4

From Sequence to Function:

Case Studies in Structural and Functional Genomics

One of the main challenges facing biology is to assign biochemical and cellular functions to the thousands of hitherto uncharacterized gene products discovered by genome sequencing. This chapter discusses the strengths and limitations of the many experimental and computational methods, including those that use the vast amount of sequence information now available, to help determine protein structure and function. The chapter ends with two individual case studies that illustrate these methods in action, and show both their capabilities and the approaches that still must be developed to allow us to proceed from sequence to consequence.

4-0 Overview: From Sequence to Function in the Age of Genomics

Genomics is making an increasing contribution to the study of protein structure and function

The relatively new discipline of **genomics** has great implications for the study of protein structure and function. The genome-sequencing programs are providing more amino-acid sequences of proteins of unknown function to analyze than ever before, and many computational and experimental tools are now available for comparing these sequences with those of proteins of known structure and function to search for clues to their roles in the cell or organism. Also underway are systematic efforts aimed at providing the three-dimensional structures, subcellular locations, interacting partners, and deletion phenotypes for all the gene products in several model organisms. These databases can also be searched for insights into the functions of these proteins and their corresponding proteins in other organisms.

Sequence and structural comparison can usually give only limited information, however, and comprehensively characterizing the function of an uncharacterized protein in a cell or organism will always require additional experimental investigations on the purified protein *in vitro* as well as cell biological and mutational studies *in vivo*. Different experimental methods are required to define a protein's function precisely at biochemical, cellular, and organismal levels in order to characterize it completely, as shown in Figure 4-1.

In this chapter we first look at methods of comparing amino-acid sequences to determine their similarity and to search for related sequences in the sequence databases. Sequence comparison alone gives only limited information at present, and in most cases, other experimental and structural information is also important for indicating possible biochemical function and mechanism of action. We next provide a summary of some of the genome-driven experimental tools for probing function. We then describe computational methods that are being developed to deduce the protein fold of an uncharacterized protein from its sequence. The existence of large families of structurally related proteins with similar functions, at least at the biochemical level, is enabling sequence and structural motifs characteristic of various functions to be identified. Protein structures can also be screened for possible ligand-binding sites and catalytic active sites by both computational and experimental methods.

As we see next, predicting a protein's function from its structure alone is complicated by the fact that evolution has produced proteins with almost identical structures but different functions, proteins with quite different structures but the same function, and even multifunctional proteins which have more than one biochemical function and numerous cellular and physiological functions. We shall also see that some proteins can adopt more than one stable protein fold, a change which can sometimes lead to disease.

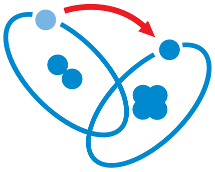
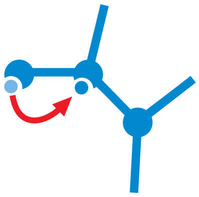
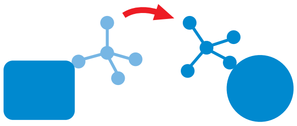
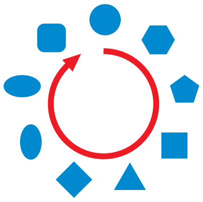
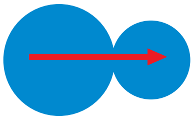
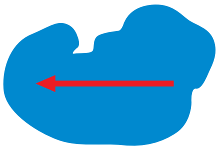
The chapter ends with two case histories illustrating how a range of different approaches were combined to determine aspects of the functions of two uncharacterized proteins from the genome sequences of *E. coli* and yeast, respectively.

Figure 4-1 **Time and distance scales in functional genomics** The various levels of function of proteins encompass an enormous range of time (scale on the left) and distance (scale on the right). Depending on the time and distance regime involved, different experimental approaches are required to probe function. Since many genes code for proteins that act in processes that cross multiple levels on this diagram (for example, a protein kinase may catalyze tyrosine phosphorylation at typical enzyme rates, but may also be required for cell division in embryonic development), no single experimental technique is adequate to dissect all their roles. In the age of genomics, interdisciplinary approaches are essential to determine the functions of gene products.

Definitions

genomics: the study of the DNA sequence and gene content of whole genomes.

Overview: From Sequence to Function in the Age of Genomics 4-0

Time		Process	Example System	Example Detection Methods	Distance
10^{-15} sec		electron transfer	photosynthetic reaction center	optical spectroscopy	1 Å
10^{-9} sec		proton transfer	triosephosphate isomerase	fast kinetics	2–10 Å
10^{-6} sec		fastest enzyme reactions	catalase, fumarase, carbonic anhydrase	kinetics	
10^{-3} sec		typical enzyme reactions	trypsin, protein kinase A, ketosteroid isomerase	kinetics, time-resolved X-ray, nuclear magnetic resonance	
sec		slow enzyme reactions/cycles	cytochrome P450, phosphofructokinase	kinetics, low T X-ray, nuclear magnetic resonance, mass spectroscopy	Å – nm
min/ hour		protein synthesis/ cell division	budding yeast cell	light microscopy, genetics, optical probes	nm – µm
day/ year		embryonic development	mouse embryo	genetics, microscopy, microarray analysis	µm – m

References

Brazhnik, P. *et al.*: **Gene networks: how to put the function in genomics.** *Trends Biotechnol.* 2002, **20**:467–472.

Chan, T.-F. *et al.*: **A chemical genomics approach toward understanding the global functions of the target of rapamycin protein (TOR).** *Proc. Natl Acad. Sci. USA* 2000, **97**:13227–13232.

Guttmacher A.E. and Collins, F.S.: **Genomic medicine—a primer.** *N. Engl. J. Med.* 2002, **347**:1512–1520.

Houry, W.A. *et al.*: **Identification of *in vivo* substrates of the chaperonin GroEL.** *Nature* 1999, **402**:147–154.

Koonin E.V. *et al.*: **The structure of the protein universe and genome evolution.** *Nature* 2002, **420**:218–223.

O'Donovan, C. *et al.*: **The human proteomics initiative (HPI).** *Trends Biotechnol.* 2001, **19**:178–181.

Oliver S.G.: **Functional genomics: lessons from yeast.** *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* 2002, **357**:17–23.

Quevillon-Cheruel, S. *et al.*: **A structural genomics initiative on yeast proteins.** *J. Synchrotron. Radiat.* 2003, **10**:4–8.

Tefferi, A. *et al.*: **Primer on medical genomics parts I–IV.** *Mayo. Clin. Proc.* 2002, **77**:927–940.

Tong, A.H. *et al.*: **Systematic genetic analysis with ordered arrays of yeast deletion mutants.** *Science* 2001, **294**:2364–2368.

von Mering, C. *et al.*: **Comparative assessment of large-scale data sets of protein–protein interactions.** *Nature* 2002, **417**:399–403.

4-1 Sequence Alignment and Comparison

```
S.c. Kss1 INNQNSGFSTLSDDHVQYFTYQILRALKSIHSAQVI
H.s. Erk2 LKTQH-----LSNDHICYFLYQILRGLKYIHSANVL

          HRDIKPSNLLLNSNCDLKVCDFGLARCLASSSDSRET
          HRDLKPSNLLLNTTCDLKICDFGLARVA----DPDHD
```

Figure 4-2 Pairwise alignment Part of an alignment of the amino-acid sequences of the kinase domains from two ERK-like kinases of the MAP kinase superfamily, Erk2 from humans and Kss1 from yeast. The region shown covers the kinase catalytic loop and part of the activation loop (see Figure 3-24). Identical residues highlighted in purple show the extensive similarity between these two homologous kinases (their evolutionary relationship can be seen in Figure 4-5). To maximize similarity, a small number of gaps have had to be inserted in the human sequence.

Sequence comparison provides a measure of the relationship between genes

The comparison of one nucleotide or amino-acid sequence with another to find the degree of similarity between them is a key technique in present-day biology. A marked similarity between two gene or protein sequences may reflect the fact that they are derived by evolution from the same ancestral sequence. Sequences related in this way are called **homologous** and the evolutionary similarity between them is known as **homology**. Unknown genes from newly sequenced genomes can often be identified by searching for similar sequences in databases of known gene and protein sequences using computer programs such as BLAST and FASTA. Sequences of the same protein from different species can also be compared in order to deduce evolutionary relationships. Two genes that have evolved fairly recently from a common ancestral gene will still be relatively similar in sequence to each other; those that have a more distant common ancestor will have accumulated many more mutations, and their evolutionary relationship will be less immediately obvious, or even impossible to deduce from sequence alone.

Alignment is the first step in determining whether two sequences are similar to each other

A key step in comparing two sequences is to match them up to each other in an **alignment** that shows up any similarity that is present. Alignments work on the general principle that two homologous sequences derived from the same ancestral sequence will have at least some identical residues at the corresponding positions in the sequence; if corresponding positions in the sequence are aligned, the degree of matching should be statistically significant compared with that of two randomly chosen unrelated sequences.

As a quantitative measure of similarity, a pairwise alignment is given a score, which reflects the degree of matching. In the simplest case, where only identical matched residues are counted, the fraction of identical amino acids or nucleotides gives a similarity measure known as **percent identity**. When protein sequences are being compared, more sophisticated methods of assessing similarity can be used. Some amino acids are more similar to each other in their physical-chemical properties, and consequently will be more likely to be substituted for each other during evolution (see section 1-1). Most of the commonly used alignment programs give each aligned pair of amino acids a score based on the likelihood of that particular match occurring. These scores are usually obtained from reference tables of the observed frequencies of particular substitutions in sets of known related proteins (see Figure 1-6). The individual scores for each position are summed to give an overall similarity score for the alignment.

In practice, insertions and deletions as well as substitutions will have occurred in two homologous sequences during their evolution. This usually results in two gene or protein sequences of different lengths in which regions of closely similar sequence are separated by dissimilar regions of unequal length. In such cases, portions of the sequence are slid over each other when making the alignment, in order to maximize the number of identical and similar amino acids. Such sliding creates gaps in one or other of the sequences (Figure 4-2). Experience tells us that closely related sequences do not, in general, have many insertions or deletions relative to each other. Because any two sequences could be broken up randomly into as many gaps as needed to maximize matching, in which case the matching would have no biological significance, gaps are subject to a penalty when scoring sequence relatedness.

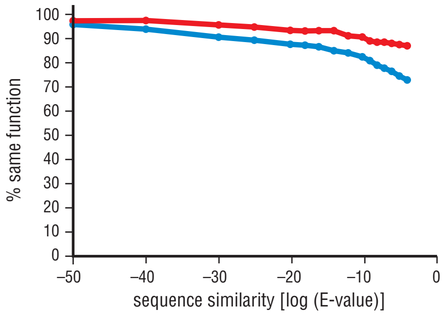


Figure 4-3 Plot of percentage of protein pairs having the same biochemical function as sequence changes When a series of sequences of homologous proteins are compared, it is observed that as sequence similarity (measured by the E-value from a sequence comparison) decreases, the probability that homologs will have the same function also decreases. The red curve corresponds to single-domain proteins, the blue curve to multidomain proteins. Up to an E-value of approximately 10^{-10}, the likelihood of an identical function is reasonably high, but then it starts to decrease substantially, especially for multidomain proteins.

Definitions

alignment: procedure of comparing two or more sequences by looking for a series of characteristics (residue identity, similarity, and so on) that match up in both and maximize conservation, in order to assess overall similarity.

conserved: identical in all sequences or structures compared.

E-value: the probability that an **alignment** score as good as the one found between two sequences would be found in a comparison between two random sequences; that is, the probability that such a match would occur by chance.

evolutionary distance: the number of observed changes in nucleotides or amino acids between two related sequences.

Hidden Markov Model: a probabilistic model of a sequence **alignment**.

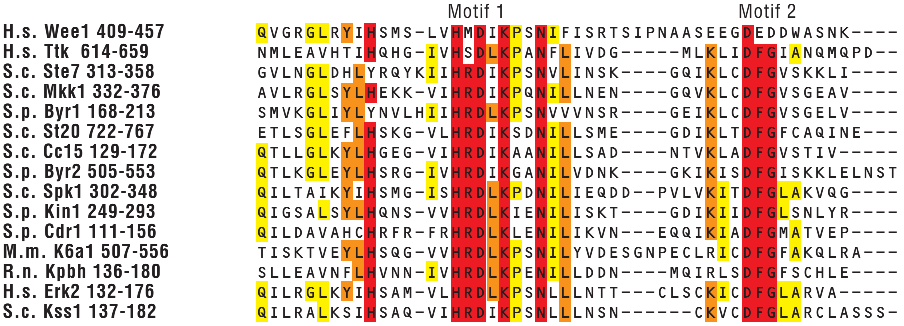
homologous: describes genes or proteins related by divergent evolution from a common ancestor.

homology: the similarity seen between two gene or protein sequences that are both derived by evolution from a common ancestral sequence.

multiple sequence alignment: alignment of more than two sequences to maximize their overall mutual identity or similarity.

pairwise alignment: alignment of two sequences.

percent identity: the percentage of columns in an **alignment** of two sequences that contain identical amino acids. Columns that include gaps are not counted.



```
                                        Motif 1                            Motif 2
H.s. Wee1 409-457    QVGRGLRYIHSMS-LVHMDIKPSNIFISRTSIPNAASEEGDEDDWASNK----
H.s. Ttk  614-659    NMLEAVHTIHQHG-IVHSDLKPANFLIVDG-----MLKLIDFGIANQMQPD--
S.c. Ste7 313-358    GVLNGLDHLYRQYKIIHRDIKPSNVLINSK----GQIKLCDFGVSKKLI----
S.c. Mkk1 332-376    AVLRGLSYLHEKK-VIHRDIKPQNILLNEN----GQVKLCDFGVSGEAV----
S.p. Byr1 168-213    SMVKGLIYLYNVLHIIHRDLKPSNVVVNSR----GEIKLCDFGVSGELV----
S.c. St20 722-767    ETLSGLEFLHSKG-VLHRDIKSDNILLSME----GDIKLTDFGFCAQINE---
S.c. Cc15 129-172    QTLLGLKYLHGEG-VIHRDIKAANILLSAD----NTVKLADFGVSTIV-----
S.p. Byr2 505-553    QTLKGLEYLHSRG-IVHRDIKGANILVDNK----GKIKISDFGISKKLELNST
S.c. Spk1 302-348    QILTAIKYIHSMG-ISHRDLKPDNILIEQDD--PVLVKITDFGLAKVQG----
S.p. Kin1 249-293    QIGSALSYLHQNS-VVHRDLKIENILISKT----GDIKIIDFGLSNLYR----
S.p. Cdr1 111-156    QILDAVAHCHRFR-FRHRDLKLENILIKVN---EQQIKIADFGMATVEP----
M.m. K6a1 507-556    TISKTVEYLHSQG-VVHRDLKPSNILYVDESGNPECLRICDFGFAKQLRA---
R.n. Kpbh 136-180    SLLEAVNFLHVNN-IVHRDLKPENILLDDN----MQIRLSDFGFSCHLE----
H.s. Erk2 132-176    QILRGLKYIHSAM-VLHRDLKPSNLLLNTT---CLSCKICDFGLARVA-----
S.c. Kss1 137-182    QILRALKSIHSAQ-VIHRDIKPSNLLLNSN------CKVCDFGLARCLASSS-
```

Figure 4-4 Multiple alignment A small part of a large multiple alignment of more than 6,000 protein kinase domains in the Pfam database (http://pfam.wustl.edu), displaying part of the region shown in Figure 4-2. Residues identical in all or almost all sequences in the complete alignment are highlighted in red, the next most highly conserved in orange and those next most conserved in yellow. The alignment reveals residues and sequence motifs that are common to all protein kinase catalytic domains and can be used to identify additional members of the family. One is motif 1, which identifies the catalytic loop and contains a conserved aspartic acid (D) important to catalytic function. H.s.: human; S.c.: *Saccharomyces cerevisiae*; S.p.: *Schizosaccharomyces pombe*; M.m.: *Mus musculus*, mouse; R.n.: *Rattus norvegicus*, rat.

Various algorithms have been used to align sequences so as to maximize matching while minimizing gaps. The most powerful is the **Hidden Markov Model**, a statistical model that considers all possible combinations of matches, mismatches and gaps to generate the "best" alignment of two or more sequences. Use of such models provides a third score to go along with percent identity and the similarity score. This score is usually expressed as the probability that the two sequences will have this degree of overall similarity by chance; the lower the score, the more likely the two sequences are to be related. Two virtually identical sequences tend to have probability scores (known in this context as **E-values**) of 10^{-50} or even lower. When the E-value for a sequence comparison is greater than about 10^{-10}, the two sequences could still be related and could have similar structures, but the probability that the two proteins will differ in function increases markedly, especially for multidomain proteins (Figure 4-3).

Multiple alignments and phylogenetic trees

The alignment process can be expanded to give a **multiple sequence alignment**, which compares many sequences (Figure 4-4). Such multiple sequence alignments are arrived at by successively considering all possible pairwise alignments. In effect, one mutates one sequence into all the others to try and determine the most likely evolutionary pathway, given the likelihoods of the various possible substitutions. As more sequences are added to the multiple alignment, such a model becomes "trained" by the evolutionary history of the family of proteins being compared. From this alignment one can see that certain residues are identical in all the sequences. Any residue, or short stretch of sequence, that is identical in all sequences in a given set (such as that of a protein family) is said to be **conserved**. Multiple alignments tend to give a better assessment of similarity than pairwise alignments and can identify distantly related members of a gene family that would not be picked up by pairwise alignments alone.

Multiple sequence alignments of homologous proteins or gene sequences from different species are used to derive a so-called **evolutionary distance** between each pair of species, based, in this instance, on the degree of difference (rather than similarity) between each sequence pair. Given that sequences that diverged earlier in time will be more dissimilar to each other than more recently diverged sequences, these distances can be used to construct **phylogenetic trees** that attempt to reflect evolutionary relationships between species, or, as in the tree illustrated here (Figure 4-5), individual members of a protein superfamily. The tree that emerges, however, will be influenced by the particular tree-building algorithm used and the evolutionary assumptions being made. As the rates of change of protein sequences can vary dramatically, depending on, among other things, the function of the proteins in question and large-scale genomic rearrangements, these specific assumptions are crucial to evaluating the results of phylogenetic analysis.

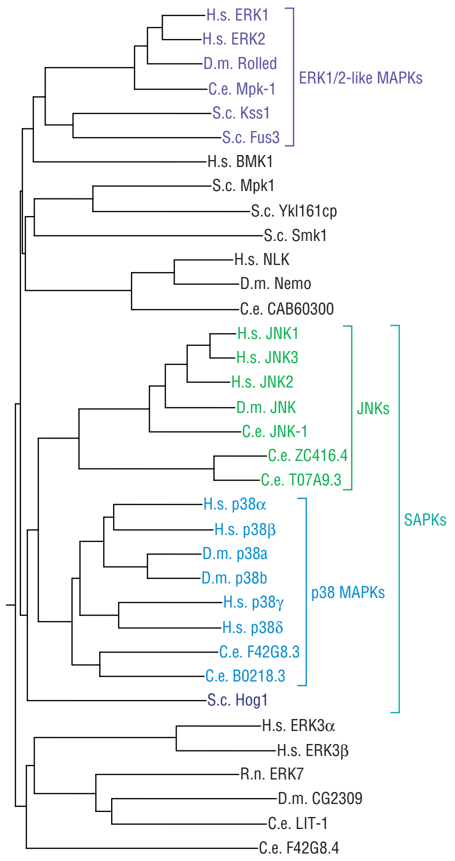


Figure 4-5 Phylogenetic tree comparing the three major MAP kinase subgroups The three major subgroups of MAP kinases (ERKs, JNKs and p38) are well conserved throughout evolution. This dendrogram shows the evolutionary relationships between the ERKs, JNKs and p38 in the budding yeast *S. cerevisiae* (S.c.), the nematode worm *Caenorhabditis elegans* (C.e.), the fruit fly *Drosophila melanogaster* (D.m.) and humans (H.s.). The mammalian MAPK ERK7 was isolated from the rat (R.n.). No human homolog has yet been identified. SAPK stands for stress-activated protein kinases, a general name for the JNK and p38 families. (Kindly provided by James E. Ferrell Jr.)

phylogenetic tree: a branching diagram, usually based on the evolutionary distances between sequences, that illustrates the evolutionary history of a protein family or superfamily, or the relationships between different species of organism.

References

Gerstein, M. and Honig, B.: **Sequences and topology.** *Curr. Opin. Struct. Biol.* 2001, **11**:327–329.

Mount, D.W.: *Bioinformatics: sequence and genome analysis* (Cold Spring Harbor Laboratory Press, New York, 2001).

Wilson, C. *et al.*: **Assessing annotation transfer for genomics: quantifying the relations between protein sequence, structure and function through tradition and probabilistic scores.** *J. Mol. Biol.* 2000, **297**:233–249.

The Pfam database : http://pfam.wustl.edu

Some Examples of Small Functional Protein Domains

Domain	Function
SH2	binds phosphotyrosine
SH3	binds proline-rich sequences
Pleckstrin homology (PH)	binds to G proteins and membranes
WD40	protein–protein interaction
DH	guanine nucleotide exchange
EF-hand	binds calcium
Homeobox	binds DNA
TRBD	binds tRNA
Helix-turn-helix	binds DNA
PUA	RNA modification

Figure 4-6 Representative examples of small functional domains found in proteins These domains are characterized by degenerate sequence motifs that extend over the whole domain. For the structures of some of these domains see Figure 3-2. See Figure 1-46 for an indication of how these domains are combined in proteins.

Structural data can help sequence comparison find related proteins

Some sequence-comparison methods also try to include secondary and tertiary structural information. Because different secondary structural elements can be formed from very similar segments of sequence (see section 4-14), using structural information in the description of the reference protein could, in theory at least, help exclude proteins with somewhat similar sequences but very different structures. It is also known that even similar proteins can have shifts in the relative positions of sequence segments, dictated by differences in secondary-structure packing and the positioning of functionally important groups. This makes the similarity at the sequence level very difficult to determine. For example, there are prokaryotic SH3 domains which, like their eukaryotic relatives, bind to proline-rich sequences. Straightforward sequence alignment does not indicate any relationship between the prokaryotic and eukaryotic domains; however, when the alignment is performed by comparing residues in the corresponding secondary structure elements of the prokaryotic and eukaryotic domains, some regions of sequence conservation appear. A number of small functional domains that can be characterized in this way are listed in Figure 4-6.

Prediction of secondary structure and tertiary structure from sequence alone, by methods such as that of Chou-Fasman and profile-based threading (see sections 1-8 and 4-7), is more accurate when multiple sequences are compared. Both secondary and tertiary structures are determined by the amino-acid sequence; however, there is an interplay between the intrinsic secondary structure propensities of the amino acids and the energetics of the local interactions within a tertiary structure. Tertiary interactions can override a preferred conformation for a residue or segment of residues, and this effect can differ within different local structural contexts. This effect can be taken into account if multiple structures resulting from multiple sequences are available for a *superfamily* of proteins. Therefore, knowledge of the variability of a sequence that can form closely similar structures can improve the performance of prediction methods based on statistical analysis of sequences. Interestingly, all methods for predicting protein structure from sequence seem to have a maximum accuracy of about 70%. The reason for this barrier is unclear.

