

4-3 Deriving Function from Sequence

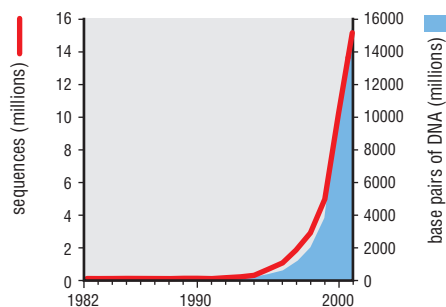


Figure 4-9 The growth of DNA and protein sequence information collected by GenBank over 20 years. There has been an exponential increase in both base pairs of DNA sequence and coding sequences, especially since 1994 when various genomics projects were initiated. (Information from <http://www.ncbi.nlm.nih.gov/Genbank/genbankstats.html>)

Sequence information is increasing exponentially

During the past decade, more than 800 organisms have been the object of genome-sequencing projects. We now know the complete DNA sequences of the genomes of over 100 species of bacteria and archaea, including some important pathogens, and three yeasts, and have partial or complete genome sequences of a number of protozoan parasites. Among multicellular organisms, the genomes of the nematode worm (*Caenorhabditis*), the fruit fly (*Drosophila*) and the plants *Arabidopsis thaliana* and rice have also been completely sequenced. The human genome sequence is now completely finished and a draft mouse genome sequence has also been completed. The growth of sequence information is exponential, and shows no sign of slowing down (Figure 4-9). However, in all these organisms the biochemical and cellular functions of a large percentage of the proteins predicted from these sequences are at present unknown.

It is hardly surprising, therefore, that much effort is being expended on the attempt to define the structures and functions of proteins directly from sequence. Such efforts are based on comparison of sequences from many different organisms using computational tools such as BLAST to retrieve related sequences from the databases (see section 4-1). Attempts to derive function from sequence depend on the basic assumption that proteins that are related by sequence will also be related by structure and function. In this chapter, we will show that the assumption of structural relatedness is usually valid, but that function is less reliably determined by such methods. Structure and function can be derived in this way only for sequences that are quite closely related to those encoding proteins of known structure and function, and sometimes not even then.

As one proceeds from prokaryotes to eukaryotes, and from single-celled to multicellular organisms, the number of genes increases markedly (Figure 4-10), by the addition of genes such as those involved in nuclear transport, cell-cell communication, and innate and acquired immunity. The number of biochemical functions also increases. With increasing evolutionary distance, sequences of proteins with the same structure and biochemical function can diverge so greatly as to render any relationship extremely difficult to detect. Consequently, defining functions for gene products from higher organisms by sequence comparisons alone will be difficult until even more sequences and structures are collected and correlated with function.

In some cases function can be inferred from sequence

If a protein has more than about 40% sequence identity to another protein whose biochemical function is known, and if the functionally important residues (for example, those in the active site of an enzyme) are conserved between the two sequences, it has been found that a reasonable working assumption can be made that the two proteins have a common biochemical function (Figure 4-11). The 40% rule works because proteins that are related by descent and have the same function in different organisms are likely still to have significant sequence similarity, especially in regions critical to function. Sequence comparison will not, however, detect proteins of identical structure and biochemical function from organisms so remote from one another on the evolutionary tree that virtually no sequence identity remains. Moreover, identity of biochemical function does not necessarily mean that the cellular and other higher-level

Genome Sizes of Representative Organisms

Organism	Genome size (base pairs)	Number of genes
<i>Mycoplasma genitalium</i>	45.8×10^5	483
<i>Methanococcus jannaschii</i>	1.6×10^6	1,783
<i>Escherichia coli</i>	4.6×10^6	4,377
<i>Pseudomonas aeruginosa</i>	6.3×10^6	5,570
<i>Saccharomyces cerevisiae</i>	1.2×10^7	6,282
<i>Caenorhabditis elegans</i>	1.0×10^8	19,820
<i>Drosophila melanogaster</i>	1.8×10^8	13,601
<i>Arabidopsis thaliana</i>	1.2×10^8	25,498
<i>Homo sapiens</i>	3.3×10^9	~30,000 (?)

Figure 4-10 Table of the sizes of the genomes of some representative organisms. The first four organisms are prokaryotes. A continuous update on sequencing projects, both finished and in progress, may be found at <http://www.genomesonline.org>

References

Brenner, S.: **Theoretical biology in the third millennium.** *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* 1999, **354**: 1963–1965.

Brizuela, L. et al.: **The FLEXGene repository: exploiting the fruits of the genome projects by creating a needed resource to face the challenges of the post-genomic era.** *Arch. Med. Res.* 2002, **33**:318–324.

Domingues, F.S. et al.: **Structure-based evaluation of sequence comparison and fold recognition align-**

ment accuracy. *J. Mol. Biol.* 2000, **297**:1003–1013.

Hegyil, H. and Gerstein, M.: **Annotation transfer for genomics: measuring functional divergence in multi-domain proteins.** *Genome Res.* 2001, **11**:1632–1640.

