also seriously lacking in logic of organization, juxtaposing as it does 'apples and oranges' of cell function. There is need to expand, rearrange and further supplement the classification system so that the guide to understanding the role of gene products has a higher degree of scientific validity.

Beginnings

We need to be able to attach more than one kind of designation to a gene product. For instance, the current annotation of the functions of *E. coli* genes mixes two kinds of definitions, one is the type of protein, such as enzyme, regulator, transport protein, the other is the type of cellular function, such as electron transport, carbohydrate degradation, macromolecular biosynthesis [5•,18]. Ashburner [14••] has defined the former as 'functional primitives' and the latter as 'process'. Any one gene product carries designators for both systems. Both kinds of description are useful.

Moving beyond the current approach, we need to recognize that a protein can often play a part in more than one process. If metabolism is a process and adaptation is a process, then some reactions of trehalose could be classified as part of both carbohydrate metabolism and the adaptive osmoprotection process. A permease for the transport of galactosides could be classified in both transport and carbohydrate metabolism. The lac repressor could be classified as functioning in both carbohydrate metabolism and osmoprotection. The multisubunit enzyme succinic dehydrogenase could be classified as a whole in metabolism (as an enzyme in the tricarboxylic acid cycle) yet two subunits could be classified as members of the electron transport chain and another subunit is part of the membrane, and could therefore be classified in cell structure. Flagellar proteins could be designated as part of cell structure and they could also be

considered an essential part of the process of motility. A final example within the metabolism category is the enzyme acetate kinase, which functions when acetate is the only carbon source as an early enzyme in the utilization of acetate for growth. When the cell is living via fermentation, however, the same enzyme can function in the opposite direction to generate ATP from acetyl phosphate, producing acetate as an end product of fermentation. Multiple classification schemes will capture this complexity.

I propose a multidimensional system in which gene products are labeled with more than one attribute. One dimension could be 'metabolism', including all the metabolic processes of synthesis and degradation of small molecules, and also the synthesis, modification, maturation, translocation, repair and recombination of large molecules. Any one gene product can be assigned to more than one metabolic role. Several kinds of databases of metabolism exist. EcoCyc is a computable system of metabolic reactions, compounds, genes and pathways of *E. coli* [6]. KEGG is based on a simpler program and includes metabolic reactions from many organisms [19] The WIT site offers a comprehensive collection of ASCII files representing reactions, pathways, and the properties of enzymes from many organisms [20].

When using 'metabolism' as one of the dimensions of several functional categories, all proteins involved in supporting any biochemical tranformation would be included — not just the enzymes of the reaction pathway but also the related regulators and the transport systems specific to the uptake of relevant substrates.

The second dimension could be the type of 'regulation'. Even though individual regulators are associated with the metabolic activity they affect in the metabolism dimension, in this dimension, all regulators would be organized by the type of mechanism. Some are transcriptional activators or repressors, some are members of the more complex two component sensor-responder systems, some are narrow in effect, some are broad and some are global. Regulators of different kinds interact with different molecules — through a site in DNA or through protein–protein interactions.

A third dimension could be 'transport'. Again, a specific transporter will already have been assigned a metabolic role. In this classification, the type of transporter for a molecular mechanism would be captured. For instance, a transporter could be classified as either a phosphotransferase transport system enzyme, a high-affinity binding protein, a secretion system for specific proteins, or a porin in the outer membrane.

Another dimension could be 'structure' which would contain structural elements of the cell and would also specify the cellular 'location' for any gene product as, for instance, cytoplasmic, inner membrane, periplasmic, outer membrane, appendage or extracellular coating. Many metabolic enzymes and transporter proteins are located at the inner membrane and many form an integral part of the membrane. They are therefore both part of a metabolic pathway and part of the structure of the membrane. Porin proteins function as channels and are also part of the outer membrane.

All of these categories are relevant to higher organisms but, in addition, the higher organisms would have a greater complexity than unicellular organisms. Structure, for instance, applies not only to cell structure but to the entire gross anatomy of a differentiated organism.

Additional categories are needed for 'other processes' besides metabolism, such as 'cell division', 'genetic exchange', the 'adaptation strategies' of the cell (e.g. osmoregulation and SOS repair systems), and 'motility and chemotaxis' (to include the relevant two component regulation systems as well as the motor systems). Other categories will be necessary in order to describe other organisms. *B. subtilis*, for instance, needs categories for 'sporulation', germination' and 'transformation'. For yeast, additional categories, such as 'protein destination', 'intracellular transport' and 'biogenesis of cell components' are needed. When applying all relevant categories that characterize any one gene product, one sees that many gene products could be assigned labels from several of the categories, thus building a more complete picture of the multiple roles of the gene products within the life of the cell.