Sequence and structural motifs and patterns can identify proteins with similar biochemical functions

Sometimes, only a part of a protein sequence can be aligned with that of another protein. Such **local alignments** can identify a functional module within a protein. These function-specific blocks of sequence are called **functional motifs**. There are two broad classes. Short, contiguous motifs usually specify binding sites and can be found within the context of many structures (Figure 4-7). Discontinuous short binding motifs also occur but are often harder to identify by sequence comparisons. Discontinuous or non-contiguous motifs are composed of short stretches of conserved sequence, or even individual conserved residues, separated by stretches

Some Examples of Short Sequence Motifs and Their Functions

Contiguous motif	Consensus sequence	Function
Walker (P loop)	[A/G]XXXXGK[S/T]	binds ATP or GTP
Zn finger	$CX_{2-4}CX_{12}HX_{3-5}H$	binds Zn in a DNA-binding domain
Osteonectin	$CX[D/N]XXCXXG[K/R/H]XCX_{6-7}PXCXCX_{3-5}CP$	binds calcium and collagen
DEAD box helicase	XXDEAD[R/K/E/N]X	ATP-dependent RNA unwinding
MARCKS	GQENGNV[K/R]	substrate for protein kinase C
Calsequestrin	[E/Q][D/E]GL[D/N]FPXYDGXDRV	binds calcium

Figure 4-7 Representative examples of short contiguous binding motifs These motifs are determined by comparison of numerous different versions of the given motif from different proteins. Each motif represents a so-called consensus sequence reflecting the residue most likely to occur at each position. Where two or more residues are equally likely at the same position they are shown in square brackets. X can be any amino acid. The subscript numbers represent repeated residues.

Definitions

BLAST: a family of programs for searching protein and DNA databases for sequence similarities by optimizing a specific similarity measure between the sequences being compared.

functional motif: sequence or structural motif that is always associated with a particular biochemical function.

local alignment: alignment of only a part of a sequence with a part of another.

profile: a table or matrix of information that characterizes a protein family or superfamily. It is typically composed of sequence variation or identity with respect to a reference sequence, expressed as a function of each position in the amino-acid sequence of a protein. It can be generalized to include structural information. Three-dimensional profiles express the three-dimensional structure of a protein as a table which represents the local environment and conformation of each residue.

References

Aitken, A.: **Protein consensus sequence motifs.** *Mol. Biotechnol.* 1999, **12**:241–253.

Altschul, S.F. *et al.*: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Res.* 1997, **25**:3389–3402.

Elofsson, A. *et al.*: **A study of combined structure/sequence profiles.** *Fold. Des.* 1996, **1**:451–461.

Falquet, L. *et al.*: **The PROSITE database, its status in**

of non-conserved sequence. Such discontinuous patterns can also represent catalytic sites; examples are the motifs characterizing the serine proteases and glycosyltransferases. For example, catalases, which are heme-containing enzymes that degrade hydrogen peroxide, can be identified by the discontinuous motifs RXFX**Y**XD[A/S/T][Q/E/H] where the bold Y is the heme iron ligand tyrosine and [I/F]X[R/H]X_4[E/Q]RXX**H**X_2[G/A/S], where the bold H is an essential catalytic histidine. Finally, there are some motifs that extend over the entire sequence of a domain and are highly degenerate. These characterize small protein domains such as SH2 and SH3 (see Figure 4-6). A web server that can be used to find all types of motifs is the PROSITE database, which as of early 2003 contained 1,585 different recognizable motifs.

Protein-family profiles can be generated from multiple alignments of protein families for which representative structures are known

Because functionally important residues must necessarily be conserved over evolution, when multiple sequences from different organisms can be aligned, the probability of recognizing related proteins or a similar biochemical function even at very low overall sequence identity increases dramatically. Specialized computer programs such as PSI-BLAST have been developed for this purpose. This looks for a set of particular sequence features—a **profile**—that characterizes a protein family. Such profiles are obtained from a multiple alignment as described in Figure 4-8. A profile is derived from a position-specific score matrix (PSSM) and this method is used in PSI-BLAST (position-specific iterated **BLAST**). Motif or profile search methods are frequently much more sensitive than pairwise comparison methods (such as ordinary BLAST) at detecting distant relationships. PSI-BLAST may not be as sensitive as the best available dedicated motif-search programs, but its speed and ease of use has brought the power of these methods into more common use.

During evolution, certain positions in a sequence change more rapidly than others. Functionally and structurally important residues tend to be conserved, although the former can change if the specificity or biochemical activity of a protein changes over time; this is how new families branch off from old ones, building up a large superfamily. The concept of a position-based matrix of information to represent a sequence can be generalized to include structural information, which changes more slowly, as well as sequence similarity. This information is used to refine PSSMs such as the one shown in Figure 4-8 to provide a more accurate profile. Profile-based comparison methods differ in two major respects from other methods of sequence comparison. First, any number of known sequences can be used to construct the profile, allowing more information to be used to test the target sequence than is possible with pairwise alignment methods. This is done in PSI-BLAST, where the number of sequences grows with each iteration as more distantly related sequences are found, increasing the informational content of the profile on each iteration. The profile can include penalties for insertion or deletion at each position, which enables one to include information derived from the secondary structure and other indicators of tertiary structure such as the pattern of hydrophobicity or even the local environment around each residue in the comparison. Evolutionary information can also be incorporated.

Profile construction allows the identification of sequences that are compatible with a specific tertiary structure even when sequence identity is too low to be detected with statistical significance. This is the theoretical basis for the profile-based threading method of assigning folds to sequences of unknown proteins (see section 4-7). However, if such a match is not found, it is not an indication that the sequence is incompatible with the protein fold or that two sequences do not have the same structure. False negatives are common in profile-based methods.

Constructing a Family Profile

Position	1	2	3	4	5
	C	C	G	T	L
	C	G	H	S	V
	G	C	G	S	L
	C	G	G	T	L
	C	C	G	S	S

Position	1	2	3	4	5
Prob(C)	0.8	0.6	-	-	-
Prob(G)	0.2	0.4	0.8	-	-
Prob(H)	-	-	0.2	-	-
Prob(S)	-	-	-	0.6	0.2
Prob(T)	-	-	-	0.4	-
Prob(L)	-	-	-	-	0.6
Prob(V)	-	-	-	-	0.2

Figure 4-8 Construction of a profile In this simple example, five homologous sequences five residues long are compared and a matrix is constructed that expresses, for each position, the probability of a given amino acid being found at that position in this family, which is simply a fraction representing the frequency of occurrence. This position-specific score matrix (PSSM) represents the "profile" of this sequence family. An unknown sequence can be scanned against this profile to determine the probability that it belongs to the family by multiplying the individual probabilities of each residue in its sequence, selected from the profile, to obtain a total probability. This can be compared to the value generated by scanning random sequences against the same profile, to assess the significance of the value. For example, the sequence CCHTS would have a probability score of 0.8 × 0.6 × 0.2 × 0.4 × 0.2 = 0.0077, comparable to the score of the aligned sequence CGHSV (0.8 × 0.4 × 0.2 × 0.6 × 0.2). The sequence CLHTG would have a score of zero (0.8 × 0.0 × 0.2 × 0.4 × 0.0).

2002. *Nucleic Acids Res.* 2002, **30**:235–238.

Gaucher, E.A. *et al.*: **Predicting functional divergence in protein evolution by site-specific rate shifts.** *Trends Biochem. Sci.* 2002, **27**:315–321.

Gribskov, M. *et al.*: **Profile analysis: detection of distantly related proteins.** *Proc. Natl Acad. Sci. USA* 1987, **84**:4355–4358.

Kawabata, T. *et al.*: **GTOP: a database of protein structures predicted from genome sequences.** *Nucleic Acids Res.* 2002, **30**:294–298.

Kunin, V. *et al.*: **Consistency analysis of similarity between multiple alignments: prediction of protein function and fold structure from analysis of local sequence motifs.** *J. Mol. Biol.* 2001, **307**:939–949.

Jones, D.T. and Swindells, M.B.: **Getting the most from PSI-BLAST.** *Trends Biochem. Sci.* 2002, **27**:161–164.

Pellegrini, M. *et al.*: **Assigning protein functions by comparative genome analysis: protein phylogenetic profiles.** *Proc. Natl Acad. Sci. USA* 1999, **96**:4285–4288.

Snel, B. *et al.*: **The identification of functional modules from the genomic association of genes.** *Proc. Natl Acad. Sci. USA* 2002, **99**:5890–5895.

von Mering, C. *et al.*: **STRING: a database of predicted functional associations between proteins.** *Nucleic Acids Res.* 2003, **31**:258–261.

The PROSITE database:
http://ca.expasy.org/prosite

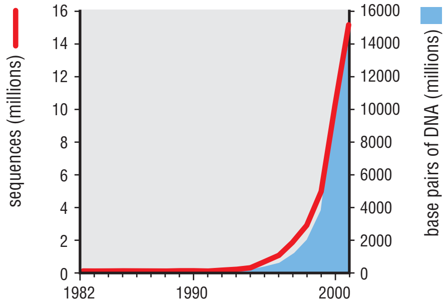


Figure 4-9 The growth of DNA and protein sequence information collected by GenBank over 20 years There has been an exponential increase in both base pairs of DNA sequence and coding sequences, especially since 1994 when various genomics projects were initiated. (Information from http://www.ncbi.nlm.nih.gov/Genbank/genbankstats.html)

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
Mycoplasma genitalium	45.8 x 10^5	483
Methanococcus jannaschii	1.6 x 10^6	1,783
Escherichia coli	4.6 x 10^6	4,377
Pseudomonas aeruginosa	6.3 x 10^6	5,570
Saccharomyces cerevisiae	1.2 x 10^7	6,282
Caenorhabditis elegans	1.0 x 10^8	19,820
Drosophila melanogaster	1.8 x 10^8	13,601
Arabidopsis thaliana	1.2 x 10^8	25,498
Homo sapiens	3.3 x 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://ergo.integratedgenomics.com/GOLD/

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

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

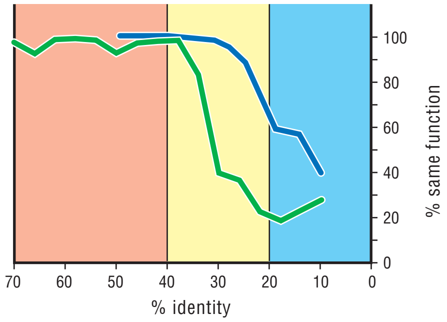
Genomic and protein resources on the Internet:

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

http://ergo.integratedgenomics.com/GOLD/

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 reliably 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 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.

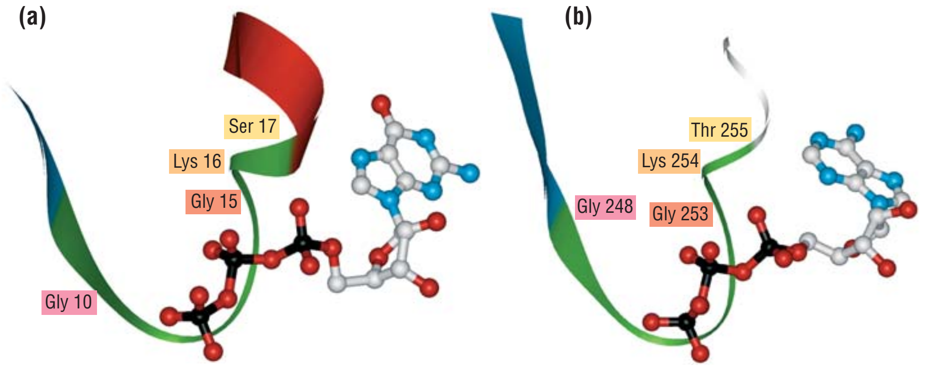


Figure 4-12 The P loop of the Walker motif A contiguous sequence block, the so-called Walker A block or P loop, is a stretch of sequence with a consensus pattern of precisely spaced phosphate-binding residues; this is found in a number of ATP- or GTP-binding proteins, for example ATP synthase, myosin heavy chain, helicases, thymidine kinase, G-protein alpha subunits, GTP-binding elongation factors, and the Ras family. The consensus sequence is: [A or G]XXXXGK[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 1aq2).

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.

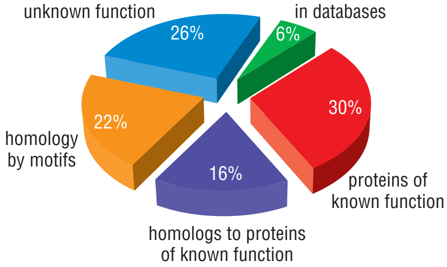


Figure 4-13 Analysis of the functions of the protein-coding sequences in the yeast genome Some are known experimentally, some are surmised from sequence comparison with proteins of known function in other organisms, and some are deduced from motifs that are characteristic of a particular function. Some 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.

4-4 Experimental Tools for Probing Protein Function

Gene function can sometimes be established experimentally without information from protein structure or sequence homology

The explosive growth of sequence information has driven the development of new experimental methods for obtaining information relevant to the function of a gene. Many of these methods are high throughput: they can be applied to large numbers of genes or proteins simultaneously. Consequently, databases of information about the expression level, cellular localization, interacting partners and other aspects of protein behavior are becoming available for entire genomes. Such data are then combined with the results of more classical biochemical and genetic experiments to suggest the function of a gene of interest. The order of experiments is flexible and many will be carried out in parallel. Here we review some of the most common techniques. Most of them require either cloned DNA or protein samples (which sometimes must be purified) for the gene(s) of interest. The rest of the chapter discusses methods, both experimental and computational, that attempt to derive functional information primarily from either protein sequence or protein structure data.

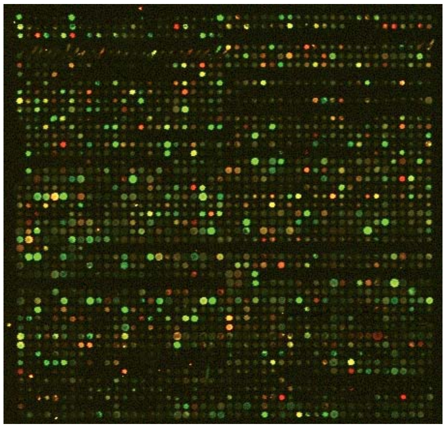


Figure 4-14 DNA microarray Part of a microarray chip showing changes in gene expression when yeast cells are treated with a drug. Genes whose expression increases on drug treatment appear as red spots; those that decrease are green; those that do not change are yellow. Some genes do not appear because they are not expressed under these conditions. Each spot represents a single gene.

One valuable clue to function is the expression pattern of the gene(s) in question. Experience suggests that genes of similar function often display similar patterns of expression: for example, proteins that are involved in chromosome segregation tend to be expressed at the same phase of the cell cycle, while proteins involved in response to oxidative stress usually are expressed—or their expression levels are greatly increased—when cells are subjected to agents that produce oxidative damage (hydrogen peroxide, superoxide, nitric oxide, and so on). Expression can be measured at the level of mRNA or protein; the mRNA-based techniques, such as **DNA microarrays** (Figure 4-14) and SAGE (serial analysis of gene expression), tend to be easier to carry out, especially on a genome-wide scale. Microarray technology, in particular, can provide expression patterns for up to 20,000 genes at a time. It is based on the fidelity of hybridization of two complementary strands of DNA. In its simplest form, the technique employs synthetic "gene chips" that consist of thousands of oligonucleotide spots on a glass slide, one for each gene of the genome. Complementary DNA, labeled with a fluorescent dye, is then made from the mRNA from two different states of the cells being analyzed, one labeled with a red probe (the test state) and the other labeled with a green probe (the reference state). Both are mixed and applied to the chip, where they hybridize to the DNA in the spots. If the level of mRNA for a particular gene is increased in the test cells relative to the reference cells that spot will show up as red; if the level is unchanged the spot will be yellow, and if the mRNA has decreased the spot will be green. In theory, differences of 3–4-fold or greater in mRNA level can be detected reliably with this technique, but in practice the threshold for significance is often 5–10-fold. Any important change in expression must always be verified by **northern blot** analysis.

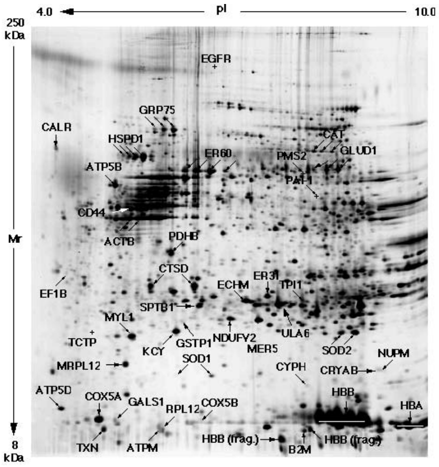


Figure 4-15 2-D protein gel Some spots have been identified and labeled.

Protein expression in the cell can be monitored by antibody binding, but this method is only useful for one protein at a time. High throughput can be achieved by two-dimensional gel electrophoresis, which can separate complex protein mixtures into their components, whose identity can be determined by cutting out the bands and measuring the molecular weight of each protein by mass spectrometry (Figure 4-15). In addition to the amount of protein present, this method can also detect covalent modifications of the protein. The technique is powerful but is also relatively slow and expensive, cannot resolve all the proteins in a cell extract, and can fail to detect proteins that are only present in a few copies per cell. Experience suggests that mRNA levels determined by microarray are good predictors of relative protein levels as determined by two-dimensional gels for the most abundant proteins in a cell; the correlation breaks down for scarcer proteins. Efforts are underway to develop protein microarrays (so-called "protein chips") that can rapidly measure the levels of larger numbers of proteins.

Definitions

DNA microarray: an ordered array of nucleic acid molecules, either cDNA fragments or synthetic oligonucleotides, where each position in the array represents a single gene.

gene knockout: inactivation of the function of a specific gene in a cell or organism, usually by recombination with a marker sequence but sometimes by antisense DNA, RNA interference, or by antibody binding to the gene product. The phenotype resulting from the knockout can often provide clues to the function of the gene.

northern blot: technique for detecting and identifying individual RNAs by hybridization to specific nucleic acid probes, after separation of a complex mixture of mRNAs by electrophoresis and blotting onto a nylon membrane.

RNA interference (RNAi): Abolition of the expression of a gene by a small (~22 base pair) double-stranded RNA.

yeast two-hybrid: a method for finding proteins that interact with another protein, based on activation of a reporter gene in yeast.

References

Colas, P. and Brent, R.: **The impact of two-hybrid and related methods on biotechnology.** *Trends Biotechnol.* 1998, **16**:355–363.

Gasch, A.P. *et al.*: **Genomic expression programs in the response of yeast cells to environmental changes.** *Mol. Biol. Cell* 2000, **11**:4241–4257.

Kallal, L. and Benovic, J.L.: **Using green fluorescent proteins to study G-protein-coupled receptor localization and trafficking.** *Trends Pharmacol. Sci.* 2000, **21**:175–180.

The phenotype produced by inactivating a gene, a **gene knockout**, is highly informative about the cellular pathway(s) in which the gene product operates (Figure 4-16). Knockouts can be obtained by classical mutagenesis, targeted mutations, **RNA interference (RNAi)**, the use of antisense message RNA, or by antibody binding. Microarray analysis on the knockout, comparing the pattern of gene expression in the presence and absence of the gene, will often provide a wealth of information about how the cell responds to its expression, as will studies of changes in protein expression and modification. Of course, the phenotype is an overall response to the loss of the gene product, not a direct readout of biochemical or cellular function. In addition, expressing the gene at high levels in tissues or organisms where it is normally not expressed significantly (ectopic expression) frequently also produces an interesting, and informative, phenotype.

The location of a protein in the cell often provides a valuable clue to its functions. If a gene product is nuclear, cytoplasmic, mitochondrial, or localized to the plasma membrane, for example, and especially if that localization changes in different states of the cell, then inferences about the pathways in which the protein participates can be drawn. A number of techniques exist for determining location, all dependent on attachment of a tag sequence to the gene in question. A commonly used method is to fuse the sequence encoding green fluorescent protein (GFP) to one end of the gene sequence for the protein in question, and then use the intrinsic fluorescence of GFP to monitor where the protein is in the cell (Figure 4-17). Of course, care must be taken that the fusion does not interfere with folding or localization of the gene product.

Many proteins do not function on their own; they are part of a complex of two or more gene products. If the function of one of the interacting proteins is known, then the fact that it binds to a given protein will help reveal the latter's function. Interacting proteins can be found by the **yeast two-hybrid** system. This exploits the fact that transcriptional activators are modular in nature. Two physically distinct domains are necessary to activate transcription: (1) a DNA-binding domain (DBD) that binds to the promoter; and (2) an activation domain that binds to the basal transcription apparatus and activates transcription. In the yeast two-hybrid system, the gene for the target protein is cloned into a "bait" vector next to a sequence encoding the DBD of a given transcription factor. cDNAs encoding potential interactor proteins (the "prey") are cloned separately into another set of plasmids in-frame with the sequence encoding the activation domain of the transcription factor. A bait plasmid and a prey plasmid are introduced together into yeast cells, where the genes they carry are translated into proteins (all combinations of bait and prey are tested in parallel experiments). To form a working transcription factor within the yeast cell, the DBD and the activation domain must be brought together, and this can only happen if the protein carrying the activation domain interacts with the protein fused to the DBD (Figure 4-18). The complete transcription factor can then activate a reporter gene, producing enzyme activity, for example, or cell growth in the absence of a nutrient. Although the two-hybrid screen is a powerful and rapid method for detecting binding partners, it is plagued by false positives and irreproducibility, so any putative interaction must be verified by direct methods such as isolation of the protein complex and identification of its components by antibody binding.

Many other techniques exist and can be employed as needed. Among them are techniques for identifying possible substrates and regulatory molecules. Some of the most popular of these are surface plasmon resonance to detect ligand binding, and purification and direct assay of possible biochemical function *in vitro*. More are being developed as the need for methods to probe function increases. Many of these, like the techniques described here, will produce large databases, so computational analysis of and correlation between such databases will be of great importance for functional genomics.

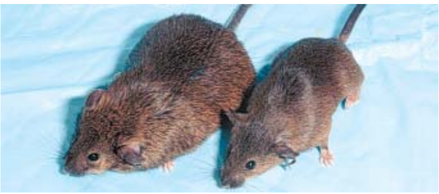


Figure 4-16 The phenotype of a gene knockout can give clues to the role of the gene The mouse on the right is normal; the mouse on the left lacks the gene that encodes pro-opiomelanocortin (POMC), which, among other things, affects the regulation of energy stores and has been linked to obesity. Photograph kindly provided by Ute Hochgeschwender. (From Yaswen, L. *et al.*: **Obesity in the mouse model of pro-opiomelanocortin deficiency responds to peripheral melanocortin.** *Nat. Med.* 1999, **5**:1066–1070.)

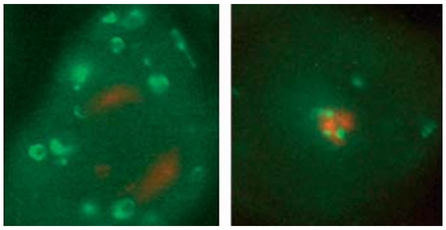


Figure 4-17 Protein localization in the cell The protein has been fused to GFP (green); the nucleus is stained red. In different stages of the cell cycle the protein is either cytoplasmic (left) or localized to the nucleus (right). Photographs kindly provided by Daniel Moore and Terry Orr-Weaver. (From Kerrebrock, A.W. *et al.*: *Cell* 1995, **83**:247–256.)