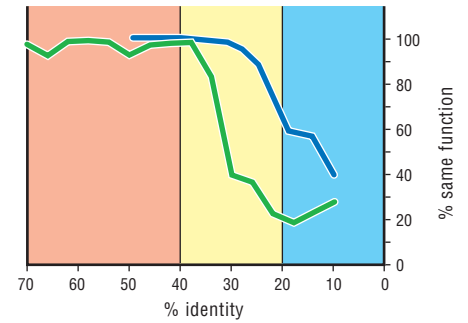
Genomic and protein resources on the Internet:

http://bioinfo.mbb.yale.edu/lectures/spring2002/show/index_2

<http://www.genomesonline.org>

<http://www.ncbi.nlm.nih.gov/Genbank/genbankstats.html>

Figure 4-11 Relationship of sequence similarity to similarity of function The percentage of protein pairs with the same precise biochemical function is plotted against the sequence identity (enzymes, blue curve; non-enzymes, green curve). The orange area represents proteins whose fold and function can be reliably predicted from sequence comparison. The yellow area represents proteins whose fold can be predicted from sequence but whose precise function cannot. The blue area represents proteins for which neither the fold nor the function can reliably be predicted from sequence. Note that below about 40% identity, the probability of making an incorrect functional assignment increases dramatically. Adapted from an analysis by Mark Gerstein (http://bioinfo.mbb.yale.edu/lectures/spring2002/show/index_2).



functions of the proteins will be similar. Such functions are expressed in a particular cellular context and many proteins, such as hormones, growth factors and cytokines, have multiple functions in the same organism (see section 4-13).

Local alignments of functional motifs in the sequence (see section 4-2) can often identify at least one biochemical function of a protein. If the sequence motif is large enough and contiguous, it can identify an entire domain or structural module with a recognizable fold and function. For example, helix-turn-helix motifs (see Figure 1-50) and zinc finger motifs (see Figure 1-49) are

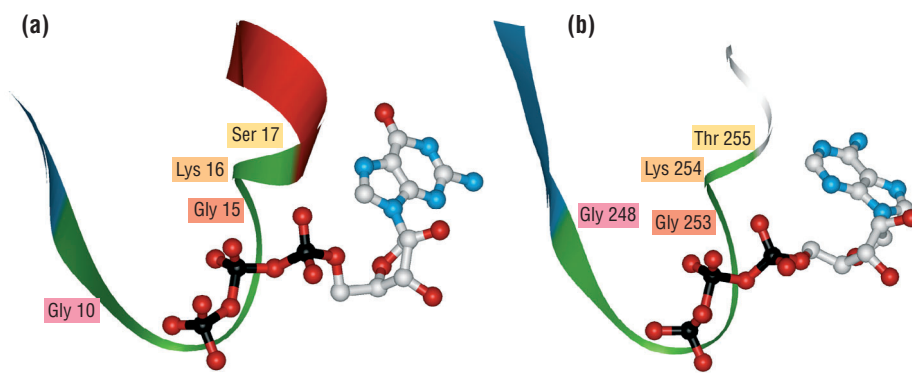


Figure 4-12 The P loop of the Walker motif

A contiguous sequence block, the so-called Walker A block or P loop, is a stretch of sequence with a consensus pattern of precisely spaced phosphate-binding residues; this is found in a number of ATP- or GTP-binding proteins, for example ATP synthase, myosin heavy chain, helicases, thymidine kinase, G-protein alpha subunits, GTP-binding elongation factors, and the Ras family. The consensus sequence is: [A or G]XXXGK[S or T]; this forms a flexible loop between alpha-helical and beta-pleated-sheet domains of the protein in question. The proteins may have quite different overall folds. The triphosphate group of ATP or GTP is bound by residues from the P loop. Shown are the interactions (a) of GTP with the P loop of the signaling protein H-Ras (PDB 1qra) and (b) of ATP with the P loop of a protein kinase (PDB 1q2).

often recognizable in the sequence and are diagnostic for, respectively, small secondary structure elements and small domains that potentially bind DNA. The SH2 and SH3 domains present in many signal transduction proteins can also often be recognized by characteristic stretches of sequence. When present, such sequences usually indicate domains that are involved in the recognition of phosphotyrosines or proline-rich sequences, respectively, in dynamic protein-protein interactions. The so-called Walker motif, which identifies ATP- and GTP-binding sites, is also easily identified at the level of sequence, although its presence does not reveal what the nucleotide binding is used for and it is found in many different protein folds. The Walker motif is actually three different, non-contiguous stretches of sequence, labeled Walker A, B, and C. Of these, the Walker A motif, or P loop, which defines the binding site for the triphosphate moiety, is the easiest to recognize (Figure 4-12 and see Figure 4-7). The B and C motifs interact with the base of the nucleotide.

Sequence comparison is such an active area of research because it is now the easiest technique to apply to a new protein sequence. Figure 4-13 shows an analysis of the functions of all the known or putative protein-coding sequences in the yeast genome: some of these are experimentally established, but a large proportion are inferred only by overall sequence similarity to known proteins (labeled homologs in the figure) or by the presence of known functional motifs, and 32% of them are unknown. Similar distributions are observed for many other simple organisms. For more complex organisms, the proportion of proteins of unknown function increases dramatically. Current efforts are focused on ways of identifying structurally and functionally similar proteins when the level of sequence identity is significantly below the 40% threshold. As we shall see, identification of structural similarity is easier and more robust than the identification of functional similarity.

Figure 4-13 Analysis of the functions of the protein-coding sequences in the yeast genome Some are known experimentally, some are surmised from sequence comparison with proteins of known function in other organisms, and some are deduced from motifs that are characteristic of a particular function. Some of these surmised functions may not be correct, and a large percentage of the coding sequences cannot at present be assigned any function by any method.