For multicellular organisms, a more complex scheme of function will be needed. Michael Ashburner has developed not only the detailed classification of *Drosophila* gene functions mentioned above, but also a detailed vocabulary and anatomy of the fly that will help immeasurably in making connections between those functions that are common for *Drosophia* and more primitive unicellular organisms [21•]. This work presages the needs of the yet more complex mammalian systems. In order to facilitate comparability among organisms, perhaps we can now agree to organize classes of gene function along similar lines, and to describe functions that are held in common in similar terms, with similar organization, and a similar vocabulary.

Other beginnings

We may ask what other beginnings have been made to systemize biological information along these lines across all kingdoms. Many biochemical reactions relevant to metabolism have been systematized in a numerical hierarchical scheme. It is tempting to take advantage of the large body of work carried out by the Enzyme Commission (EC) of the International Union of Biochemistry and Molecular Biology [22]. This EC system presents reactions in an arbitrary left to right direction and organizes the reactions into six major classes. Each of the six major categories are broken down further into subcategories hierarchically, to a total of four levels represented by four numbers in sequence. Each reaction is therefore specified by four numerals, such as 3.4.5.21. Enzyme nomenclature is actu-

ally a classification of reactions, not of individual enzymes. The classification system is not intended to give information on the properties of the various enzyme catalysts that exist in different organisms, nor does it give any information on the commonly used, biologically relevant direction of a reaction.

Beyond that, the greatest drawback to adopting the numerical system as biologically meaningful is the lack of information on the mechanism of reaction. It is the case, however, that the biologically relevant similarity of proteins, even at the sequence level, is strongly correlated with similarity of their mechanism of action. One wants to bring the similarity of mechanisms of action into the equation when attempting to connect proteins with similar functions. The classification reflects the kinds of changes occurring to the substrates in the reaction and what the products of the reaction are, but it is important to understand that they do not reflect the mechanism of reaction.

Another complication of the EC numbering system is the fact that genes, proteins and reactions do not always have a 1:1:1 relationship, so that one EC number does not always apply to one gene that encodes one gene product that catalyses one reaction. On the contrary, one reaction that has one particular EC number may be catalysed by a multifunctional polypeptide chain that also carries out other reactions with other EC numbers. Likewise, more than one gene encoding more than one polypeptide chain may be required to form a complex that carries out one reaction with one EC number.

This discussion of the meaning of EC numbers is intended to illustrate the complexities inherent in efforts to describe complex biological phenomena systematically. Because they do not portray reaction mechanisms, properties of enzymes or relationships with genetic determination, EC numbers have limitations as handles for revealing biochemical or genetic equivalence among organisms or even within one organism. To be aware of this kind of limitation is to be aware of the difficulties and demands of any system that undertakes systematic expression of gene and gene product function.

Difficulty with comparability can be illustrated by glutamine amidotransferase subunits. The glutamine amidotransferase subunit of anthranilate synthetase (one of the TrpD subunit functions) and the glutamine amidotransferase subunit of carbamyl-phosphate (a CarA subunit function) carry out similar reactions using a similar mechanism. This relatedness is lost when only the holoenzyme reaction is numbered, in the first case as EC number 4.1.3.27 and, in the second case, as EC number 6.3.5.5. Yet, the CarA and TrpD proteins are members of a distinct class of proteins with glutamine amidase activity. They are recognizably similar in amino acid sequence, reflecting the similar function and mechanism of reaction, yet this similarity is not revealed by the holoenzyme EC numbers.

Recently, transport proteins have been categorized using a hierarchical numbering system similar to that of the EC system [23]. Transport processes and types of proteins are probably inherently more consistent and straightforward than enzymes, and their mechanisms of action are for the most part known. Thus, they lend themselves to this kind of organizing approach.

The classification of proteins into families according to their secondary or tertiary structure seems to be the way to recognize connections between distant relatives. This is the path to follow in order to establish comparability of proteins in all organisms. Unless two proteins are very close in sequence, the prediction of structure based upon primary sequence is firm mainly at the secondary-structure level [24•]. In lucky cases, however, an unknown protein will be demonstrably similar to a protein whose tertiary structure is known. The SCOP database presents protein families that have similar three-dimensional structures [25•]. Its creators describe it as "a detailed and comprehensive description of the structural and evolutionary relationships between all proteins whose structure is known". CATH is a hierarchical classification of domain structures [26]. The HSSP database combines information on sequence similarities with known three-dimensional structures [27•]. The related Dali program classifies fold structures. The Dali/FSSP web site offers a "network of links between neighbors in fold space, between domains and proteins, and between structures and sequences" [28•]. Through sequence similarities, one may find a similarity between an unknown protein and a protein of known structure. As always, care is necessary when similarities fall in the twilight zone. Thus, tentative classification into a protein family could be particularly useful as an attribute describing any protein gene product.