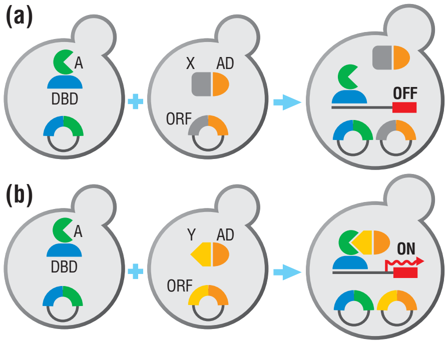


Figure 4-18 Two-hybrid system for finding interacting proteins The "bait" vector expresses a transcription factor DNA-binding domain (DBD, blue) fused to the test protein (protein A, green). The "prey" expression vectors each contain an individual open reading frame (ORF) of interest placed adjacent to the sequence encoding the activation domain (AD) of the same transcription factor (orange). **(a)** When a bait and a prey vector are introduced into a yeast cell, the DBD and its attached protein A binds to the reporter gene (red). If the protein encoded by the ORF (protein X, grey) does not interact with the bait, the reporter gene is not activated. **(b)** If the prey protein (Y, yellow) does interact with protein A, the two parts of the transcription factor are reunited and the reporter gene is expressed.

Kerrebrock, A.W. *et al.*: **Mei-S332, a *Drosophila* protein required for sister-chromatid cohesion, can localize to meiotic centromere regions.** *Cell* 1995, **83**:247–256.

Patterson, S.D.: **Proteomics: the industrialization of protein chemistry.** *Curr. Opin. Biotechnol.* 2000, **11**:413–418.

Phizicky, E.M. and Fields, S.: **Protein–protein interactions: methods for detection and analysis.** *Microbiol. Rev.* 1995, **59**:94–123.

Reymond, M.A. *et al.*: **Standardized characterization of gene expression in human colorectal epithelium by two-dimensional electrophoresis.** *Electrophoresis* 1997, **18**:2842–2848.

Schena, M. *et al.*: **Quantitative monitoring of gene expression patterns with a complementary DNA microarray.** *Science* 1995, **270**:467–470.

Sherlock, G. *et al.*: **The Stanford Microarray Database.** *Nucleic Acids Res.* 2001, **29**:152–155.

A movie of protein transport in a live cell is at:
http://elab.genetics.uiowa.edu/ELABresearch.htm

Evolution has produced a relatively limited number of protein folds and catalytic mechanisms

Although the total number of different enzymatic activities in any living cell is large, they involve a smaller number of classes of chemical transformation (see, for example, section 2-10). For each of these transformations, there is an even smaller number of different catalytic mechanisms by which they can be achieved. This all suggests that most enzymes should be related in both sequence and structure to many others of similar mechanism, even where their substrates are different. Such structural relatedness has indeed been observed: there are only a limited number of protein structural *superfamilies* and the proteins in the same superfamily often share some features of their mechanisms. In practice, however, detecting these structural and functional relationships from sequence alone is fraught with complications.

As described in section 4-1, two proteins with high sequence identity throughout can be assumed to have arisen by **divergent evolution** from a common ancestor and can be predicted to have very similar, if not identical, structures. In general, if the overall identity between the two sequences is greater than about 40% without the need to introduce an inordinate number of gaps in the alignment, and if this identity is spread out over most of the sequence, then the expectation is that they will code for proteins of similar overall fold (Figure 4-19). However, problems in deducing evolutionary relationships and in predicting function from sequence and structure arise when the situation is less clear-cut. And even proteins with greater than 90% sequence identity, which must have very similar structures and active sites, can in rare cases operate on quite different substrates.

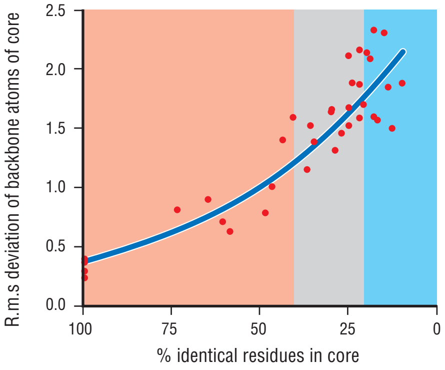


Figure 4-19 Relationship between sequence and structural divergence of proteins The percent identity of the protein cores of 32 pairs of proteins from eight different structural families was plotted against their structural divergence as measured by the root-mean-square difference in spatial positions of backbone atoms. A striking relationship is found, which holds for all the families studied. As the sequences diverge, the structures diverge, but not at the same rate. Small differences in sequence have little effect on structure, but structural divergence increases exponentially as sequence divergence becomes greater. Sequences with greater than 40% identity are generally considered to be homologous and the probability that they will have the same overall structure is also very high. For proteins with sequence identities below about 20%, evolution has usually altered much of the structure, and homology cannot be determined with any certainty. In between is a "grey area", where the overall identity between two sequences is less than about 40% but greater than about 20%, and when it may be impossible from sequence comparisons alone to determine that two proteins are related. Data from Lesk, A.M., *Introduction to Protein Architecture* (Oxford University Press, Oxford, 2001).

Proteins that differ in sequence and structure may have converged to similar active sites, catalytic mechanisms and biochemical function

The structure of the active site determines the biochemical function of an enzyme, and in many homologous proteins active-site residues and structure are conserved even when the rest of the sequence has diverged almost beyond recognition. One might therefore suppose that all proteins with similar active sites and catalytic mechanisms would be homologs. This is, however, not the case. If two such proteins have quite distinct protein folds as well as low sequence similarity, it is likely that they are examples of **convergent evolution**: that is, they did not diverge from a common ancestor but instead arose independently and converged on the same active-site configuration as a result of natural selection for a particular biochemical function. Clear examples of convergent evolution are found among the serine proteases and the aminotransferases, which include proteins of quite different structure and fold, but with similar catalytic sites and biochemical function; these are considered in detail later in the chapter (see sections 4-8 and 4-12, respectively).

Proteins with low sequence similarity but very similar overall structure and active sites are likely to be homologous

It can be difficult to discern homology from sequence gazing alone, because sequence changes much more rapidly with evolution than does three-dimensional structure (Figure 4-20). In fact, proteins with no detectable sequence similarity at all, but with the same structures and biochemical functions, have been found. Among numerous examples are the glycosyltransferases, which transfer a monosaccharide from an activated sugar donor to a saccharide, protein, lipid, DNA or small-molecule acceptor. Some glycosyltransferases that operate on different

Definitions

convergent evolution: evolution of structures not related by ancestry to a common function that is reflected in a common structure.

divergent evolution: evolution from a common ancestor.

References

Bullock, T.L. *et al*.: **The 1.6 angstroms resolution crystal structure of nuclear transport factor 2 (ntf2).** *J. Mol. Biol.* 1996, **260**:422–431.

Hasson, M.S. *et al*.: **The crystal structure of benzoylformate decarboxylase at 1.6Å resolution: diversity of catalytic residues in thiamine diphosphate dependent enzymes.** *Biochemistry* 1998, **37**:9918–9930.

Irving, J.A. *et al*.: **Protein structural alignments and functional genomics.** *Proteins* 2001, **42**:378–382.

Kim, S.W. *et al*.: **High-resolution crystal structures of delta5-3-ketosteroid isomerase with and without a reaction intermediate analogue.** *Biochemistry* 1997, **36**:14030–14036.

Lesk, A.M.: *Introduction to Protein Architecture* (Oxford University Press, Oxford, 2001).

Lundqvist, T. *et al*.: **Crystal structure of scytalone dehydratase—a disease determinant of the rice pathogen, *Magnaporthe grisea*.** *Structure* 1994, **2**:937–944.

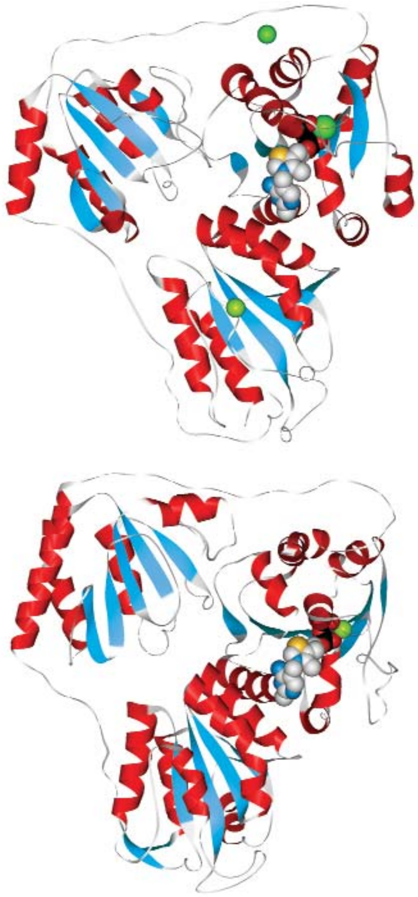


Figure 4-20 **Ribbon diagram of the structure of a monomer of benzoylformate decarboxylase (BFD) and pyruvate decarboxylase (PDC)** BFD (top) and PDC (bottom) share a common fold and overall biochemical function, but they recognize different substrates and have low (21%) sequence identity. The bound thiamine pyrophosphate cofactor is shown in space-filling representation in both structures. The green spheres are metal ions. (PDB 1bfd and 1pvd)

substrates and show no significant sequence identity nevertheless contain a structurally very similar catalytic domain and are thought to have a common ancestor.

In some—probably most—cases, low sequence homology combined with high structural similarity reflects selective conservation of functionally important residues in genuinely homologous, but highly diverged, sequences. Mandelate racemase, muconate lactonizing enzyme and enolase display very little overall sequence identity but have similar structures and active sites (see section 4-11). The reactions they catalyze share a core step and this step is catalyzed in the same way by all three enzymes, implying that they have probably diverged from a common ancestor.

Convergent and divergent evolution are sometimes difficult to distinguish

In other cases, however, there is spatial equivalence at the functional site, but little or no sequence conservation of the functionally important residues. In such cases, distinguishing between convergent and divergent evolution may be difficult. For example, the enzymes benzoylformate decarboxylase (BFD) and pyruvate decarboxylase (PDC) have only about 21% overall sequence identity but have essentially identical folds (Figure 4-20). The catalytic amino-acid side chains are conserved in spatial position in the three-dimensional structure but not in the sequence. It is possible that the two proteins evolved independently and converged to the same chemical solution to the problem of decarboxylating an alpha-ketoacid. But their great similarity in overall structure would seem to indicate that they diverged from a common ancestor. The level of sequence identity between them is, however, too low to distinguish between these two possibilities with confidence.

Divergent evolution can produce proteins with sequence and structural similarity but different functions

Conversely, there are proteins with very different biochemical functions but which nevertheless have very similar three-dimensional structures and enough sequence identity to imply homology. Such cases suggest that structure also diverges more slowly than function during evolution. For example, steroid-delta-isomerase, nuclear transport factor-2 and scytalone dehydratase share many structural details (Figure 4-21) and are considered homologous, yet the two enzymes—the isomerase and the dehydratase—have no catalytically essential residue in common. This suggests that it is general features of the active-site cavity of this enzyme scaffold that have the potential ability to catalyze different chemical reactions that proceed via a common enolate intermediate, given different active-site residues. The third protein in this homologous set—nuclear transport factor-2—is not an enzyme at all, as far as is known, but its active-site-like cavity contains residues that are present in the catalytic sites of both enzymes. Thus, determination of function from sequence and structure is complicated by the fact that proteins of similar structure may not have the same function even when evolutionarily related.

Murzin, A.G.: **How far divergent evolution goes in proteins.** *Curr. Opin. Struct. Biol.* 1998, **8**:380–387.

Patthy, L.: *Protein Evolution* (Blackwell Science, Oxford, 1999).

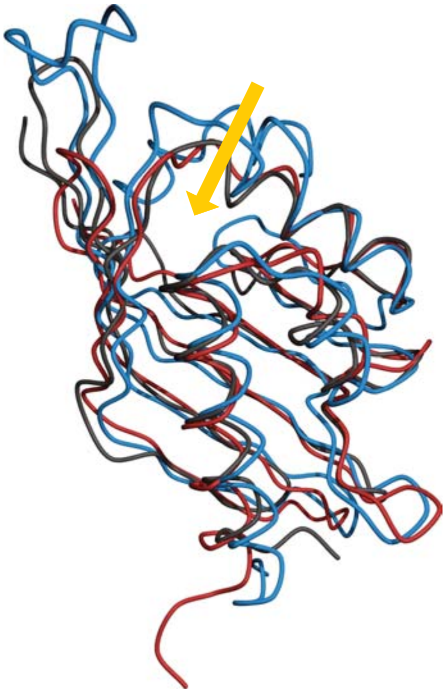


Figure 4-21 **Superposition of the three-dimensional structures of steroid-delta-isomerase, nuclear transport factor-2 and scytalone dehydratase** The active site is indicated by an arrow. (PDB 8cho, 1oun, 1std)

4-6 Structure from Sequence: Homology Modeling

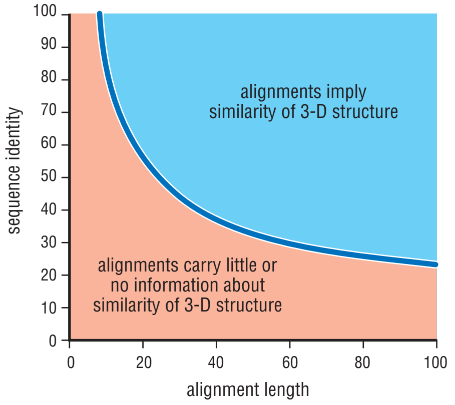


Figure 4-22 The threshold for structural homology Sequence space, plotted as a function of length of the segment being aligned and the percent identity between the two sequences, can be divided into two regions. The upper region (above the curve) shows where sequence similarity is likely to yield enough structural similarity for homology modeling to work. The lower region is highly problematic. At present 25% of known protein sequences fall in the safe area, implying 25% of all sequences can be modeled reliably.

Structure can be derived from sequence by reference to known protein folds and protein structures

Because structure changes more slowly than sequence, if there is a high degree of sequence identity between two proteins, their overall folds will always be similar. But at sequence identity of less than around 40% (see Figure 4-19), structures can be markedly different from each other. In practice, however, structural similarity often extends to lower levels of sequence identity, depending on how the identical residues are distributed. And there are many cases of two proteins having virtually identical overall folds and closely related functions despite having no statistically significant degree of sequence identity/similarity. The real problem in deducing structure from sequence is how to treat these difficult cases.

There are at present about 20,000 entries in the Protein Data Bank representing, depending on how one classifies them, 1,000–2,000 distinct structural "domains", that is unique folds. It has been estimated that the total number of unique folds will be at most several thousand. One of the major goals of work in structural genomics is to determine structures representative of all unique folds so that the structure of any unknown sequence can be modeled. Currently, the known protein structures and canonical protein folds are used to derive structure from sequence by two quite different approaches. The first is described here; the second is the subject of the next section.

Homology modeling is used to deduce the structure of a sequence with reference to the structure of a close homolog

The technique of **homology modeling** aims to produce a reasonable approximation to the structure of an unknown protein by comparison with the structure of a known sequence homolog (a protein related to it by divergent evolution from a common ancestor). Structures that have diverged too far from each other cannot be modeled reliably; the arrangements in space of their secondary structure elements tend to shift too much. In practice, a sequence with greater than about 40% amino-acid identity with its homolog, and with no large insertions or deletions having to be made in order to align them (Figure 4-22), can usually produce a predicted structure equivalent to that of a medium-resolution experimentally solved structure.

Higher-resolution models can be obtained, in principle at least, when there are a number of aligned sequences. To exploit such information better, a technique was developed that uses evolutionary data for a protein family to measure statistical interactions between amino-acid positions. The technique is based on two hypotheses that derive from empirical observation of

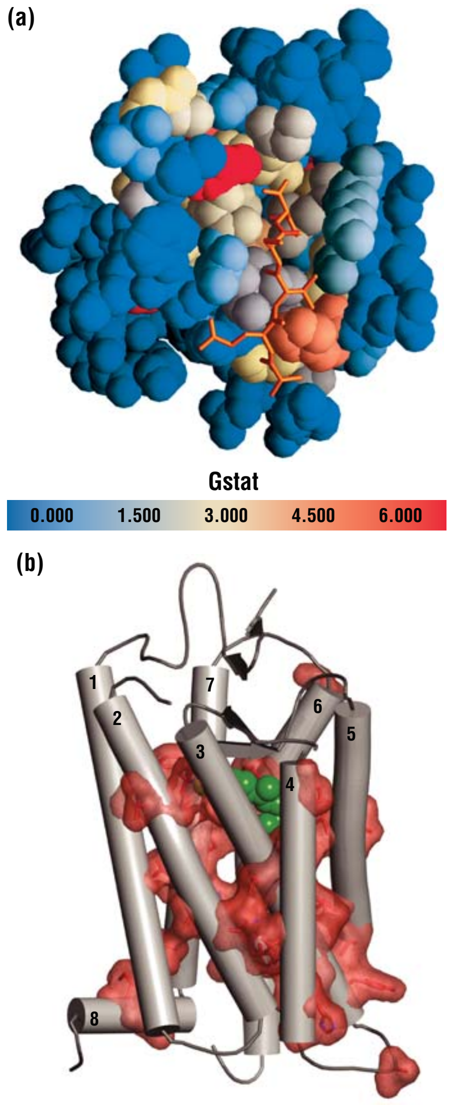


Figure 4-23 Evolutionary conservation and interactions between residues in the protein-interaction domain PDZ and in rhodopsin (a) Highly conserved regions of the PDZ domain were determined using a representative known structure plus information from a structure-based multiple alignment of 274 PDZ-domain sequences, which show a low degree of sequence similarity. This analysis shows that the peptide-binding groove is the most conserved portion of this protein family. Evolutionary conservation is measured by Gstat, a statistical "energy" function: the larger the value of Gstat for a position, the more highly conserved the position is. These data are plotted onto the three-dimensional structure to show the protein interaction surface of the fold, which has a co-crystallized peptide ligand (orange wire model). The high Gstat values for the residues in the groove are consistent with the intuitive expectation that functionally important sites on a protein tend to have a higher than average degree of conservation. **(b)** The structure of the integral membrane protein rhodopsin with the cluster of conserved interacting residues shown in red surrounded by brown van der Waals spheres. This connected network of coevolving residues connects the ligand-binding pocket (green) with known protein-binding regions through a few residues mediating packing interactions between the transmembrane helices. Graphics kindly provided by Rama Ranganathan.

Definitions

homology modeling: a computational method for modeling the structure of a protein based on its sequence similarity to one or more other proteins of known structure.

References

Al-Lazikani, B. *et al.*: **Protein structure prediction.** *Curr. Opin. Chem. Biol.* 2001, **5**:51–56.

Baker, D. and Sali, A.: **Protein structure prediction and structural genomics.** *Science* 2001, **294**:93–96.

Cardozo, T. *et al.*: **Estimating local backbone structural deviation in homology models.** *Comput. Chem.* 2000, **24**:13–31.

Cline, M. *et al.*: **Predicting reliable regions in protein sequence alignments.** *Bioinformatics* 2002, **18**:306–314.

Fetrow, J.S. *et al.*: **Genomic-scale comparison of sequence- and structure-based methods of func-**

sequence evolution. First, a lack of evolutionary constraint at one position should cause the distribution of observed amino acids at that position in the multiple sequence alignment to approach their mean abundance in all proteins, and deviances from the mean values should quantitatively represent conservation. Second, the functional coupling of two positions, even if distantly located in the structure, should mutually constrain evolution at the two positions, and this should be represented in the statistical coupling of the underlying amino-acid distributions in the multiple sequence alignment, which can then be mapped onto the protein (Figure 4-23a). For rhodopsin and for the PDZ domain family, this analysis predicted a set of coupled positions for binding-site residues (shown in red on the figure) that includes unexpected long-range interactions (Figure 4-23b). Mutational studies confirmed these predictions, demonstrating that Gstat, the statistical energy function reflecting conservation, is a good indicator of coupling in proteins. When this technique is used in combination with homology modeling, it can indicate which residues are most likely to remain in conserved positions, even at low levels of sequence identity, and it can also suggest mutagenesis experiments to verify modeled interactions.

What can be done with such models? In some cases they have proven accurate enough to be of value in structure-based drug design. They can be used to predict which amino acids may be in the catalytic site or molecular recognition site if those sites are in the same place in the modeled and experimentally determined protein structures, but they cannot be used to find new binding sites that have been added by evolution. At present, there is no well established way to interrogate an experimentally determined structure, much less a purely modeled structure, and locate such sites from first principles (although some promising new methods are described in section 4-9). Homology models cannot be used to study conformational changes induced by ligand binding, pH changes, or post-translational modification, or the structural consequences of sequence insertions and deletions. At present, computational tools to generate such changes from a starting model are not reliable.

A striking example of the limitations of homology modeling is shown by comparison of the experimentally determined crystal structures of the catalytic domains of the serine protease precursors chymotrypsinogen, trypsinogen, and plasminogen. These protein family members share a high degree of overall sequence identity (over 40%), and an attempt to model the structure of plasminogen from the structure of either of the other two should produce the correct fold. A distinctive difference between plasminogen and the other two zymogens is a complete lack of activity, whereas each of the other two precursors has some activity. This observation cannot be explained from a homology model: the arrangements of residues in the catalytic site will be similar to those of the model template. This is a fundamental limitation of homology modeling: the model is biased toward the structure of the template even in detail. The crystal structure of plasminogen shows that its inactivity is due to blockage of the substrate-binding pocket by a tryptophan residue which is conserved in the sequences of all family members but whose spatial position is different in plasminogen as a result of sequence differences elsewhere in the structure (Figure 4-24).

Homology models also usually cannot be docked together to produce good structures of protein–protein complexes; not only are the docking algorithms unreliable, but the likelihood of significant conformational changes when proteins associate makes it impossible to know whether one is docking the right structures. The same considerations mean that, unless the two homologs have the same oligomeric states, it will not be possible to predict the quaternary structure of a protein from sequence. In short, many, if not most, of the things that biologists want to do with a protein structure cannot be done with confidence using homology models alone. However, even an imperfect homology model may be of use as a guide to planning and interpreting experiments—for example, which amino acid to mutate.

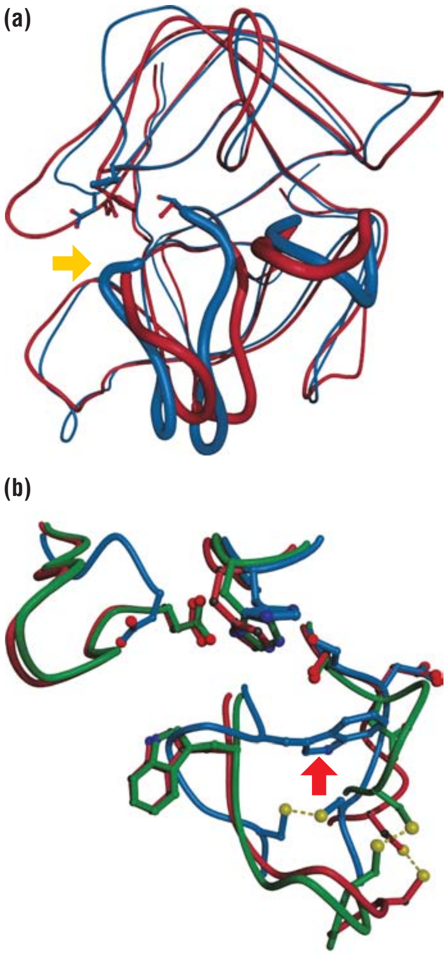


Figure 4-24 Structural changes in closely related proteins (a) The structures of plasminogen (blue) and chymotrypsinogen (red) are very similar, as befits their high sequence identity. Yet the small differences in the positions of loops have important functional consequences, as seen in (b). **(b)** Although chymotrypsinogen (red), chymotrypsin (green) and plasminogen (blue) have about the same degree of sequence identity to one another, the active sites of chymotrypsinogen and chymotrypsin differ from that of plasminogen, where a change in the conformation of the loop indicated by the yellow arrow in (a) has caused a tryptophan residue (Trp 761, red arrow), conserved in both sequences, to adopt a different conformation, where it blocks the substrate-binding pocket. (PDB 2cga, 1ab9 and 1qrz)

tion prediction: does structure provide additional insight? *Protein Sci.* 2001, **10**:1005–1014.

Irving, J.A. *et al.*: **Protein structural alignments and functional genomics.** *Proteins* 2001, **42**:378–382.

Lockless, S.W. and Ranganathan, R.: **Evolutionarily conserved pathways of energetic connectivity in protein families.** *Science* 1999, **286**:295–299.

Marti-Renom, M.A., *et al.*: **Comparative protein structure modeling of genes and genomes.** *Ann. Rev. Biophys.* 2000, **29**:291–325.

Peisach, E. *et al.*: **Crystal structure of the proenzyme domain of plasminogen.** *Biochemistry* 1999, **38**:11180–11188.

Yang, A.S. and Honig, B.: **An integrated approach to the analysis and modeling of protein sequences and structures I-III.** *J. Mol. Biol.* 2000, **301**:665–711.

Protein Data Bank website:
http://www.rcsb.org/pdb/index.html

Profile-based threading tries to predict the structure of a sequence even if no sequence homologs are known

The most important method that has been developed so far for the identification of a protein fold from sequence information alone in the absence of any apparent sequence identity to any other protein, is the method of "profile-based threading". In this method, a computer program forces the sequence to adopt every known protein fold in turn, and in each case a scoring function is calculated that measures the suitability of the sequence for that particular fold (Figure 4-25).

The function provides a quantitative measure of how well the sequence fits the fold. The method is based on the assumption that three-dimensional structures of proteins have characteristics that are at least semi-quantitatively predictable and that reflect the physical-chemical properties of strings of amino acids in sequences as well as limitations on the types of interactions allowed within a folded polypeptide chain. Does, for example, forcing the sequence to adopt particular secondary structures and intra-protein interactions place hydrophobic residues on the inside and helix-forming residues in helical segments? If so, the score will be relatively high.

Experience with profile-based threading has shown that a high score, indicating a good fit to a particular fold, can always be trusted. On the other hand, a low score only indicates that a fit was not found; it does not necessarily indicate that the sequence cannot adopt that fold. Thus, if the method fails to find any fold with a significantly high score, nothing has been learned about the sequence. Despite this limitation, profile-based threading is a powerful method that has been able to identify the general fold for many sequences. It cannot provide fine details of the structure, however, because at such low levels of sequence identity to the reference fold the local interactions and side-chain conformations will not necessarily be the same.

The Rosetta method attempts to predict protein structure from sequence without the aid of a homologous sequence or structure

Ideally, one would like to be able to compute the correct structure for any protein from sequence information alone, even in the absence of homology. Ongoing efforts to achieve this "holy grail" of structure prediction have met with mixed success. Periodically these methods are tested against proteins of known but unpublished structures in a formal competition called CASP (critical assessment of techniques for protein structure prediction). Perhaps the most promising at the moment is the Rosetta method. One of the fundamental assumptions underlying Rosetta is that the distribution of conformations sampled for a given short segment of the sequence is reasonably well approximated by the distribution of structures adopted by that sequence and closely related sequences in known protein structures. Fragment libraries for short segments of the chain are extracted from the protein structure database. At no point is knowledge of the overall native structure used to select fragments or fix segments of the structure. The conformational space defined by these fragments is then searched using a Monte Carlo procedure with an energy function that favors compact structures with paired strands and buried hydrophobic residues. A total of 1,000 independent simulations are carried out for each query sequence, and the resulting structures are clustered. One selection method was simply to choose the centers of the largest clusters as the highest-confidence models. These cluster centers are then rank-ordered according to the size of the clusters they represent, with the cluster centers representing the largest clusters being designated as the highest-confidence models. Before clustering, most structures produced by Rosetta are incorrect (that is, good

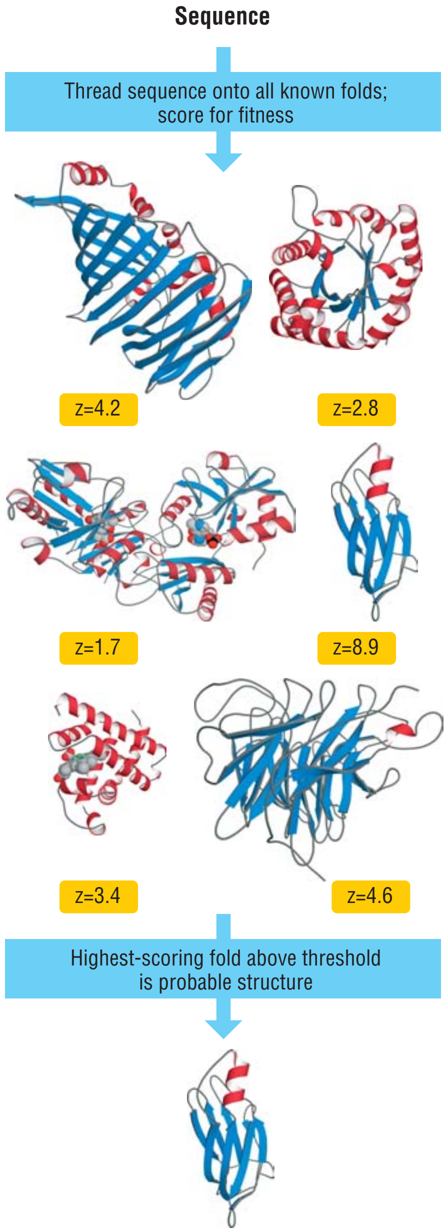

Sequence
Thread sequence onto all known folds; score for fitness
z=4.2
z=2.8
z=1.7
z=8.9
z=3.4
z=4.6
Highest-scoring fold above threshold is probable structure

Figure 4-25 The method of profile-based threading A sequence of unknown structure is forced to adopt all known protein domain folds, and scored for its suitability for each fold. The z-value relates the score for the query sequence to the average score for a set of random sequences with the same amino-acid composition and sequence length. A very high z-score indicates that the sequence almost certainly adopts that fold. Sequences can be submitted online for threading by PSIPRED (http://bioinf.cs.ucl.ac.uk/psipred/index.html).

References

Bonneau, R. *et al.*: **Rosetta in CASP4: Progress in *ab initio* protein structure prediction.** *Proteins* 2001, **45(S5)**:119–126.

Bowie, J.U. *et al.*: **A method to identify protein sequences that fold into a known three-dimensional structure.** *Science* 1991, **253**:164–170.

de la Cruz, X. and Thornton, J.M.: **Factors limiting the performance of prediction-based fold recognition methods.** *Protein Sci.* 1999, **8**:750–759.

Fischer, D. and Eisenberg, D.: **Protein fold recognition using sequence-derived predictions.** *Protein Sci.* 1996, **5**:947–955.

Miller, R.T. *et al.*: **Protein fold recognition by sequence threading: tools and assessment techniques.** *FASEB J.* 1996, **10**:171–178.

Simons, K.T. *et al.*: **Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions.** *J. Mol. Biol.* 1997, **268**:209–225.

structures account for less than 10% of the conformations produced); for this reason, most conformations generated by Rosetta are referred to as decoys (Figure 4-26). The problem of discriminating between good and bad decoys in Rosetta populations is still under investigation. Still, in some test calculations, the best cluster center has been shown to agree fairly well with the overall fold of the protein (Figure 4-27).

Both the Rosetta method and the method of profile-based threading suffer from some of the same limitations that beset homology modeling. The issue of false positives and negatives is significant, because the failure to generate a model does not mean one cannot be generated, nor that the structure is a novel one. And the generation of a model does not mean it is right, either overall or, more usually, in detail. At best one should look to these methods, at least for the present, for rough indication of fold class and secondary structure topology. And it is important to remember that all methods of model building based on a preexisting structure, whether found by sequence homology or by threading, suffer from massive feedback and bias. The structure obtained will always look like the input structure, because the computational tools for refining the model are unable to generate the kinds of shifts in secondary structure position and local tertiary structure conformations that are likely to exist between two proteins when their overall sequence identity is low (see Figure 4-19). *Ab initio* methods like Rosetta at least do not suffer from this problem, whatever their other limitations.

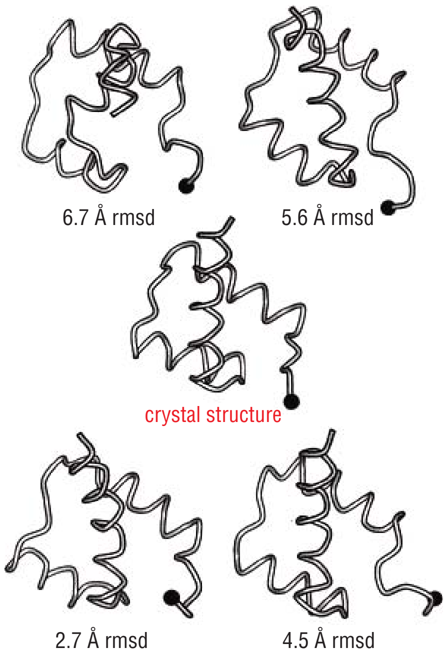


Figure 4-26 Some decoy structures produced by the Rosetta method The structure at the center is the target, the experimentally determined structure of a homeodomain. The other structures are generated by the Monte Carlo approach in Rosetta, using only the sequence of the protein. Although some of the structures are quite far from the true structure, others are close enough for the fold to be recognizable. Rmsd is the root mean square deviation in α-carbon positions between the computed structure and the experimentally determined structure. (Taken from Simons, K.T. *et al.*: *J. Mol. Biol.* 1997, **268**:209–225.)

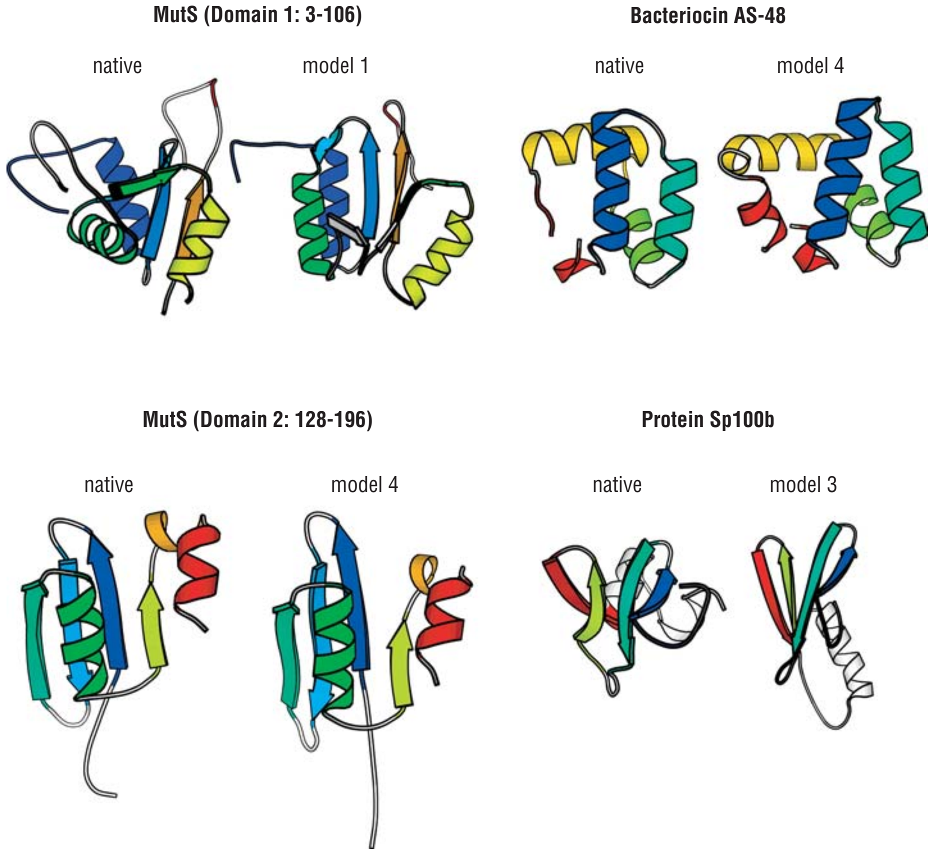


Figure 4-27 Examples of the best-center cluster found by Rosetta for a number of different test proteins The level of agreement with the known native structure varies, but in many cases the overall fold is predicted well enough to be recognizable. Note, however, that the relative positions of the secondary structure elements are almost always shifted at least somewhat from their true values. Graphics kindly provided by Richard Bonneau and David Baker. (Adapted from Bonneau, R. *et al.*: *Proteins* 2001, **45(S5)**:119–126.)

Simons, K.T. *et al.*: **Prospects for *ab initio* protein structural genomics.** *J. Mol. Biol.* 2001, **306**: 1191–1199.

URL for threading website:
http://bioinf.cs.ucl.ac.uk/psipred/index.html

URL for CASP:
http://moult.carb.nist.gov/casp

4-8 Deducing Function from Structure: Protein Superfamilies

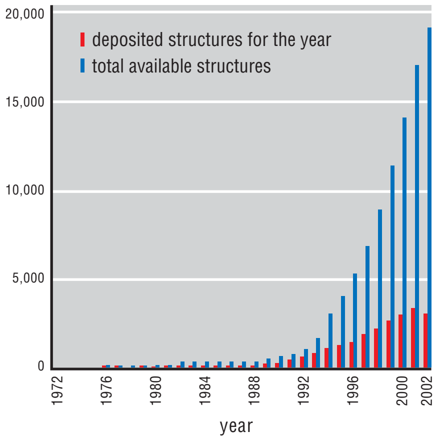


Figure 4-28 Growth in the number of structures in the protein data bank Both yearly and cumulative growth is shown. (Taken from the Protein Data Bank website: http://www.rcsb.org/pdb/holdings.html)

Members of a structural superfamily often have related biochemical functions

In contrast to the exponential increase in sequence information, structural information, which is still chiefly obtained by X-ray crystallography and NMR, has up to now been increasing at a much lower rate (Figure 4-28). One goal of the structural genomics initiatives that have been implemented since the release of the first complete genome sequence is to increase the rate at which experimentally derived structures of the gene products are produced. The driving force behind these initiatives is the assumption that, in addition to defining the ensemble of all possible protein folds, comprehensive structural information could provide a firmer basis than sequence for functional predictions, as three-dimensional structure changes much more slowly than sequence during evolution. A good reason for optimism that these assumptions will hold true is the existence of **superfamilies** of proteins with related structures and biochemical functions.

A superfamily is loosely defined as a set of homologous proteins with similar three-dimensional structures and related, though not necessarily identical, biochemical functions. Almost all superfamilies exhibit some functional diversity, which is generated by local sequence variation and/or domain shuffling. Within enzyme superfamilies, for example, substrate diversity is common, while parts of the reaction chemistry are highly conserved. In many enzyme superfamilies, the sequence positions of catalytic residues vary from member to member, despite the fact that they have equivalent functional roles in the proteins. These variations may make the assignment of a protein to a superfamily from sequence comparison alone problematic or impossible. Although some superfamily members may be similar in sequence, it is the structural and functional relationships that place a protein in a particular superfamily. Within each superfamily, there are **families** with more closely related functions and significant (>50%) sequence identity.

Because the total number of protein folds and the total number of biochemical functions is smaller than the total number of genes in biology, if a protein can be assigned to a superfamily from sequence or structural information, at the very least the number of its possible functions can be narrowed down, and in some instances it may be possible to assign a function precisely.

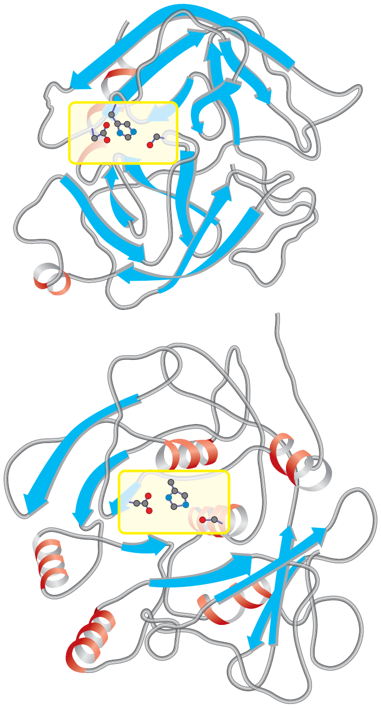


Figure 4-29 The overall folds of two members of different superfamilies of serine proteases The enzymes are chymotrypsin (top) and subtilisin (bottom). The residues in the catalytic triad are indicated for each.

The four superfamilies of serine proteases are examples of convergent evolution

Striking examples of similarities in biochemical function but quite different biological roles come from enzymes where the chemical reactions are the same but the substrates can differ considerably. There are, for instance, many hundreds of enzymes that hydrolyze peptide bonds in protein and polypeptide substrates, but they can be grouped into a small number of classes, each with its own characteristic chemical mechanism. The most numerous class comprises the serine proteases, in which the side-chain hydroxyl group of a serine residue in the active site attacks the carbonyl carbon atom of the amide bond that is to be hydrolyzed. Two other characteristic residues, a histidine and an aspartic acid (or a glutamic acid), are involved in assisting this hydrolysis, forming a catalytic triad. Serine proteases fall into several structural superfamilies, which are recognizable from their amino-acid sequences and the particular disposition of the three catalytically important residues in the active site (Figure 4-29). Each serine protease superfamily has many members but there is no obvious relationship between the superfamilies, either in sequence or structure. The three residues of the catalytic triad are

Definitions

family: a group of homologous proteins that share a related function. Usually these will also have closely related sequences. Members of the same enzyme family catalyze the same chemical reaction on structurally similar substrates.

superfamily: proteins with the same overall fold but with usually less than 40% sequence identity. The nature of the biochemical functions performed by proteins in the same superfamily are more divergent than those within families. For instance, members of the same enzyme superfamily may not catalyze the same overall reaction, yet still retain a common mechanism for stabilizing chemically similar rate-limiting transition-states and intermediates, and will do so with similar active-site residues.

References

Gerlt, J.A. and Babbitt, P.C.: **Divergent evolution of enzymatic function: mechanistically diverse superfamilies and functionally distinct suprafamilies.** *Annu. Rev. Biochem.* 2001, **70**:209–246.

Krem, M.M. *et al.*: **Sequence determinants of function and evolution in serine proteases.** *Trends Cardiovasc. Med.* 2000, **10**:171–176.

Perona, J.J. and Craik, C.S.: **Evolutionary divergence of substrate specificity within the chymotrypsin-like**

found in a different order in different locations along the sequence in each superfamily: nevertheless, in the tertiary structure they come together in a similar configuration. Presumably, the existence of a similar active site is due to convergent evolution, while within each superfamily, divergent evolution has produced distinct individual proteases with very similar structures but different substrate specificity.

A given serine protease can be highly specific for a particular target amino-acid sequence—although some are relatively nonspecific—so that, in general, the substrate(s) for such a protease cannot be predicted from knowledge of the sequence or even the structure of the enzyme. However, observation of a serine protease fold combined with the right active-site residues is diagnostic for a protease. Serine proteases participate in such diverse cellular functions as blood clotting, tissue remodeling, cell-cycle control, hormone activation and protein turnover. A small number of members of some families within the serine protease superfamily have lost one or more of the catalytic residues and perform non-catalytic functions such as forming a structural matrix.

Another large enzyme superfamily with numerous different biological roles is characterized by the so-called polymerase fold, which resembles an open hand (Figure 4-30). DNA polymerases of all types from all organisms studied so far appear to share this fold, as do RNA polymerases and viral reverse transcriptases. In most cases, sequence comparisons alone do not make this functional distinction, and in some cases no sequence similarity between the families of the polymerase fold superfamily is apparent. However, structural evidence indicates that every enzyme that transcribes or replicates nucleic acid polymers has probably descended from a common ancestor.

There are many other examples of related but differing functions among members of a superfamily. It should be borne in mind, however, that in some proteins with similar function, equivalent active-site residues come from different positions in the sequence, obscuring superfamily membership until structural information is obtained.

Very closely related protein families can have completely different biochemical and biological functions

There are some well-known cases of significant differences in function at very high levels of sequence identity. One example is the crystallins, which appear to have evolved from several different enzymes. Although some crystallins retain more than 50% sequence identity to these enzymes, they function as structural proteins, not enzymes, in the eye lens.

And with increasing numbers of structures being solved for proteins of known function, the functional diversity of many other protein superfamilies has been revealed. Thornton and co-workers have assessed the functional variation within homologous enzyme superfamilies containing two or more enzymes. Combining sequence and structure information to identify relatives, the majority of superfamilies display variation in enzyme function, with 25% of the superfamilies having members of different enzyme types. For example, the α/β hydrolase superfamily has at least four different functions; the ferredoxin superfamily has at least three. For single- and multidomain enzymes, difference in biochemical function is rare above 40% sequence identity, and above 30% the overall reaction type tends to be conserved, although the identity of the substrate may not be. For more distantly related proteins, sharing less than 30% sequence identity, functional variation is significant, and below this threshold, structural data are essential for understanding the molecular basis of observed functional differences.