Conclusions

An effort to coordinate existing information about comparable functions of genes and gene products in different organisms should lay a useful foundation for assimilating the massive information expected to result from the current genomic sequencing projects. The disciplines of comparative biology and evolutionary biology will be enriched by such data. Relatedness can be established from similarities of primary sequences, secondary and tertiary structures of proteins, and descriptions of the cellular functions of gene products. The capacity to extract biological meaning rests to some extent on establishing comparable systems of descriptors for the genes and gene products of different organisms. This will probably be a major effort, and now is the time to begin.

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. Bachmann BJ: **Linkage map of *Escherichia* K-12, edition 8.** *Microbiol Rev* 1990, **54**:130-197.

2. Berlyn MKB, Low KB, Rudd KE, Singer M: **Linkage map of *Escherichia coli* K-12, edition 9.** In *Escherichia coli and Salmonella*: *Cellular and Molecular Biology,* edn 2. Edited by Neidhardt FN, Curtiss R III, Lin ECC, Ingraham JL, Low KB, Magasanik B, Reznikoff W, Riley M, Schaechter M, Umbarger E. Washington, DC: American Society for Microbiology Press; 1996:1715-1902.

3. Ingraham JL, Maaloe O, Neidhardt FC: **Appendix B.** In *Growth of the Bacterial Cell.* Sunderland, Massachusetts: Sinauer Associates Incorporated; 1983:391-403.

4. Riley M: **Functions of the gene products of *E. coli*.** *Microbiol Rev* 1993, **57**:862-952.

5. • Riley M, Labedan B: **E. coli gene products: physiological functions and common ancestries.** In *Escherichia coli and Salmonella*: *Cellular and Molecular Biology,* edn 2. Edited by Neidhardt FN, Curtiss R III, Lin ECC, Ingraham JL, Low KB, Magasanik B, Reznikoff W, Riley M, Schaechter M, Umbarger E. Washington, DC: ASM Press; 1996:2118-2202.

The basic scheme for organizing *E. coli* cellular functions – connecting gene products and genes.

6. Karp P, Riley M, Paley SM, Pellegrini-Toole A, Krummenacker M: **EcoCyc: encyclopedia of *Escherichia* genes and metabolism**. *Nucleic Acids Res* 1998, **26**:50-53. [URL: http://www.ecocyc.PangeaSystems.com/ecocyc/ecocyc.html]

7. Riley M: **Genes and proteins of *Escherichia coli* K-12 (GenProtEC).** *Nucleic Acids Res* 1998, **26**:54. [URL: http://mbl.edu/html/ecoli.html]

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

9. Tomb JF, White O, Kerlavage AR, Clayton RA, Sutton GG, Fleischmann RD, Ketchum KA, klenk HP, Gill S, Dougherty BA *et al.*: **The complete genome sequence of the gastric pathogen *Helicobacter pylori*.** *Nature* 1997, **388**:539-547. [URL: http://www.tigr.org/tdb/mdb/hpdb/hpdb.html]

10. Riley M, Labedan B: **Protein evolution viewed through *Escherichia coli* protein sequences: introducing the notion of a structural segment of homology, the module**. *J Mol Biol* 1997, **268**:857-868.

11. Kunst F, Ogasawara N, Moszer I, Albertini AM, Alloni G, Azevedo V, Bertero MG, Bessieres P, Bolotin A, Borchert S *et al.*: **The complete genome sequence of the gram-positive bacterium *Bacillus subtilis*.** *Nature* 1997, **390**:249-256. [URL: http://www.pasteur.fr/Bio/SubtiList/help/classif-search.html]

12. • Mewes HW, Albermann K, Heumann K, Liebl S, Pfeiffer F: **MIPS: a database for protein sequences, complete genomes.** *Nucleic Acid Res* 1998, **25**:28-30. [URL: http://www.mips.biochem.mpg.de/mips/yeast/index.htmlx]

This database provides several 'catalogs' which are ways of viewing gene products in different ways. In the functional catalog, sometimes useful pictures explain the inter-relationship of reactions and protein interactions. Pathway catalogs are under construction. Information provided includes prosite motifs, EC numbers, types of proteins, protein complexes (with links to documentation).