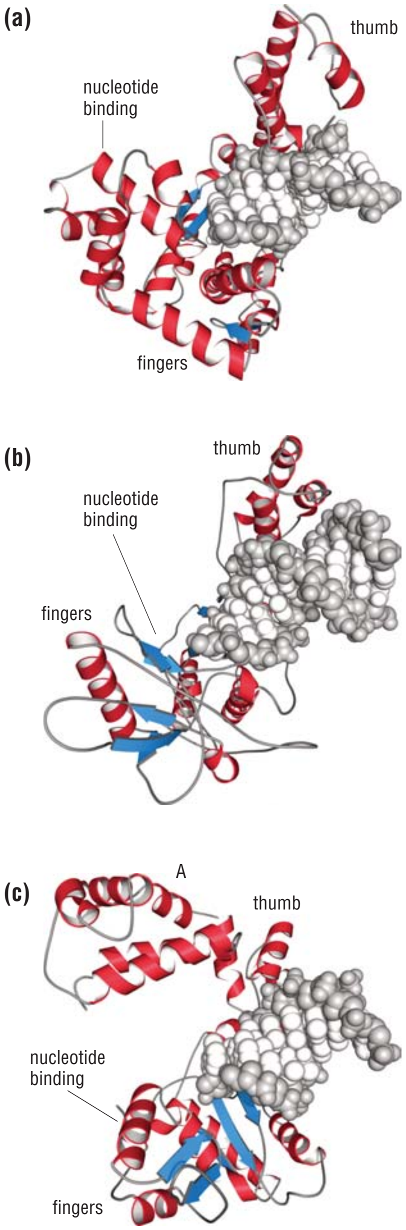


Figure 4-30 A comparison of primer–template DNA bound to three DNA polymerases **(a)** Taq DNA polymerase bound to DNA. The DNA stacks against the "fingers" and is contacted across the minor groove by the "thumb" domain. **(b)** The binary complex of HIV-1 reverse transcriptase and DNA. This structure does not have a nucleotide-binding alpha helix in the fingers domain. Instead, a beta hairpin probably performs this function. **(c)** The ternary complex of rat DNA polymerase β with DNA and deoxy-ATP (not shown). Although this polymerase has an additional domain (A), the "thumb" domain similarly binds the DNA primer–template in the minor groove, while the "fingers" present a nucleotide-binding alpha helix at the primer terminus. (PDB 1tau, 2hmi and 8icp)

serine protease fold. *J. Biol. Chem.* 1997, **272**: 29987–29990.

Siezen, R.J. and Leunissen, J.A.: **Subtilases: the superfamily of subtilisin-like serine proteases.** *Protein Sci.* 1997, **6**:501–523.

Steitz, T.A.: **DNA polymerases: structural diversity and common mechanisms.** *J. Biol. Chem.* 1999, **274**:17395–17398.

Todd, A.E. *et al.*: **Evolution of function in protein superfamilies, from a structural perspective.** *J. Mol. Biol.* 2001, **307**:1113–1143.

Evolution of function in protein superfamilies, from a structural perspective:

http://www.biochem.ucl.ac.uk/bsm/FAM-EC/

4-9 Strategies for Identifying Binding Sites

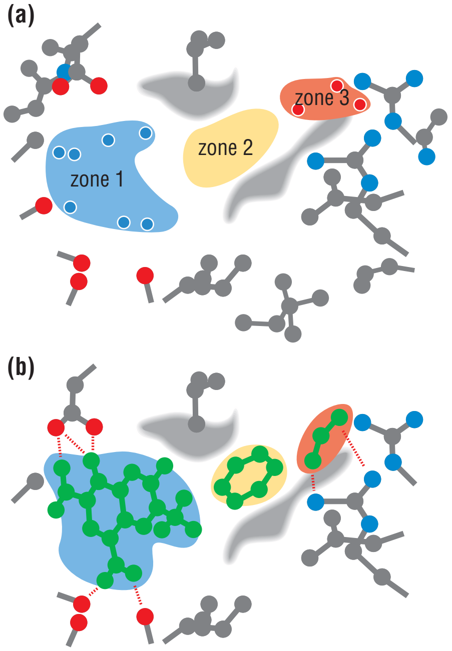


Figure 4-31 Example of the use of GRID Three different types of probes have been used to locate binding sites for functional groups in the active site of the enzyme dihydrofolate reductase. **(a)** Zone 1 (blue) is a good site for binding electrostatically positive groups, with the energy function minima from an amino probe shown in blue dots. It was also identified with a carbon probe as being a good pocket for shape complementarity. Zone 2 (yellow) is a good site for hydrophobic interaction, as illustrated by the hydrophobic molecular surface (grey shapes) in that region. Zone 3 (red) is a good binding site for electrostatically negative groups, with minima from a carboxylate probe shown by the red dots. **(b)** Overlay of three pieces of a known inhibitor of dihydrofolate reductase onto the zones of favorable interaction energy found by GRID. Figures adapted from http://thalassa.ca.sandia.gov/~dcroe/builder.html

Binding sites can sometimes be located in three-dimensional structures by purely computational means

Whether the three-dimensional structure of a new protein is determined experimentally or computationally (see sections 4-6 and 4-7)—there will always be the problem of finding those sites on the protein surface that are involved in its biochemical and cellular functions. This problem is particularly acute when the structure reveals a polypeptide chain fold that has never been seen before, and there is no obvious cofactor or other bound ligand to identify a functional site. But it can also apply to structures of proteins with known folds, as proteins often have more than one function (see section 4-13) and even a familiar protein may have acquired additional functions and functional sites in the context of a different organism.

To some extent, the characteristics of binding sites that were discussed in section 2-4 can be used to identify regions of a protein's surface that are good candidates for functional sites. These characteristics include concavity—binding sites are usually depressions rather than protrusions or flat areas, although many exceptions are known—as well as a higher than average amount of exposed hydrophobic surface area. However, such generalizations usually only narrow down the possibilities. What is needed is a method, ideally a computational method so that it can be used with homology models as well as with experimentally derived structures, that can scan the surface of a protein structure and locate those sites that have evolved to interact with small molecules or with other macromolecules. Some success has been obtained with methods that scan the surface and look for sites of specific shape. Residue conservation analysis can also be quite revealing if there are enough homologous sequences (see Figure 4-23).

Several other computational methods have been proposed and tested on experimentally determined structures. Most use a "probe" molecule and an energy function that describes the interaction of the probe with the residues on the protein surface. Binding sites are identified as regions where the computed interaction energy between the probe and the protein is favorable for binding. Two widely used methods, GRID and MCSS (multiple conformations simultaneous search), use this strategy. In GRID, the interaction of the probe group with the protein structure is computed at sample positions on a lattice throughout and around the macromolecule, giving an array of energy values. The probes, which are usually used singly, include water, the methyl group, the amine NH_2 group, the carboxylate group and the hydroxyl group, among others. Contour surfaces at various energy levels are calculated for each probe for each point on the lattice and displayed by computer graphics together with the protein structure (Figure 4-31). Contours at negative energy levels delineate regions of attraction between probe and protein that could indicate a binding site, as such contours are found at known ligand-binding clefts.

The approach taken by MCSS is similar in principle but differs in detail and can take into account the flexibility of both the probe molecule and the protein. The resulting distribution map of regions on the protein surface where functional groups show a favorable interaction energy can be used for the analysis of protein–ligand interactions and for rational drug design.

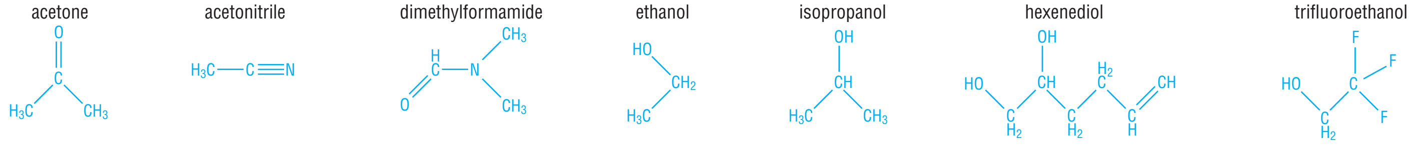


Figure 4-32 Some organic solvents used as probes for binding sites for functional groups

References

Allen, K.N. *et al.*: **An experimental approach to mapping the binding surfaces of crystalline proteins.** *J. Phys. Chem.* 1996, **100**:2605–2611.

Aloy, P. *et al.*: **Automated structure-based prediction of functional sites in proteins: applications to assessing the validity of inheriting protein function from homology in genome annotation and to protein docking.** *J. Mol. Biol.* 2001, **311**:395–408.

Bitetti-Putzer, R. *et al.*: **Functional group placement in protein binding sites: a comparison of GRID and MCSS.** *J. Comput. Aided Mol. Des.* 2001, **15**:935–960.

Byerly, D.W. *et al.*: **Mapping the surface of *Escherichia coli* peptide deformylase by NMR with organic solvents.** *Protein Sci.* 2002, **11**:1850–1853.

Dennis, S. *et al.*: **Computational mapping identifies the binding sites of organic solvents on proteins.** *Proc. Natl Acad. Sci. USA* 2002, **99**:4290–4295.

English, A.C. *et al.*: **Experimental and computational mapping of the binding surface of a crystalline protein.** *Protein Eng.* 2001, **14**:47–59.

Goodford, P.J.: **A computational procedure for determining energetically favorable binding sites on biologically important macromolecules.** *J. Med. Chem.* 1985, **28**:849–857.

Laskowski, R.A. *et al.*: **Protein clefts in molecular recognition and function.** *Protein Sci.* 1996, **5**:2438–2452.

Liang, J. *et al.*: **Anatomy of protein pockets and cavities: measurement of binding site geometry and implications for ligand design.** *Protein Sci.*

Computational methods are, however, extremely inefficient when no information is available to limit the regions of the protein surface to be scanned. Scanning an entire protein surface with either GRID or MCSS yields hundreds of possible binding sites of roughly equivalent energy. Thus, these tools are most useful in cases where the active site is already known and one wants to determine what sorts of chemical groups might bind there.

Experimental means of locating binding sites are at present more accurate than computational methods

MSCS (multiple solvent crystal structures) is a crystallographic technique that identifies energetically favorable binding sites and orientations of small organic molecules on the surface of proteins; this experimental method can find likely functional sites on the surface of any protein that can be crystallized. The method involves soaking protein crystals in an organic solvent that mimics a functional group on a ligand: thus, ethanol will probe for hydroxymethyl-binding sites such as those that interact with a threonine side chain; dimethylformamide identifies binding sites that interact with the C=O and N–H groups of peptides, and so on (Figure 4-32). Determination of the protein structure at resolutions of the order of 2 Å in the presence of the solvent probe reveals the solvent-binding sites (Figure 4-33). If the experiment is repeated with several different probes, it is found that they cluster in only a few binding sites, regardless of their polarity. These sites are the functional sites on the surface of the protein. In contrast to the computational methods, MSCS involves direct competition between the probe and the bound water on the surface of the protein. As it is displacement of this water that drives ligand-binding events (see section 2-4), MSCS finds a much more restricted set of binding sites than do the computational methods.

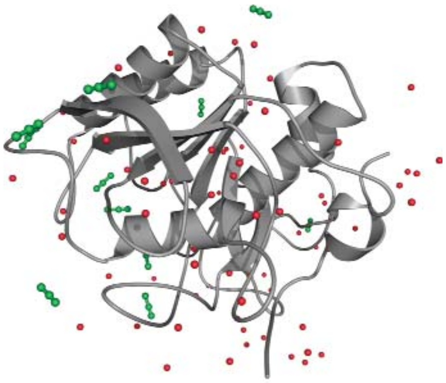


Figure 4-33 Structure of subtilisin in 100% acetonitrile Crystal structure of the serine protease subtilisin in 100% acetontitrile. The organic solvent, shown as green rods, binds at only a few sites on the protein surface, including the active site, which is approximately left of center in the figure. The red spheres are bound water molecules, which are not displaced even by this water-miscible organic solvent at 100% concentration. These bound waters should be considered an integral part of the folded structure of the protein. (PDB 1be6)

High-resolution structures of crystals of the well-studied enzyme thermolysin soaked in acetone, acetonitrile, or phenol show probe molecules clustering in the main specificity pocket of the thermolysin active site and in a buried sub-site, consistent with structures of known protein–ligand complexes of thermolysin (Figure 4-34). When the experimentally determined solvent positions within the active site were compared with predictions from GRID and MCSS, both these computational methods found the same sites but gave fewer details of binding. And both GRID and MCSS predicted many other sites on the protein surface, not observed experimentally, as equally favorable for probe binding.

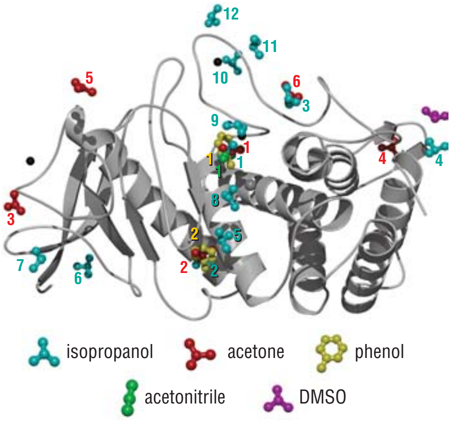


Figure 4-34 Ribbon representation showing the experimentally derived functionality map of thermolysin The binding sites for different organic solvent molecules were obtained by X-ray crystallography of crystals of thermolysin soaked in the solvents. The same probe molecules bound to different positions are numbered to identify their site of binding. The active-site zinc ion and the bound calcium ions are shown as grey and black spheres, respectively. Dimethyl sulfoxide (DMSO, purple) is present in the crystallization conditions of thermolysin; one molecule binds per molecule protein. Graphic kindly provided by Roderick E. Hubbard. (Adapted from English *et al.*: *Protein Eng.* 2001, **14**:47–59.)

Related experimental methods using NMR instead of X-ray crystallography study the binding of small-molecule compounds as well as organic solvents as probes. The experimental methods are accurate but cannot be used on homology models. Computational methods are not accurate enough to discriminate among possible binding sites. What is needed is a computational analog of the experimental methods. One such new computational mapping strategy has been tested recently with promising results. Using eight different ligands for lysozyme and four for thermolysin, the computational search finds the consensus site to which all the ligands bind, whereas positions that bind only some of the ligands are ignored. The consensus sites turn out to be pockets of the enzymes' active sites, lined with partially exposed hydrophobic residues and with some polar residues toward the edge. Known substrates and inhibitors of hen egg-white lysozyme and thermolysin interact with the same side chains identified by the computational mapping, but the computational mapping did not identify the precise hydrogen bonds formed and the unique orientations of the bound substrates and inhibitors.

1998, **7**:1884–1897.

Liepinsh, E. and Otting, G.: **Organic solvents identify specific ligand binding sites on protein surfaces.** *Nat. Biotechnol.* 1997, **15**:264–268.

Mattos, C. and Ringe, D.: **Locating and characterizing binding sites on proteins.** *Nat. Biotechnol.* 1996, **14**:595–599.

Mattos, C. and Ringe, D.: **Proteins in organic solvents.** *Curr. Opin. Struct. Biol.* 2001, **11**:761–764.

Miranker, A. and Karplus, M.: **Functionality maps of binding sites: a multiple copy simultaneous search method.** *Proteins* 1991, **11**:29–34.

Shuker, S.B. *et al.*: **Discovering high-affinity ligands for proteins: SAR by NMR.** *Science* 1996, **274**:1531–1534.

URL for GRID:
http://thalassa.ca.sandia.gov/~dcroe/builder.html

4-10 Strategies for Identifying Catalytic Residues

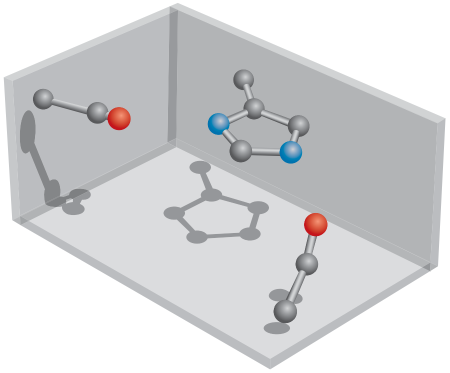


Figure 4-35 **An active-site template** The geometry of the catalytic triad of the serine proteases as used to locate similar sites in other proteins. Adapted from the rigid active-site geometries website: http://www.biochem.ucl.ac.uk/bsm/PROCAT/PROCAT.html.

Site-directed mutagenesis can identify residues involved in binding or catalysis

Locating binding sites, either experimentally or computationally (see section 4-9), does not automatically indicate which residues in those sites are responsible for ligand binding or, in the case of enzymes, catalysis. The standard experimental method for identifying these residues is alanine-scanning mutagenesis, in which candidate amino acids are replaced by alanine by site-directed mutagenesis of the gene. The effect of this side-chain excision on function—usually binding or catalysis—of the expressed mutant protein is then assayed. When combined with genetic assays for the *in vivo* phenotype of the mutated protein and information from the pH/rate profile of a catalytic reaction, for example, such experiments can reveal which side chains in a binding site may actually perform chemistry on a bound ligand.

Active-site residues in a structure can sometimes be recognized computationally by their geometry

But with the advent of genome-wide sequencing and protein-structure determination, a computational tool is needed that can identify such residues rapidly and automatically. This would be particularly useful in cases where the protein in question has no known function and the location of the active site is uncertain, because knowledge of the residues that can carry out chemistry might indicate what type of chemistry the protein actually performs.

The simplest of the computational methods searches the structure for geometrical arrangements of chemically reactive side chains that match those in the active sites of known enzymes. This rigid active-site approach has successfully identified the catalytic triad of a serine protease (Figure 4-35) in an enzyme of unknown function, but has not been used extensively to probe for other functions. Because it relies solely on geometry and not on position in the sequence, this method could find serine protease catalytic sites in any protein fold in which they occur. A more sophisticated variation uses a three-dimensional descriptor of the functional site of interest, termed a "fuzzy functional form", or FFF, to screen the structure. FFFs are based on the geometry, residue identity, and conformation of active sites using data from known crystal structures of members of a functional family and experimental biochemical data. The descriptors are made as general as possible ("fuzzy") while still being specific enough to identify the correct active sites in a database of known structures.

These fuzzy functional descriptors can identify active sites not only in experimentally determined structures, but also from predicted structures provided by *ab initio* folding algorithms (see sections 1-9, 4-7) or threading algorithms (see section 4-7). A disulfide oxidoreductase FFF has been successfully applied to find other disulfide oxidoreductases in a small structural database and, more recently, has been used to scan predicted protein structures derived from the entire *Bacillus subtilis* genome. A total of 21 candidate disulfide oxidoreductases were found, of which six turned out to be false positives. The method did not miss any of the known disulfide oxidoreductases and identified at least two potential new ones.

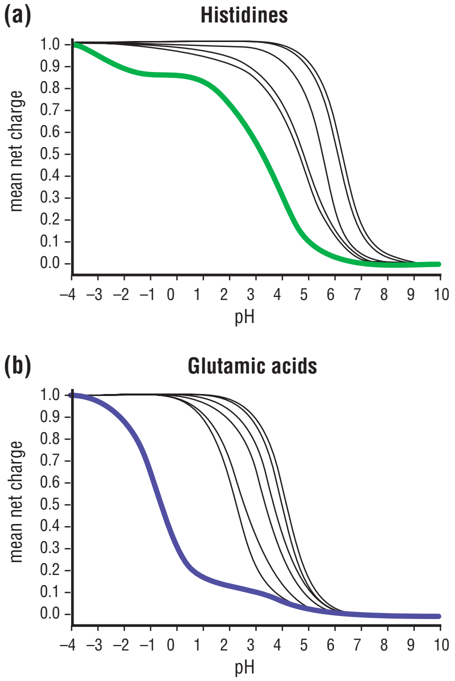


Figure 4-36 **Theoretical microscopic titration curves** Calculated curves for two types of ionizable residues—histidines and glutamic acids—in the structure of the glycolytic enzyme triosephosphate isomerase. In each case, most of the residues behave similarly, with a sharp change in charge as a function of pH, but one residue of each type (histidine 95 (curve in green) and glutamic acid 165 (curve in blue)) displays abnormal behavior.

References

Bartlett, G.J. *et al.*: **Analysis of catalytic residues in enzyme active sites.** *J. Mol. Biol.* 2002, **324**:105–121.

Di Gennaro, J.A. *et al.*: **Enhanced functional annotation of protein sequences via the use of structural descriptors.** *J. Struct. Biol.* 2001, **134**:232–245.

Ewing, T.J. *et al.*: **DOCK 4.0: search strategies for automated molecular docking of flexible molecule databases.** *J. Comput. Aided Mol. Des.* 2001, **15**:411–428.

Fetrow, J.S. *et al.*: **Structure-based functional motif identifies a potential disulfide oxidoreductase active site in the serine/threonine protein phosphatase-1 subfamily.** *FASEB J.* 1999, **13**:1866–1874.

Jones, S. and Thornton, J.M.: **Prediction of protein–protein interaction sites using patch analysis.** *J. Mol. Biol.* 1997, **272**:133–143.

Laskowski, R.A. *et al.*: **Protein clefts in molecular recognition and function.** *Protein Sci.* 1996, **5**:2438–2452.

Ondrechen, M.J. *et al.*: **THEMATICS: a simple computational predictor of enzyme function from structure.** *Proc. Natl Acad. Sci. USA* 2001, **98**:12473–12478.

Reva, B. *et al.*: **Threading with chemostructural restrictions method for predicting fold and functionally significant residues: application to dipeptidylpeptidase IV (DPP-IV).** *Proteins* 2002, **47**:180–193.

Sheinerman, F.B. *et al.*: **Electrostatic aspects of protein–protein interactions.** *Curr. Opin. Struct. Biol.* 2000, **10**:153–159.

Wallace, A.C. *et al.*: **Derivation of 3D coordinate**

The limitation with all computational methods of this type is that they can only find active-site residues that conform to known active sites. A protein with a novel function will yield no result, or worse, an incorrect identification with a known site. A more general approach has been developed that does not depend on previous knowledge of active-site geometries. This employs theoretical microscopic titration curves (THEMATICS), to identify active-site residues that are potentially involved in acid-base chemistry in proteins of known structure. Location of such residues automatically determines the position of the active site, as well as providing a clue to the biochemical function of the protein.

In THEMATICS, the mean net charge of potentially ionizable groups in each residue in the protein structure is calculated as a function of pH. The resulting family of curves for each type of residue (Figure 4-36) is then analyzed for deviations from ideal behavior. A small fraction (3–7%) of all curves for all residues differ from the others in having a flat region where the residue is partially protonated over a wide pH range. Most residues with these perturbed curves occur in active sites (Figure 4-37). The method is successful for proteins with a variety of different chemistries and structures and has a low incidence of false positives. Of course, identification of acid-base residues in active sites does not necessarily establish what the overall chemical reaction must be. Most enzymatic reactions use one or more acid-base steps but catalyze other chemistries as well. Since the pK_a values of catalytic residues are likely to be perturbed by electrostatic interactions, once they are identified, computational tools such as GRASS can be used to compute and display such interactions.

Docking programs model the binding of ligands

Even when it is possible to identify active sites and to draw some conclusions about the likely chemistry they will perform, it is still necessary to determine on what substrate(s) that chemistry will operate. At present, there is no method, experimental or computational, that will enable one to find the most likely substrate for any particular active site. Some approaches involving mass spectroscopy to identify ligands pulled out of cellular extracts by the protein in question are under development, and peptide substrates for protein kinases can often be found by screening combinatorial peptide libraries, but a computational method would be most general. One promising approach is that of the program DOCK, where the shape of the binding site on a protein is represented as a set of overlapping spheres, in which the centers of the spheres become potential locations for ligand atoms. Each ligand is divided into a small set of rigid fragments that are docked separately into the binding site, allowing a degree of flexibility at the positions that join them. The fragments are rejoined later in the calculation and an energy minimum calculated for the rejoined ligand in the receptor site. The method can find binding geometries for the ligand similar to those observed crystallographically, as well as other geometries that provide good steric fit, and has been used to find possible new compounds for drug development. In such applications, each of a set of small molecules from a structural database is individually docked to the receptor in a number of geometrically permissible orientations. The orientations are evaluated for quality of fit, with the best fits being kept for examination by molecular mechanics calculations.

The method cannot take unknown conformational changes of the protein into account. In principle, it could be used to find candidate substrates for any active site, but in practice it is too computationally cumbersome, and all potential ligands are not contained in any database. The method also gives not one but many possible ligands from any database, and the energy function used to evaluate binding cannot discriminate among them.

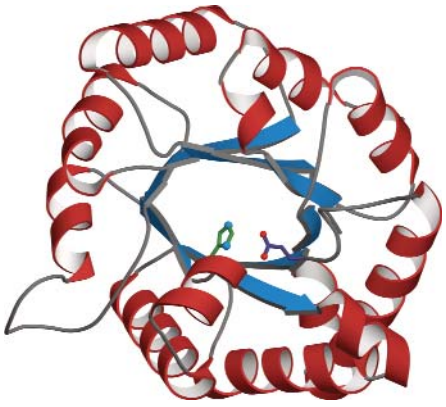


Figure 4-37 Residues that show abnormal ionization behavior with changing pH define the active site The locations of the two abnormally titrating residues in Figure 4-36 are shown on the three-dimensional structure of triosephosphate isomerase. The histidine (green) and glutamic acid (blue) that are partially protonated over a wide range of pH are both located in the active site and both are important in catalysis.

templates for searching structural databases: application to Ser-His-Asp catalytic triads in the serine proteinases and lipases. *Protein Sci.* 1996, **5**:1001–1013.

Zhang, B. *et al.*: **From fold predictions to function predictions: automation of functional site conservation analysis for functional genome predictions.** *Protein Sci.* 1999, **8**:1104–1115.

Rigid active-site geometries:
http://www.biochem.ucl.ac.uk/bsm/PROCAT/PROCAT.html

GRASS, a protein surface visualization tool:
http://trantor.bioc.columbia.edu/

DOCK:
http://www.cmpharm.ucsf.edu/kuntz/dock.html

FFF:
http://www.geneformatics.com

4-11 TIM Barrels: One Structure with Diverse Functions

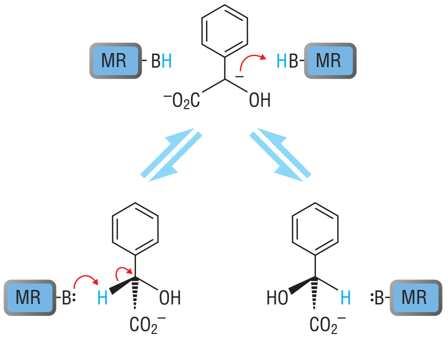


Figure 4-38 The chemical reaction catalyzed by mandelate racemase R-mandelate (left) and S-mandelate (right) can be converted into each other through the intermediate in the center. The enzyme (MR) catalyzes the reaction by removing a proton from a carbon atom adjacent to a carboxylate group and subsequently replacing it. A basic residue (B) at one side of the active site of the enzyme removes the proton and a proton is replaced by a basic residue on the other side of the active site (not shown), which can also act as an acid. The red arrows indicate the movement of electron pairs. In the reverse reaction, the two residues reverse roles.

Knowledge of a protein's structure does not necessarily make it possible to predict its biochemical or cellular functions

Perhaps the most promising case for the prediction of biochemical function from structure is when two proteins show some similarity in amino-acid sequences, share the same overall tertiary structure, and have active sites with at least some residues in common. But consideration of just such an example shows that, even with favorable parameters, function cannot always be deduced from structure.

The bacterium *Pseudomonas ovalis* lives in the soil and scavenges a wide variety of organic compounds for food. One of these is mandelate, a byproduct of decaying fruit pits (Figure 4-38). Mandelate naturally exists as two mirror-image isomers, R- and S-mandelate, and *P. ovalis* can use both as a carbon source. Two enzymes essential for the bacterium to grow on either R- or S-mandelate are mandelate racemase and muconate lactonizing enzyme.

The biochemical function of mandelate racemase (MR) is to interconvert R- and S-mandelate (Figure 4-38); because only S-mandelate is a substrate for the next enzyme in the degradative pathway, this enables *P. ovalis* to metabolize all the available mandelate instead of just half. MR is a metalloenzyme. It requires a magnesium or manganese ion for catalytic activity. Muconate lactonizing enzyme (MLE), which is further along the pathway of mandelate catabolism, transforms the *cis*, *cis*-muconic acid derived from mandelate into muconolactone (Figure 4-39). This is an essential step in the overall breakdown of mandelate into acetyl-CoA, a substrate for energy production via the tricarboxylic acid cycle. MLE is also a metalloenzyme. It requires a manganese ion for activity, although magnesium can be substituted.

The substrates for MR and MLE are very different molecules, and the biochemical functions of these two proteins are also different. Their amino-acid sequences are 26% identical, which falls in the "grey area" where one cannot predict for certain that two proteins will have any domains with a similar fold (see Figure 4-19). Secondary-structure prediction is also uninformative. Nevertheless, when the three-dimensional structures of MR and MLE were determined by X-ray crystallography, they showed that the overall folds were essentially identical (Figure 4-40). Both enzymes are TIM-barrel proteins (see section 1-18) with an extra, mostly antiparallel, beta-sheet domain attached. MR and MLE also bind their catalytic metal ions in the same positions in the structures.

Examination of the active sites shows that this similarity is preserved in detail (Figure 4-41). The amino acids that bind the metal ion are conserved between MR and MLE with one exception, and that is replaced by one with similar physical-chemical properties from a different position in the sequence. Thus, the way these two proteins bind their essential metal ions is structurally and functionally conserved, even though not all the residues involved are in exactly corresponding positions in the two sequences. Both active sites contain a pair of lysine residues in identical positions: in each case, one lysine acts as a catalytic base while the other serves to reduce the pK_a of the first through the proximity of its positive charge (see section 2-12). Opposite the lysine pair, however, the active sites of MR and MLE are different. MR has a second catalytic base, histidine 297, while MLE has a lysine residue, of uncertain role in catalysis, that occupies the same spatial position as histidine 297 but comes from sequence position 273.

The striking similarity between the active sites of MR and MLE, together with their virtually identical folds, implies that these enzymes are homologous, that is, that they diverged from a common ancestor. Nevertheless, they catalyze different chemical reactions on different substrates. Nor is there any conservation of binding specificity between them. Neither R- nor S-mandelate

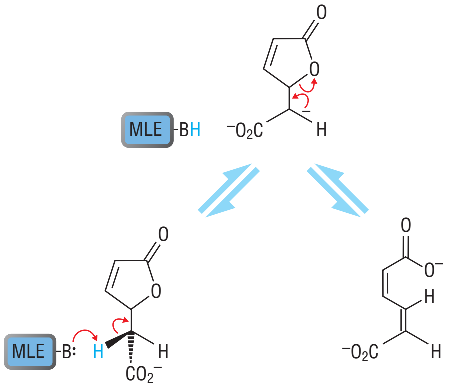


Figure 4-39 The chemical reaction catalyzed by muconate lactonizing enzyme Although the substrates are different, the core step in the catalytic mechanism of muconate lactonizing enzyme (MLE) is similar to that of mandelate racemase.

References

Babbitt, P.C. *et al.*: **A functionally diverse enzyme superfamily that abstracts the alpha protons of carboxylic acids.** *Science* 1995, **267**:1159–1161.

Gerlt, J.A. and Babbitt, P.C.: **Can sequence determine function?** *Genome Biol.* 2000, **1**:reviews0005.1–0005.10.

Gerlt, J.A. and Babbitt, P.C.: **Divergent evolution of enzymatic function: mechanistically diverse superfamilies and functionally distinct suprafamilies.** *Annu. Rev. Biochem.* 2001, **70**:209–246.

Hasson, M.S. *et al.*: **Evolution of an enzyme active site: the structure of a new crystal form of muconate lactonizing enzyme compared with mandelate racemase and enolase.** *Proc. Natl Acad. Sci. USA* 1998, **95**:10396–10401.

Nagano, N. *et al.*: **One fold with many functions: the evolutionary relationships between TIM barrel families based on their sequences, structures and functions.** *J. Mol. Biol.* 2002, **321**:741–765.

Neidhart, D.J. *et al.*: **Mandelate racemase and muconate lactonizing enzyme are mechanistically distinct and structurally homologous.** *Nature* 1990, **347**:692–694.

Petsko, G.A. *et al.*: **On the origin of enzymatic species.** *Trends Biochem. Sci.* 1993, **18**:372–376.

is a substrate or inhibitor of MLE, nor is *cis*, *cis*-muconate and muconolactone for MR. Evolutionary pressure for MR and MLE to become highly specific for their respective substrates led their substrate-binding pockets to become very different, and mutually incompatible, while the catalytic machinery remained similar. This argument implies an underlying commonality of catalytic mechanism between MR and MLE, as reflected in the conserved residues in their active sites. Both enzymes use a base—one of the pair of lysines—to abstract a hydrogen attached to a carbon atom in the substrate (see Figures 4-38 and 4-39), and in each case that carbon atom is adjacent to a carboxylate group that is coordinated to the metal ion in the active site (Figure 4-41). Presumably, the ancestral protein could carry out this chemical step on either mandelate or muconolactone or on some related molecule.

So, even if two gene products have similar sequences and share the same overall fold and some active-site residues, and even if the biochemical function of one of them is known, it is not always possible to predict the biochemical function of the other. Missing is knowledge of what molecules interact specifically with the active site of the protein of unknown function. Computational methods to determine which small molecules would bind to an active site of known shape and charge distribution do not yet exist (see section 4-9).

This particular example is not unique. MR and MLE belong to a large superfamily of TIM-barrel enzymes. All members of this superfamily have the additional beta-sheet domain, use a divalent metal ion, and have metal-binding residues in positions corresponding to those in MR and MLE. They catalyze chemical reactions as diverse as the dehydration of the sugar D-galactonate and the formation of phosphoenolpyruvate. For each of these enzymes a catalytic mechanism can be written involving base-catalyzed abstraction of a hydrogen from a carbon atom adjacent to a carboxylate group, but all of the substrates are different. A total of 21 different superfamilies of TIM-barrel enzymes have been identified on the basis of structural and functional relatedness. These 21 superfamilies include 76 difference sequence families.

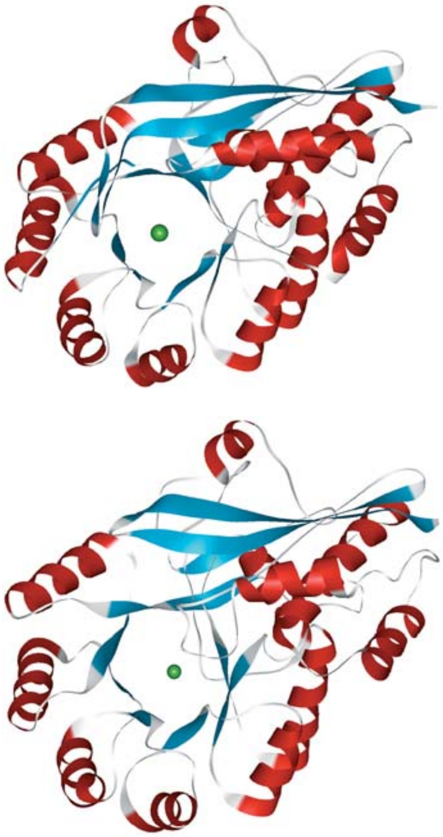


Figure 4-40 Mandelate racemase (top) and muconate lactonizing enzyme (bottom) have almost identical folds The colored spheres in the center of the structures represent the metal ions in the active sites.

Nor is the TIM barrel the only domain fold with this sort of versatility. The four-helix bundle has been found in hormones, growth factors, electron-transport proteins and enzymes. The zinc finger motif is usually found in DNA-binding proteins, where it makes sequence-specific contacts with bases in the double helix, but there are zinc finger domains that bind to RNA instead, and a number of proteins have zinc finger modules that mediate protein–protein interactions. Many other examples could be cited.

Nevertheless, there are a number of domain folds that are characteristic of a given biochemical or cellular function, and in some instances the presence of such a domain can be recognized from sequence information alone, allowing one to determine at least partial biochemical function from sequence and/or structural information. For instance, the kinase fold appears to be present almost exclusively in protein kinases; the SH2 domain appears to be used exclusively to bind phosphotyrosine-containing peptides; and in eukaryotes the seven-transmembrane helix fold appears to be used only in G-protein-coupled receptors. But many folds have so diverse a range of functions that sequence and structural information alone is unlikely to be sufficient to reveal their biochemical or cellular roles.

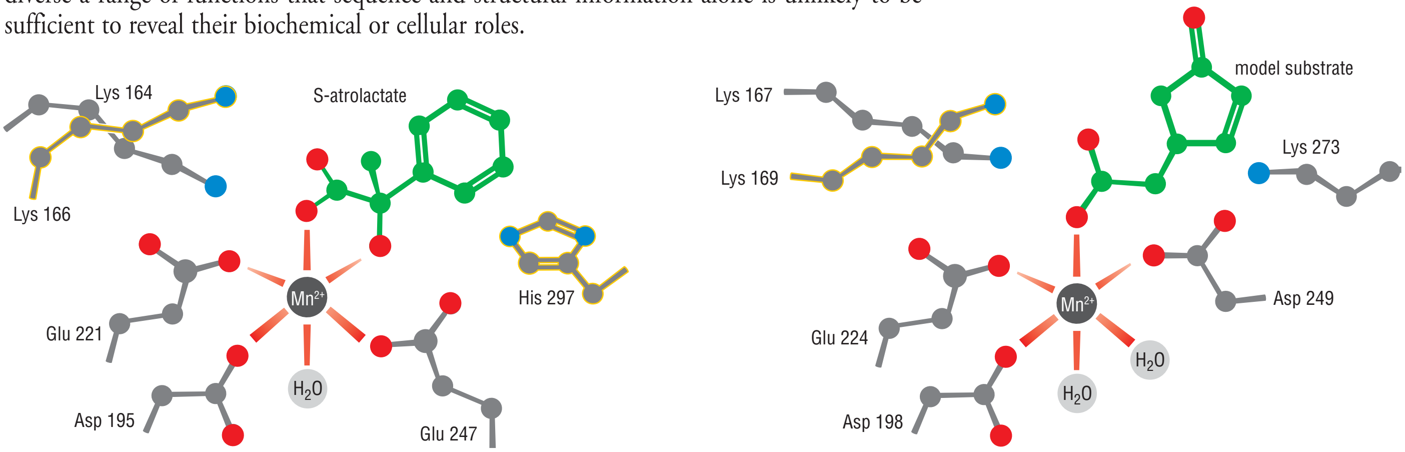


Figure 4-41 A comparison of the active sites of mandelate racemase (left) and muconate lactonizing enzyme (right) The amino acids that coordinate with the metal ion are conserved between the two enzymes, as are the catalytic residues except for histidine 297 in MR which is replaced by lysine 273 in MLE. In both cases a carboxylate group on the substrate coordinates with the metal ion. The active site of MR is shown with the inhibitor S-atrolactate bound; the MLE active site is shown with a model substrate bound. The residues shaded in yellow are the putative general acid-base catalytic residues.

4-12 PLP Enzymes: Diverse Structures with One Function

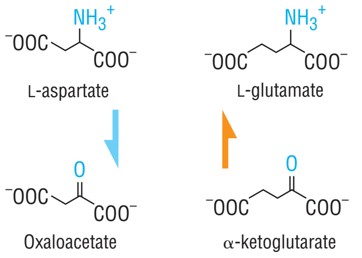


Figure 4-42 The overall reaction catalyzed by the pyridoxal phosphate-dependent enzyme L-aspartate aminotransferase

A protein's biochemical function and catalytic mechanism do not necessarily predict its three-dimensional structure

The site of biochemical function in an enzyme is characterized by a sub-site that binds the substrate and a catalytic sub-site at which the chemical reaction takes place, and these two sub-sites are usually at least partly distinct (see section 2-7). It is therefore possible to have the same arrangement of catalytic groups in combination with different arrangements of substrate-binding groups. In some cases, this produces enzymes with different specificities but which carry out the same chemistry. Such a situation is usually associated with divergent evolution from a common ancestor, in which the protein scaffold is retained but the substrate-binding sub-site is altered. It is also possible, however, for the same catalytic machinery to evolve independently on different protein scaffolds, whose substrate-binding sites may, or may not, be specific for the same substrate. This is termed convergent evolution: nature has found the same solution to the problem of catalyzing a particular reaction, but the solution has evolved in two different protein frameworks.

A prime example of convergent evolution is found among the aminotransferases. These are enzymes that "convert" one amino acid into another in a reaction known as transamination, which is central to amino-acid metabolism. In this reaction, an α-amino acid is converted to an α-keto acid, followed by conversion of a different α-keto acid to a new α-amino acid. All transaminases use the cofactor pyridoxal phosphate (PLP), derived from vitamin B_6. One of these PLP-dependent enzymes is L-aspartate aminotransferase, which converts L-aspartate to L-glutamate, via α-ketoglutarate and oxaloacetate (Figure 4-42) and is found in every living organism. Another aminotransferase, found only in bacteria, catalyzes the same reaction but is specific for the D-forms of various amino acids, including aspartate and glutamate. In bacteria, a transamination reaction involving D-glutamate, α-ketoglutarate, and pyruvate is used to produce D-alanine for synthesis of the bacterial cell wall.

Both enzymes have identical catalytic mechanisms. The amino group of the amino-acid substrate displaces the side-chain amino group of the lysine residue that binds the cofactor PLP in the active site (Figure 4-43, step 1). PLP then catalyzes a rearrangement of its new bound amino acid (step 2), followed by hydrolysis of the keto-acid portion, leaving the nitrogen of the amino acid bound to the cofactor (step 3). To regenerate the starting form of the enzyme, a different keto acid then reverses these steps and captures the bound nitrogen, producing a new amino acid and leaving the PLP once more bound to the enzyme at the active-site lysine.

A number of other amino-acid side chains in the active sites of both of these aminotransferases interact specifically with the cofactor, promoting this series of transformations over the other possible chemistries that the versatile cofactor PLP can catalyze. Yet other amino acids stabilize the position of the bound substrates and confer substrate specificity. Because these two enzymes catalyze exactly the same reaction by exactly the same mechanism, one might expect their structures to be similar. On the other hand, because one of these enzymes is specific for the commonly

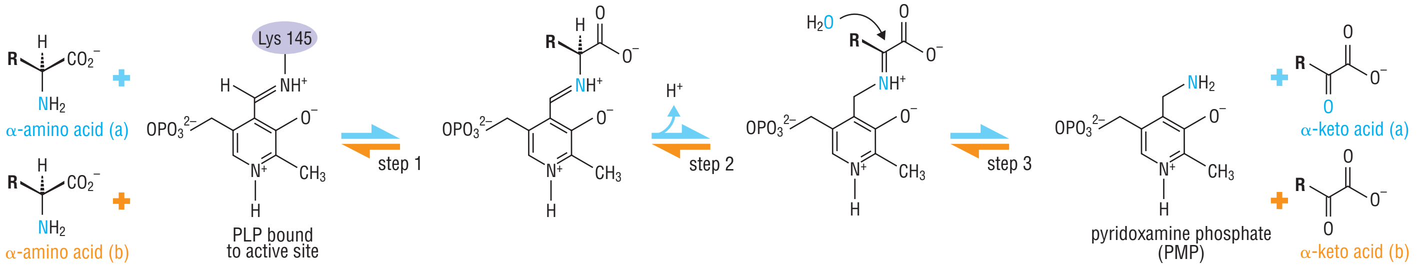


Figure 4-43 The general mechanism for PLP-dependent catalysis of transamination, the interconversion of α-amino acids and α-keto acids The amino group of the amino-acid substrate displaces the side-chain amino group of the lysine residue that holds the cofactor PLP in the active site (step 1). PLP then catalyzes a rearrangement of the amino-acid substrate (step 2), followed by hydrolysis of the keto-acid portion, leaving the nitrogen of the amino acid (blue) bound to the cofactor to form the intermediate pyridoxamine phosphate (PMP) (step 3). This forward reaction is indicated by the blue arrows. To regenerate the starting form of the enzyme, a different keto acid then reverses these steps and captures the bound nitrogen, producing a new amino acid and leaving the PLP once more bound to the enzyme at the active-site lysine (orange arrows).

References

Doolittle, R.F.: **Convergent evolution: the need to be explicit.** *Trends Biochem. Sci.* 1994, **19**:15–18.

Kirsch, J.F. *et al.*: **Mechanism of action of aspartate aminotransferase proposed on the basis of its spatial structure.** *J. Mol. Biol.* 1984, **174**:497–525.

Smith, D.L. *et al.*: **2.8-Å-resolution crystal structure of an active-site mutant of aspartate aminotransferase from *Escherichia coli*.** *Biochemistry* 1989, **28**:8161–8167.

Sugio, S. *et al.*: **Crystal structure of a D-amino acid aminotransferase: how the protein controls stereoselectivity.** *Biochemistry* 1995, **34**:9661–9669.

Yennawar, N. *et al.*: **The structure of human mitochondrial branched-chain aminotransferase.** *Acta Crystallogr. D. Biol. Crystallogr.* 2001, **57**:506–515.

Yoshimura, T. *et al.*: **Stereospecificity for the hydrogen transfer and molecular evolution of pyridoxal enzymes.** *Biosci. Biotechnol. Biochem.* 1996, **60**:181-187.

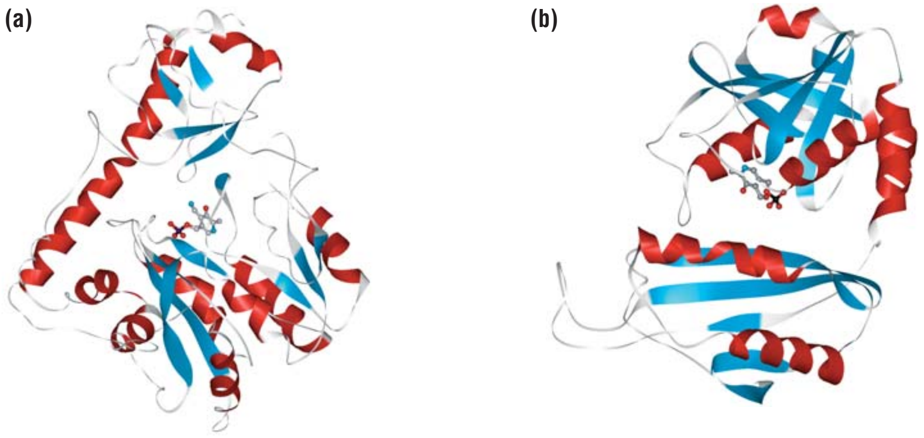


Figure 4-44 The three-dimensional structures of L-aspartate aminotransferase (left) and D-amino acid aminotransferase (right) The two proteins have completely different architectures. Not only are they different in size, they differ in their amino-acid sequence and in the folds of the protein domains. In the L-aspartate aminotransferase structure, the cofactor intermediate PMP (in a ball-and-stick representation) is shown at the active site. In the D-amino acid aminotransferase, the cofactor PLP is shown. (PDB 2aat, 1daa).

occurring L-forms of the amino acids while the other exclusively uses the rarer D-enantiomers, one might also expect that their active sites would look quite different in terms of the arrangement of side chains around the cofactor. Both these expectations are, however, far from the case.