13. • Hodges PE, Payne WE, Garrels JI: **Yeast protein database (YPD): a curated proteome database for *Saccharomyces cerevisiae.*** *Nucleic Acid Res* 1998, **25**:68-72. [URL: http://www.proteome.com/YPD_contents_by_category.html]

This database organizes gene products under several banners: localization, molecular environment, post-translational modification and gene location on chromosomes.

14. • *Drosophila* functions available by FTP from: ftp.ebi.ac.uk/pub/databases/edgp/misc/ashburner/fly_function_tree.

An incredibly detailed listing of *Drosophila* functions from Michael Ashburner. A first attempt at classifying functions of gene products for a higher organism. The next step is to populate the list with genes and proteins.

15. Carper D, Nishimura C, Shinohara T, Dietzchold B, Wistow G, Craft C, Kador P, Kinoshita JH: **Aldose reductase and p-crystallin belong to the same protein superfamily as aldehyde reductase**. *FEBS Lett* 1987, **220**:209-213.

16. • Timchenko T, Bailone A, Devoret R: **Btcd, a mouse protein that binds to curved DNA can substitute in *Escherichia coli* for H-NS, a bacterial nucleoid protein.** *EMBO J* 1996, **15**:3986-3992.

Mouse and *E. coli* both have proteins that bind to DNA. The trick is to recognize their similarity by taking care when describing the functions of the gene products in databases.

17. • On World Wide Web URL: http://info.gdb.org/Letovsky/provence.html.

This site begins the discussion on correlating information on gene products by developing a common vocabulary.

18. Blattner FR, Plunkett G III, Bloch CA, Perna NT, Burland V, Riley M, Collado-Vides J, Glasner JD, Rode CK, Mayhew G *et al.*: **The complete genome sequence of *Escherichia coli* K-12**. *Science* 1997, **277**:1453-1462. [URL: http://www.genetics.wisc.edu/html/k12.html]

19. Goto S, Bono H, Ogata H, Fujibuchi W, Nishioka T, Sato K, Kanehisa M: **Organizing and computing metabolic pathway data in terms of binary relations**. *Pac Symp Biocomput* **1997**:175-186. [URL: http://www.genome.ad.jp/kegg/docs/intro.html]

20. Selkov E Jr, Grechkin Y, Mikhailova N, Selkov E: **MPW: the metabolic pathways database.** *Nucleic Acids Res* 1998, **26**:43-45. [URL: http://www.mcs.anl.gov/home/compbio/WIT/wit.html]

21. • Proforma controlled vocabularies on World Wide Web URL: http://www.ebi.ac.uk:7081/docs/flydocs/flybase/controlled-vocabularies.txt.

An early attempt at a controlled vocabulary. The site includes components of the gross anatomy (body parts) of the fruit fly, only the metabolism is shared by lower organisms.

22. Webb EC (Ed): *Enzyme Nomenclature 1992. Recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology.* New York: Academic Press; 1992.

23. Paulsen IT, Sliwinski MK, Saier MH Jr: **Microbial genome analyses: global comparisons of transport capabilities based on phylogenies, bioenergetics and substrate specificities.** *J Mol Biol* 1998, **277**:573-592.

24. • Gerloff DL, Cohen FE, Korostensky C, Turcotte M, Gonnet GH, Benner SA: **A predicted consensus structure for the N-terminal fragment of the heat shock protein HSP90 family.** *Proteins* 1997, **27**:450-458.

The approach used here has been successful in predicting secondary structures of many proteins – this is just one example.

25. • Murzin AG, Brenner SE, Hubbard T, Chothia C: **SCOP: a structural classification of proteins database for the investigation of sequences and structures.** *J Mol Biol* 1995, **247**:536-540. [URL: http://scop.mrc-lmb.cam.ac.uk/scop/]

An indispensable reference resource. It offers a BLAST search for similar proteins whose structures are known.

26. Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, Thornton JM: **CATH – a hierarchic classification of protein domain structures**. *Structure* 1997, **5**:1093-1108

27. • Dodge C, Schneider R, Sander C: **The HSSP database of protein structure-sequence alignments and family profiles**. *Nucleic Acids Res* 1998, **26**:313-315. [URL: http://www.sander.embl-ebi.ac.uk/hssp/]

This is a comprehensive collection of data on protein families and relationships between structure and sequence alignments. Wandering around this site gives one a sense of how the field stands today and where it is going tomorrow.

28. • Holm L, Sander C: **Touring protein fold space with Dali/FSSP.** *Nucleic Acids Res* 1998, **26**:319-319. [URL: http://www.embl-ebi.ac.uk/dali]

Emphasis is on the folds that appear in various combinations in proteins. Many useful database connections are provided.