Comparison of the amino-acid sequences of the two enzymes reveals absolutely no identity; however, as we have seen, the absence of detectable sequence identity does not necessarily mean that the protein fold will be different (see section 1-16). But in this case, comparison of the three-dimensional structures of the two enzymes shows their polypeptide-chain folds to be totally different (Figure 4-44). They clearly did not evolve from a common ancestor. When one finds two sequences and two structures that are completely different, one might on the face of it expect that they represent different mechanisms for solving the problem of catalyzing the same chemical transformation. When these two structures are examined in detail, however, the active sites are found to be strikingly similar, both in the nature of the amino acids interacting with the cofactor and their positions in space (Figure 4-45). Moreover, detailed analysis of genomic sequence data suggests that all known aminotransferases possess one or other of these polypeptide-chain folds and this same active-site configuration. It appears that this constellation of catalytic groups, in combination with the intrinsic chemistry of the PLP cofactor, is especially suited to promoting transamination, and nature has independently discovered this twice, using two different protein frameworks.

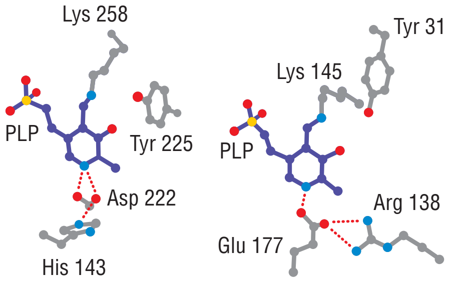


Figure 4-45 Comparison of the active sites of L-aspartate aminotransferase (left) and D-amino acid aminotransferase (right) Despite the different protein folds of these two enzymes, the active sites have converged to strikingly similar arrangements of the residues that interact with the cofactor and promote catalysis. The residues that determine which amino acid is used as a substrate and whether it will be the D- or the L-form are arranged differently in the two enzymes (not shown).

One possible explanation for the differences in three-dimensional framework is the difference in the “handedness” (or chirality) of the substrate: perhaps one fold is only suited to recognizing D-amino acids. This cannot be so, however, because there is another aminotransferase that only recognizes L-amino acids but whose sequence and polypeptide chain fold are similar to those of D-amino acid aminotransferase (Figure 4-46). These two enzymes clearly represent divergent evolution from a common ancestor. Apparently, modification of substrate specificity within the context of a given protein fold, even to the extent of reversing the handedness of the substrate, is easier than evolving a completely new catalytic mechanism.

A number of biochemical functions are carried out by enzymes that differ in their protein fold but have remarkably similar active sites. Convergent evolution to a common chemical mechanism has been observed among the serine proteases, the aminopeptidases, the NAD-dependent dehydrogenases and the sugar isomerases, to name just a few. Consequently, even if you know the biochemical function of a newly discovered protein, you cannot necessarily predict the protein fold that will carry it out. The catalytic function of enzymes can, however, sometimes be predicted by genomic analysis aimed at identification of patterns of active-site residues, and we discuss this in section 4-2.

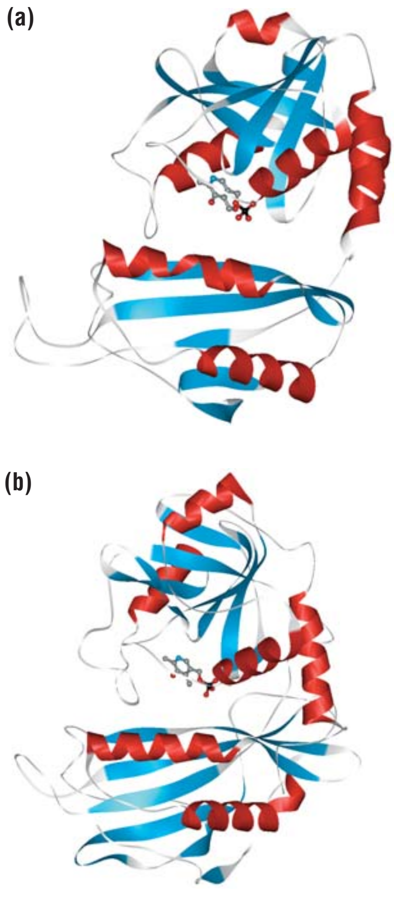


Figure 4-46 The three-dimensional structures of bacterial D-amino acid aminotransferase (top) and human mitochondrial branched-chain L-amino acid aminotransferase (bottom) These two enzymes are similar in amino-acid sequence, overall fold and active site even though their substrates are of opposite handedness. They have diverged from a common ancestor. (PDB 1daa, 1ekp).

4-13 Moonlighting: Proteins with More than One Function

In multicellular organisms, multifunctional proteins help expand the number of protein functions that can be derived from relatively small genomes

The genomes of multicellular organisms are remarkably small, in terms of number of genes (see Figure 4-10), considering the enormous increase in complexity of the organisms themselves compared with bacteria or the single-celled eukaryotes. One explanation is that the actual number of different proteins derived from a given gene can be expanded by mechanisms such as alternative splicing. An additional explanation is that, in multicellular organisms especially, a given protein may have more than one distinct biochemical and/or cellular function. The biochemical functions may include catalysis, binding, participating as a structural molecule in an assembly, or operating as a molecular switch (see sections 1-0 and 2-0). The extent of this functional diversity for any one protein is only just beginning to be appreciated.

Phosphoglucose isomerase (glucose-6-phosphate isomerase or PGI) is the second enzyme in the glycolytic pathway, a core metabolic pathway that converts glucose to pyruvate, and in which PGI converts glucose 6-phosphate to fructose 6-phosphate. The gene for PGI is thus a housekeeping gene found in nearly all organisms. Sequences of PGI from numerous organisms are well conserved, indicating that this intracellular biochemical function is catalyzed by a single type of polypeptide fold. But if one takes the protein sequence of PGI from a rabbit, say, and looks for homologous sequences in databases of other mammalian protein sequences, one finds, in addition to PGI, an identical sequence labeled neuroleukin. The protein neuroleukin was discovered as a cytokine secreted by T cells that promotes the survival of some embryonic spinal neurons and sensory nerves. It also causes B cells to mature into antibody-secreting cells. One also finds two other named protein activities with sequences identical to PGI: autocrine motility factor (AMF) and differentiation and maturation mediator (DMM). Like neuroleukin, these proteins are also secreted cytokines. AMF is produced by tumor cells and stimulates cancer-cell migration; it may be involved in cancer metastasis. DMM is isolated from culture medium in which T cells have been grown and has been shown to cause differentiation of human myeloid leukemia cells *in vitro*. Purified rabbit PGI will cause the increase in cell

Multifunctional Proteins

Protein	Function	Additional functions
Phosphoglucose isomerase	Glycolytic enzyme	Cytokine
EF-1	Elongation factor in translation	Actin-bundling protein
Cyclophilin	Peptidyl-prolyl *cis–trans* isomerase	Regulator of calcineurin
Macrophage inhibitory factor (MIF)	Activator of macrophages and T cells	Phenylpyruvate tautomerase
PutA	Proline dehydrogenase	Transcriptional repressor
Aconitase	TCA-cycle enzyme	Iron-responsive-element binding protein
Thioredoxin	Maintains SH groups in reduced state	Subunit of phage T7 DNA polymerase
Thrombin	Protease in blood clotting	Ligand for cell-surface receptor
Thymidylate synthase	Enzyme in DNA synthesis	Inhibitor of translation
FtsH	Chaperone protein in bacteria	Metalloprotease
LON	Mitochondrial protease	Chaperone protein
Methionine aminopeptidase 2	Peptidase	Protects eIF2 from phosphorylation

Figure 4-47 Some examples of multifunctional proteins with their various functions The first function column lists the biochemical function that was first identified. In most cases, this is an enzymatic activity because such activities are easily assayed. The additional functions usually depend on binding to a specific partner.

References

Cutforth, T. and Gaul, U.: **A methionine aminopeptidase and putative regulator of translation initiation is required for cell growth and patterning in *Drosophila*.** *Mech. Dev.* 1999, **82**:23–28.

Datta, B.: **MAPs and POEP of the roads from prokaryotic to eukaryotic kingdoms.** *Biochimie* 2000, **82**:95–107.

Griffith, E.C. *et al.*: **Methionine aminopeptidase (type 2) is the common target for angiogenesis inhibitors AGM-1470 and ovalicin.** *Chem. Biol.* 1997, **4**:461–471.

Haga, A. *et al.*: **Phosphohexose isomerase/autocrine motility factor/neuroleukin/maturation factor is a multifunctional phosphoprotein.** *Biochim. Biophys. Acta* 2000, **1480**:235–244.

Jeffery, C.J.: **Moonlighting proteins.** *Trends Biochem. Sci.* 1999, **24**:8–11.

Jeffery, C.J. *et al.*: **Crystal structure of rabbit phosphoglucose isomerase, a glycolytic enzyme that moonlights as neuroleukin, autocrine motility factor, and differentiation mediator.** *Biochemistry* 2000, **39**:955–964.

Liu, S. *et al.*: **Structure of human methionine aminopeptidase-2 complexed with fumagillin.** *Science* 1998, **282**:1324–1327.

Lubetsky, J.B. *et al.*: **The tautomerase active site of macrophage migration inhibitory factor is a potential target for discovery of novel anti-inflammatory agents.** *J. Biol. Chem.* 2002, **277**:24976–24982.

Sun, Y.J. *et al.*: **The crystal structure of a multifunctional**

motility seen with AMF and the dosage-dependent differentiation of human leukemia cells seen with DMM; conversely, both AMF and DMM have PGI activity. PGI, neuroleukin, AMF and DMM are the same protein, encoded by the same gene.

To carry out its cytokine and growth-factor functions it is likely that PGI/neuroleukin/AMF/DMM binds to at least one type of cell-surface receptor on a variety of target cells, and a receptor corresponding to AMF activity has been cloned from fibrosarcoma cells. Inhibitors of the PGI reaction block some, but not all, of the cytokine functions of the protein, indicating that the sites on the protein surface responsible for these different activities are at least partly distinct. Remarkably, PGI from a bacterium has been reported to have activity in the AMF assay.

The PGI reaction is extremely similar to that catalyzed by the glycolytic enzyme triosephosphate isomerase, and the catalytic mechanisms of the two reactions are identical. Nevertheless, the three-dimensional structure of PGI is completely different from that of TIM. Nothing in PGI's sequence or structure, however, suggests its additional cytokine functions; there are no obvious domains with structural similarity to any known cytokine and no sequence motifs suggestive of known growth factors or signal transduction molecules. One concludes that the extracellular cytokine functions of this protein in higher eukaryotes are at least partially independent of each other, and have evolved without gross modification of the ancestral fold.

Now that numerous genome sequences are available, each annotated according to the literature of previous studies of that organism, many other examples of multiple functions for the same protein are being discovered (Figure 4-47). One is methionine aminopeptidase type 2 (MetAP2), which catalyzes the removal of the amino-terminal methionine from the growing polypeptide chain of many proteins in eukaryotes. MetAP2 is the target for the anti-angiogenesis drugs ovalicin and fumagillin, which act by inhibiting this catalytic activity. DNA sequence analysis of MetAP2 genes reveals, as expected, sequence homologies with MetAP2 genes from other organisms but also with various eukaryotic homologs of a rat protein known originally as p67. This intracellular protein protects the alpha-subunit of eukaryotic initiation factor 2 (eIF2) from phosphorylation by its kinases. This activity of p67 is observed in different stress-related situations such as heme deficiency in reticulocytes (immature red blood cells), serum starvation and heat shock in mammalian cells, vaccinia virus infection of mammalian cells, baculovirus infection of insect cells, mitosis, apoptosis, and even possibly during normal cell growth. MetAP2 and p67 are identical proteins encoded by the same gene. Inhibitors of MetAP2 activity do not inhibit the translational cofactor activity of MetAP2, suggesting that the two functions are independent. Some mutations in the MetAP2 gene in *Drosophila* result in loss of ventral tissue in the compound eye as well as extra wing veins, whereas others impair tissue growth. However, it is not clear whether these phenotypes are due to loss of MetAP2's catalytic activity, translational cofactor activity, or both. Another example is the cytokine macrophage inhibitory factor, MIF (Figure 4-48), which also has enzymatic activity.

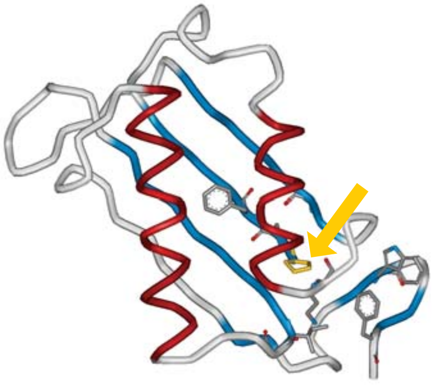


Figure 4-48 The three-dimensional structure of the monomer of macrophage inhibitory factor, MIF The protein is an important proinflammatory cytokine that activates T cells and macrophages. It is also an enzyme that catalyzes the tautomerization of phenylpyruvic acid. The residues involved in substrate binding and catalysis are shown. The proline associated with the active site is indicated by the yellow arrow. Because the active site overlaps with the binding site for receptors of the cytokine function of MIF, inhibitors of its enzymatic activity are also potential antiinflammatory drugs. (PDB 1ljt)

The term "moonlighting" has been coined to describe the performance of more than one job by the same protein. From the point of view of any one experiment, each job may appear to be the main activity and the other(s) to be the sideline. Consequently, in multicellular organisms especially, knowledge of one function of a gene product does not necessarily mean that all its functions have been determined. This fact has profound consequences for gene knockout experiments (by, for example, antisense RNA, RNA interference (RNAi), or gene disruption) as a means of determining function. The phenotype of a knockout animal or cell may be the result of the loss of all the different functions that a protein can carry out, or may differ in different tissues or under different conditions in which various functions are dominant.

protein: phosphoglucose isomerase/ autocrine motility factor/neuroleukin. *Proc. Natl Acad. Sci. USA* 1999, **96**:5412–5417.

Swope, M.D. and Lolis, E.: **Macrophage migration inhibitory factor: cytokine, hormone, or enzyme?** *Rev. Physiol. Biochem. Pharmacol.* 1999, **139**:1–32.

Watanabe, H. *et al.*: **Tumor cell autocrine motility factor is the neuroleukin/phosphohexose isomerase polypeptide.** *Cancer Res.* 1996, **56**:2960–2963.

Xu, W. *et al.*: **The differentiation and maturation mediator for human myeloid leukemia cells shares homology with neuroleukin or phosphoglucose isomerase.** *Blood* 1996, **87**:4502–4506.

4-14 Chameleon Sequences: One Sequence with More than One Fold

Some amino-acid sequences can assume different secondary structures in different structural contexts

The concept that the secondary structure of a protein is essentially determined locally by the amino-acid sequence is at the heart of most methods of secondary structure prediction; it also underlies some of the computational approaches to predicting tertiary structure directly from sequence. Although this concept appears to be valid for many sequences, as the database of protein structures has grown, a number of exceptions have been found. Some stretches of sequence up to seven residues in length have been identified that adopt an alpha-helical conformation in the context of one protein fold but form a beta strand when embedded in the sequence of a protein with a different overall fold. These sequences have been dubbed **chameleon sequences** for their tendency to change their appearance with their surroundings. One survey of all known protein structures up to 1997 found three such sequences seven residues long (Figure 4-49), 38 such sequences six residues long, and 940 chameleon sequences five residues long. Some were buried and some were on the surface; their sequences varied considerably but there tended to be a preponderance of alanines, leucines and valines and a dearth of charged and aromatic residues.

We have already seen that some segments in certain proteins can change their conformation from, for example, an alpha helix to a loop in response to the binding of a small molecule or another protein or to a change in pH. For example, when elongation factor Tu switches from its GTP-bound form to its GDP-bound form, a portion of the switch helix unravels, breaking an interaction between two domains (see section 3-9).

The ability of amino-acid sequences to convert from an alpha-helical to a beta-strand conformation has received extensive attention recently, as this structural change may induce many proteins to self-assemble into so-called amyloid fibrils and cause fatal diseases (see section 4-15). A number of sequences that are not natural chameleons can become such by a single point mutation, suggesting a possible mechanism whereby such diseases may be initiated. One example is the bacterial protein Fis, a DNA-binding protein that is implicated in the regulation of DNA replication and recombination as well as in transcriptional regulation. A peptide segment in Fis can be converted from a beta strand to an alpha helix by a single-site mutation, proline 26 to alanine. Proline 26 in Fis occurs at the point where a flexible extended beta-hairpin arm leaves the core structure (Figure 4-50a). Thus it can be classified as a "hinge proline" located at the carboxy-terminal end of one beta strand and the amino-terminal cap of the following alpha helix. The replacement of proline 26 with alanine extends the alpha helix for two additional turns in one of the dimeric subunits of Fis; therefore, the structure of the peptide from residues 22 to 26 is converted from a beta strand to an alpha helix by this one mutation (Figure 4-50b). Interestingly, this peptide in the second monomer subunit retains its beta-strand conformation in the crystal structure of Fis, suggesting that the alpha-helical and beta-sheet conformation are very similar in energy for this sequence and that only small local changes in environment are needed to cause it to flip from one form to the other.

While the conversion of a beta strand to an alpha helix in Fis is caused by a mutation, and has no implications for normal function, some proteins contain natural chameleon sequences that may be important to their function. One example is a DNA-binding transcriptional regulator from yeast, the MATα2 protein, which helps determine two differentiated cell types (mating types) in growing yeast cells by repressing genes whose expression is required for one of the two types. MATα2 binds to DNA in association with a second protein, MCM1, so that one copy of MATα2 binds on each side of MCM1 (Figure 4-51a). In the crystal structure of this complex

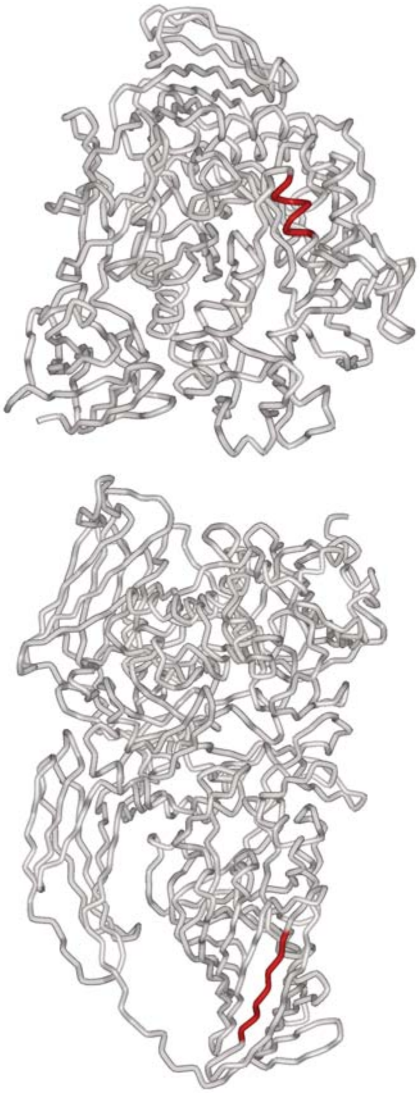


Figure 4-49 Chameleon sequences The protein backbones of the enzymes cyclodextrin glycosyltransferase (PDB 1cgu) (top) and beta-galactosidase (PDB 1bgl) (bottom), each of which contains the chameleon sequence LITTAHA (shown in red), corresponding to residues 121–127 in the sequence of cyclodextrin glycosyltransferase and residues 835–841 in beta-galactosidase. In the former structure, the sequence forms two turns of alpha helix; in the latter, it is a beta strand.

Definitions

chameleon sequence: a sequence that exists in different conformations in different environments.

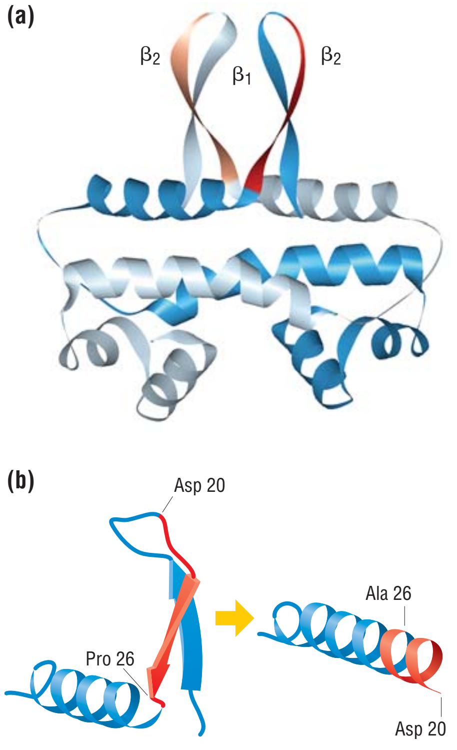


Figure 4-50 Chameleon sequence in the DNA-binding protein Fis **(a)** The structure of the dimer of the sequence-specific DNA-binding protein Fis shows a predominantly alpha-helical fold with two strands of antiparallel beta sheet, β_1 and β_2 (red), at the amino terminus. (PDB 1f36) **(b)** The replacement of proline 26 at the end of the second beta strand with an alanine converts this beta strand into two additional turns of the alpha helix that follows.

bound to DNA, an eight-amino-acid sequence adopts an alpha-helical conformation in one of two copies of the MATα2 monomer and a beta-strand conformation in the other (Figure 4-51b). Although there is no direct evidence that both forms exist in biology, such an alternative fold could have functional consequences. In most sites the sequences recognized by the MATα2 monomers are identical. However, there are separations of two to three base pairs between the MCM1- and MATα2-binding sites in the natural yeast promoters to which this transcription factor binds, and the different conformations may permit such variations to be tolerated. MATα2 can also form a complex with another transcriptional modulator, MAT**a**1, and in this context, the change in conformation may again allow MATα2 to accommodate differences in the spacing of the sites on DNA. The ability of parts of MATα2 to change conformation in different contexts could help this protein to bind to a number of sites on the genome.

To probe the context dependence of the structures of short polypeptide sequences, an 11-amino-acid chameleon sequence has been designed that folds as an alpha helix when in one position but as a beta sheet when in another position of the primary sequence of the immunoglobulin-binding domain of protein G. This protein from *Staphylococcus aureus* binds to the Fc region of IgG antibodies and is thought to protect the bacteria from these antibodies by blocking their interactions with complement and Fc receptors. Both proteins, chameleon-alpha and chameleon-beta, are folded into structures similar to native protein G except for the small region of the chameleon sequence.

These examples illustrate the general principle that the secondary structures of short peptide segments can often depend more on the tertiary structural context in which they are placed than on their intrinsic secondary structure propensities. The balance between inherent tendency and the effect of environment will be different for different sequences. If the free energies of a peptide in its alpha-helical and beta-sheet conformations are similar, then the energies of interaction between the peptide and the environment could be enough to tip the balance in favor of one or the other.

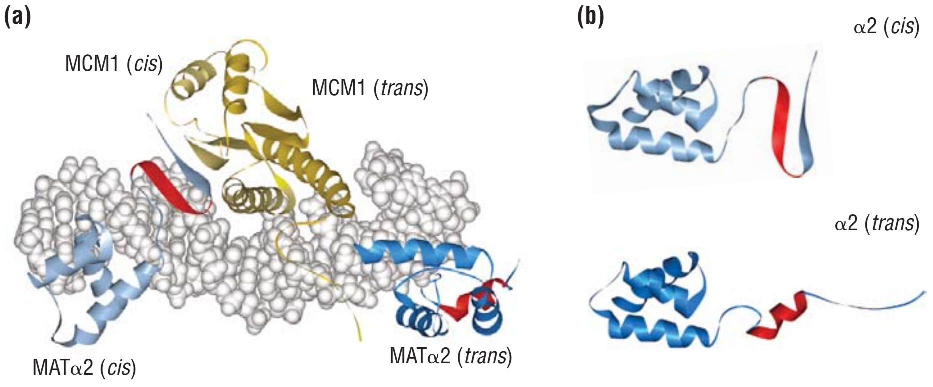


Figure 4-51 Chameleon sequence in the DNA-binding protein MATα2 from yeast **(a)** The structure of the complex of MATα2 (blue and red) with its transcriptional co-regulator MCM1 (yellow) bound to a target site in DNA. At such a site, two monomers of MATα2 bind to two (usually identical) DNA sequences on either side of two monomers of MCM1. **(b)** An eight-amino-acid sequence (red) adopts a beta-strand conformation in one MATα2 molecule (the *cis* monomer; light blue) and an alpha-helical conformation in the other (the *trans* monomer; dark blue). (PDB 1mnm)

References

Mezei, M.: **Chameleon sequences in the PDB.** *Protein Eng.* 1998, **11**:411–414.

Minor, D.L. Jr. and Kim, P.S.: **Context-dependent secondary structure formation of a designed protein sequence.** *Nature* 1996, **380**:730–734.

Smith, C.A. *et al.*: **An RNA-binding chameleon.** *Mol. Cell* 2000, **6**:1067–1076.

Sudarsanam, S.: **Structural diversity of sequentially identical subsequences of proteins: identical octapeptides can have different conformations.** *Proteins* 1998, **30**:228–231.

Tan, S. and Richmond T.J.: **Crystal structure of the yeast MATalpha2/MCM1/DNA ternary complex.** *Nature* 1998, **391**:660–666.

Yang, W.Z. *et al.*: **Conversion of a beta-strand to an alpha-helix induced by a single-site mutation observed in the crystal structure of Fis mutant Pro26Ala.** *Protein Sci.* 1998, **7**:1875–1883.

4-15 Prions, Amyloids and Serpins: Metastable Protein Folds

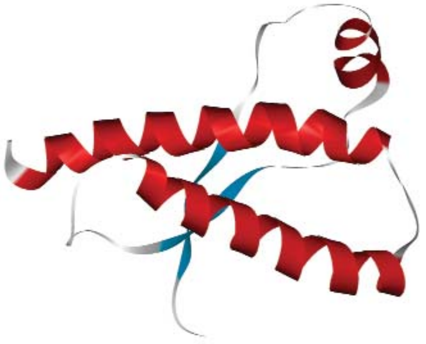


Figure 4-52 **The prion protein** This figure shows the soluble form of Syrian hamster prion protein PrPSc, which was generated by removing the amino terminus. This protein consists of residues 90–231, incorporating the region thought to be involved in the structural change. The protein was generated by refolding to produce the structure of the cellular form of the protein. The precise structure of the disease-causing form is not yet known, but is known to have much more beta sheet than the cellular form.

A single sequence can adopt more than one stable structure

The existence of chameleon sequences may reflect a general principle: that not all sequences fold into one unique structure. Some structures may be **metastable**—able to change into one or more different stable structures. But are complete protein sequences of such plasticity found in nature? It appears that the answer is yes although rarely, and that such cases are often, at least so far, associated with severe mammalian disease. The best characterized of these changeable structures is the prion associated with Creutzfeldt-Jakob disease (CJD) in humans and scrapie and bovine spongiform encephalopathy (BSE) in sheep and cattle (Figure 4-52). Prions are infectious proteins whose misfolded form is identical in sequence to the normal cellular form of the same protein. The two forms have, however, quite different conformations and physical properties. The infectious form has a propensity to form aggregrates, possesses secondary structure content that is rich in beta sheets, is partially resistant to proteolysis, and is insoluble in nonionic detergents. In contrast, the cellular form contains little beta structure, is sensitive to protease digestion, and is soluble in nonionic detergent. Contact with the infectious form causes the cellular counterpart to undergo pronounced conformational changes that lead, ultimately, to the formation of cytotoxic protein aggregates consisting almost entirely of the infectious conformation. Although the molecular events that lead to this profound conformational change are poorly understood, there is substantial evidence that the infectious form acts as a template directing the structural rearrangement of the normal form into the infectious one. Studies of synthetic peptides derived from the prion sequence indicate that a stretch of up to 55 residues in the middle of the protein has the propensity to adopt both alpha-helical and beta-sheet conformations. Presumably, the infectious form arises spontaneously in a small number of molecules as a result of this inherent plasticity.

A similar mechanism may underlie protein aggregate formation in a group of about 20 diseases called amyloidoses, which include Alzheimer's, Parkinson's, and type II diabetes. Each disease is associated with a particular protein, and extracellular aggregates of these proteins are thought to be the direct or indirect origin of the pathological conditions associated with the disease. Strikingly, the so-called amyloid fibrils characteristic of these diseases arise from well known proteins, including lysozyme and transthyretin, that have well defined, stable, non-identical folds but produce fibrous protein aggregates of identical, largely beta-sheet, structure (Figure 4-53).

Recent studies suggest that the ability to undergo a refolding leading to amyloid formation is not unique to these proteins, but can be observed in many other proteins under laboratory conditions such as low pH. One conclusion from such findings is that prions and other proteins that cause disease by this mechanism may differ from the vast array of "normal" cellular proteins only in having sequences that can undergo such refolding spontaneously under physiological conditions. Clearly, if this is the case, it is possible that a single point mutation may convert a harmless protein into one that can refold spontaneously; such mutations have been found associated with some of the amyloidoses.

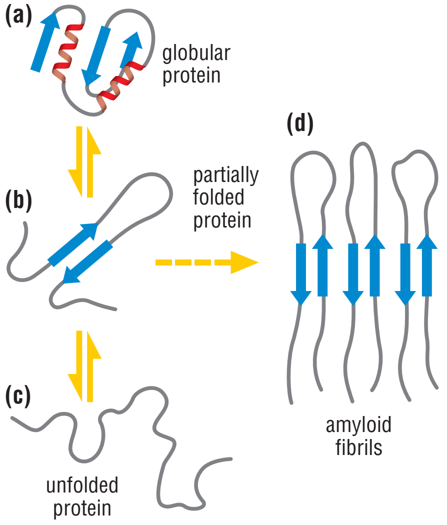


Figure 4-53 **A possible mechanism for the formation of amyloid fibrils by a globular protein** The correctly folded protein **(a)** is secreted from the cell. Under certain conditions, or because it contains a mutation, the protein unfolds partially **(b)** or completely **(c)**; the unfolded forms can also refold partially or completely. The partially unfolded form is prone to aggregation, which results in the formation of fibrils **(d)** and other aggregates that accumulate in the extracellular space.

Definitions

metastable: only partially stable under the given conditions. In the case of protein structures, a metastable fold exists in equilibrium with other conformations or with the unfolded state.

References

Cohen, B.I. *et al.*: **Origins of structural diversity within sequentially identical hexapeptides.** *Protein Sci.* 1993, **2**:2134–2145.

Dalal, S. and Regan, L.: **Understanding the sequence determinants of conformational switching using protein design.** *Protein Sci.* 2000, **9**:1651–1659.

Dobson, C.M.: **Protein misfolding, evolution and disease.** *Trends Biochem. Sci.* 1999, **24**:329–332.

Engh, R.A. *et al.*: **Divining the serpin inhibition mechanism: a suicide substrate "springe"?** *Trends Biotechnol.* 1995, **13**:503–510.

Kabsch, W. and Sander, C.: **On the use of sequence homologies to predict protein structure: identical pentapeptides can have completely different conformations.** *Proc. Natl Acad. Sci. USA* 1984, **81**:1075–1078.

Ko, Y.H. and Pedersen, P.L.: **Cystic fibrosis: a brief look at some highlights of a decade of research focused on elucidating and correcting the molecular basis of**

Structural plasticity can also be part of the normal function of a protein. Often, when this is the case, ligand binding or specific proteolytic modification is needed to drive one folded form into the other. Large, ligand-induced domain rearrangements such as that found in elongation factor Tu (see section 3-9) can be considered examples of different overall protein folds induced by the state of assembly. Perhaps the best example of a large structural rearrangement caused by limited proteolysis is found in the family of protein protease inhibitors called the serpins. Some protein protease inhibitors function as rigid substrate mimics; the serpins differ fundamentally in that the loop that recognizes and initially binds the protease active site is flexible and is cleaved by the protease. The serpin remains bound to the enzyme but the cleavage triggers a refolding of the cleaved structure that makes it more stable: one segment of the cleaved loop becomes the central strand of an existing beta sheet in the center of the serpin, converting it from a mixed beta sheet to a more stable antiparallel form (Figure 4-54); in at least one case, the downstream segment also becomes the edge strand of another beta sheet. If the cleaved serpin is released it cannot reassociate with the protease because this refolding has made the recognition strand unavailable for binding.

Mutations in serpins leading to misfolding and aggregation have also been found in some human diseases. For example, the so-called Z-variant of the serpin alpha$_1$-antitrypsin (in which glutamic acid 342 is mutated to lysine) is retained within hepatocytes as inclusion bodies; this is associated with neonatal hepatitis and cirrhosis. The inclusion bodies form because the mutation perturbs the conformation of the protein, facilitating a sequential interaction between the recognition loop of one molecule and beta-sheet A of a second; this could be thought of as a pathological case of domain swapping (see section 2-4).

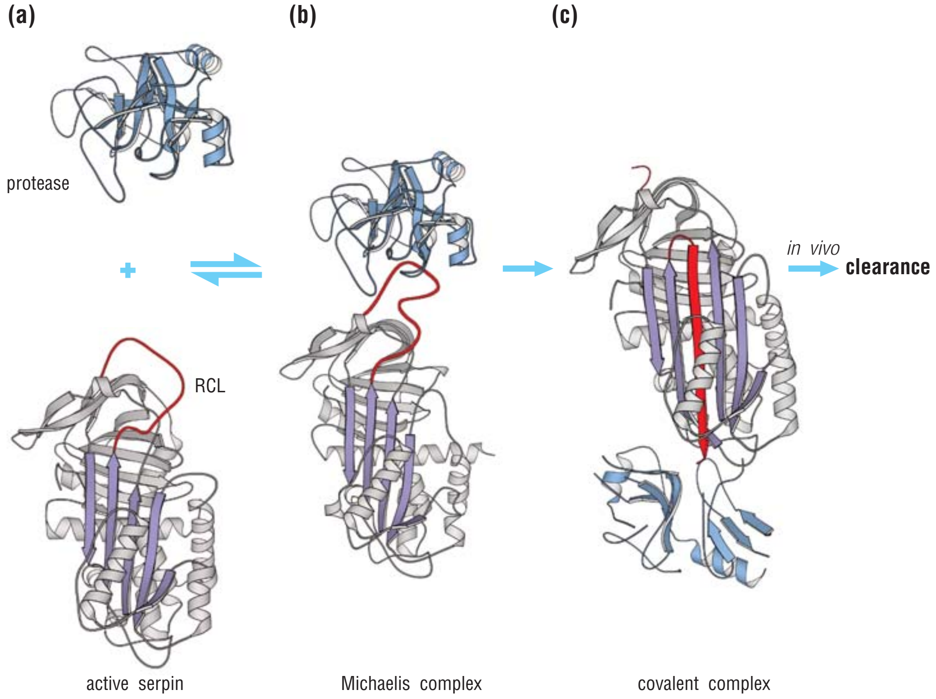


Figure 4-54 Structural transformation in a serine protease inhibitor on binding protease **(a)** The exposed reactive center loop (RCL; red) of the serpin alpha$_1$-antitrypsin (grey) is shown binding to the target protease (blue) to form an enzyme–substrate complex, the Michaelis complex **(b)**. The RCL of the serpin is then cleaved by the target protease, leading to the insertion of the unconstrained RCL into the serpin beta sheet and the formation of a covalent complex that is trapped by release of the newly formed amino terminus of the serpin **(c)**. This complex is then targeted for clearance. Adapted from Ye, S. and Goldsmith, E.J.: *Curr. Opin. Struct. Biol.* 2001, **11**:740–745.

the disease. *J. Bioenerg. Biomembr.* 2001, **33**:513–521.

Leclerc, E. *et al.*: **Immobilized prion protein undergoes spontaneous rearrangement to a conformation having features in common with the infectious form.** *EMBO J.* 2001, **20**:1547–1554.

Mahadeva, R. *et al.*: **6-mer peptide selectively anneals to a pathogenic serpin conformation and blocks polymerization. Implications for the prevention of Z alpha(1)-antitrypsin-related cirrhosis.** *J. Biol. Chem.* 2002, **277**:6771–6774.

Ye, S. and Goldsmith, E.J.: **Serpins and other covalent protease inhibitors.** *Curr. Opin. Struct. Biol.* 2001, **11**:740–745.

Zahn, R. *et al.*: **NMR solution structure of the human prion protein.** *Proc. Natl Acad. Sci. USA* 2000, **97**:145–150.

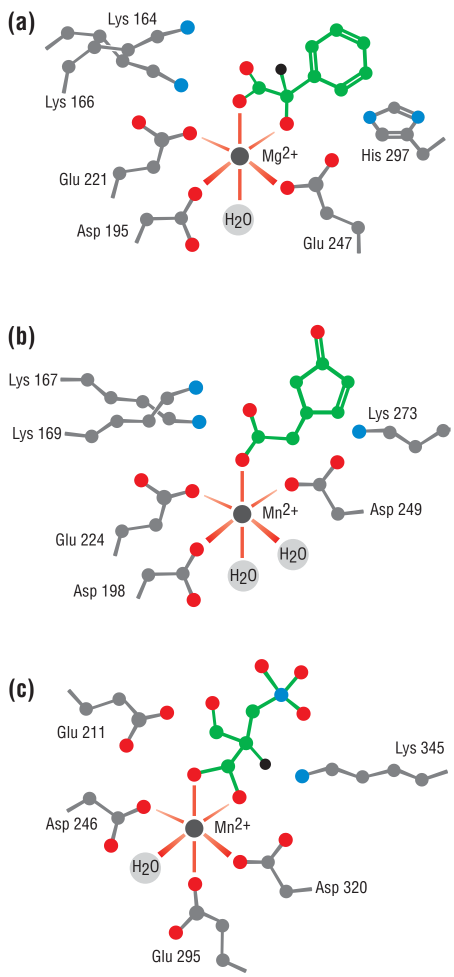


Figure 4-55 Active sites of MR, MLE, and enolase Schematic diagrams of the arrangements of the active-site residues of **(a)** mandelate racemase (MR), **(b)** muconate lactonizing enzyme (MLE) and **(c)** enolase. The types of amino acid that coordinate the divalent metal ion are conserved between the three enzymes. The other catalytic residues, however, are conserved neither in exact position nor in chemical type; nevertheless, these various residues can carry out similar chemistry. In each reaction, the carboxy group of the substrate forms a ligand to the metal ion in the active site, facilitating abstraction of a proton. The enolic intermediate resulting from this common core step is stabilized by interactions with other electrophilic groups in the active site. These groups differ among the three enzymes. The rest of the substrate-binding pocket differs considerably among the three enzymes. (PDB 1mns, 1muc and 1one)

Determining biochemical function from sequence and structure becomes more accurate as more family members are identified

The identification of the function of the *E. coli* protein f587, known originally only as an uncharacterized open reading frame in the *E. coli* genome sequence, is an illustration of the importance of lateral thinking. In this case, the sequence and structural information was eventually interpreted in the light of the known genetics and physiology of the bacterium.

The story starts with the related enzymes mandelate racemase (MR) and muconate lactonizing enzyme (MLE) (see section 4-11). The enzyme enolase, which catalyzes the conversion of 2-phospho-D-glycerate to phosphoenolpyruvate, was subsequently found to have a similar degree of sequence identity to MR and MLE (26%) and a similar distribution of conserved residues. It has the same polypeptide chain fold and many of the conserved residues map to the active-site region of the fold. Together, these three enzymes form part of the so-called enolase superfamily.

With three related sequences and structures in hand, it became apparent that not every part of the active site was preserved (Figure 4-55). All three enzymes require at least one divalent metal ion for activity, and the carboxylate ligands to these metal ions are present in all three proteins; in enolase however, glutamic acid and aspartic acid are substituted for one another. The remainder of the catalytic machinery is even more divergent in enolase. Yet, the residues are conserved in terms of their broad chemical role although not in terms of their identity. Glutamic acid 211 in enolase could in principle act as a general acid-base group, just like lysine 166 in MR and lysine 169 in MLE. On the other side of the substrate-binding pocket, lysine 345 in enolase occupies the same position as lysine 273 in MLE; the corresponding position in the structure of MR is occupied by histidine 297, which might have a different role. All three enzymes use their bound metal ions in the same way—to activate a C–H bond adjacent to a carboxylate group for abstraction of the hydrogen by a base on the enzyme.

MR has two substrates with the C–H bond in different positions (see Figure 4-38) and its function is to interconvert them; thus, two different acid-base groups are needed. In the case of enolase and MLE, only one proton needs to be abstracted and so only a single base is required. Thus, both lysine 166 and histidine 297 in MR function as acid-base groups. In MLE and enolase, however, the abstraction of a single proton is carried out only by lysine 169 and lysine 345, respectively. This conserved base-catalyzed, metal-promoted proton-transfer step is the common function linking all three enzymes.

Alignments based on conservation of residues that carry out the same active-site chemistry can identify more family members than sequence comparisons alone

If chemistry is the conserved feature, rather than the absolute identity and position in the sequence of the groups that carry out the chemistry, then a more sophisticated approach to finding homologous sequences would be to search for patterns of residues that can perform the same chemistry regardless of their specific amino-acid identities. Such a search, using a specialized computer program, identified dozens of potential members of the enolase/MR/MLE superfamily, most of which could not have been detected by conventional sequence comparison.

One of these predicted homologs, open reading frame f587 in the *E. coli* genome, coded for a protein of unknown function. Alignment of the f587 sequence with those of the other three proteins on the basis of the conservation of active-site chemical function showed that f587 contains the requisite metal-ion ligands and conservation of an active-site histidine—histidine 285—which aligns with histidine 297 of MR. From the position of this base, the prediction would be that the substrate for f587, whatever it might be, would be a carboxylate-containing molecule with a proton on an adjacent carbon that has the R-configuration, as does R-mandelate.

In well studied model organisms, information from genetics and cell biology can help identify the substrate of an “unknown” enzyme and the actual reaction catalyzed

The remaining problems—what is the substrate and overall reaction of f587—are insoluble from sequence and structural information alone. For f587, however, additional information was available. *E. coli*, like most bacteria, tends to organize its genes into operons encoding a set

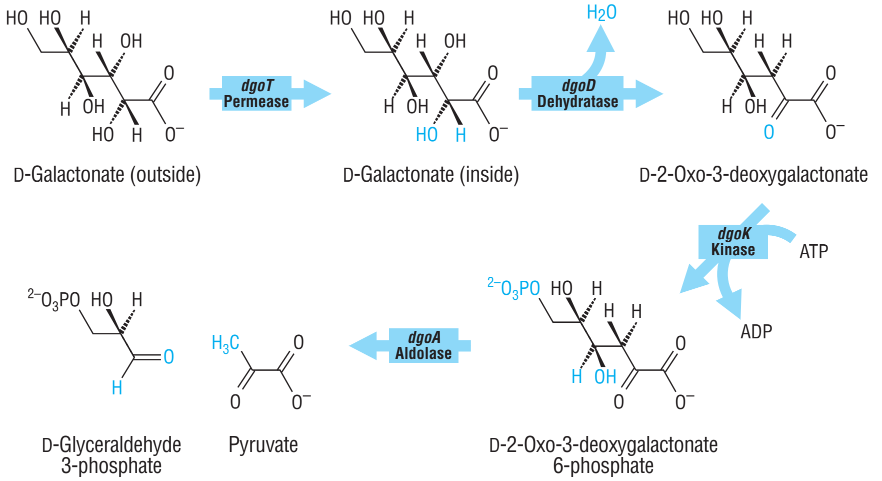


Figure 4-56 The pathway for the utilization of galactonate in *E. coli* D-galactonate is transported into the cell, dehydrated, phosphorylated, and then cleaved to produce glyceraldehyde 3-phosphate and pyruvate. This pathway allows the bacterium to grow on galactonate as a sole carbon source. The names of the genes and the enzyme activity they represent are given in the blue arrows. F587 has now been identified as the gene *dgoD*, encoding galactonate dehydratase.

of proteins that act in the same pathway. The genome sequence suggested that f587 is the fourth in a series of five open reading frames that could constitute a single operon. The other open reading frames were known to encode proteins involved in galactonate metabolism. They appear to be part of the pathway in which galactonate is imported into the cell by galactonate permease and then degraded stepwise into glyceraldehyde 3-phosphate and pyruvate. Given the proposed functions of the other proteins, the only function needed to complete this pathway was galactonate dehydratase (GalD) (Figure 4-56).

Dehydration of galactonate could be catalyzed by abstracting the 2-R proton via a catalytic base with assistance of a metal ion to activate the C–H bond. From its sequence, f587 appeared to contain all the necessary elements in the right stereochemical configuration to catalyze just such a reaction. To test the hypothesis that f587 encodes GalD, a cell-free extract of *E. coli* transformed with a plasmid overexpressing f587 was assayed for GalD activity. An alpha-keto acid was produced from galactonate, but not from other similar sugars. Subsequent purification of the f587 gene product confirmed it as GalD. Finally, the crystal structure of GalD was determined with an analog of galactonate bound. The overall polypeptide chain fold (Figure 4-57) is the same as that of MR, MLE, and enolase and the active site also resembles those of the other family members (Figure 4-58).

The case of galactonate dehydratase shows that sequence comparisons can identify overall protein structure class and locate the active-site residues, even in cases of very low sequence identity. They can also provide specific information about bound ligands—in this case, a divalent metal ion—and about the residues that bind them. Comparison of a protein of unknown function with just one other homologous sequence will usually not identify the other active-site residues or establish any commonality of function unless the overall sequence identity is very high. Comparisons of multiple sequences are much more informative, however, and will often be able to detect the functional groups that carry out the common core chemical step. When the active-site structures of at least some of the proteins being compared are known, the arrangement of these groups may also reveal the stereochemistry of the reaction being catalyzed. But what sequence and structure usually cannot do alone is to identify the substrate(s) of the reaction and the overall chemistry. Identification of chemistry is more reliable than identification of specificity. Proceeding from sequence to consequence in cases of very low sequence identity requires other sources of information, as illustrated by this case study.

References

Babbitt, P.C. and Gerlt, J.A.: **Understanding enzyme superfamilies. Chemistry as the fundamental determinant in the evolution of new catalytic activities.** *J. Biol. Chem.* 1997, **272**:30591–30594.

Babbitt, P.C. *et al.*: **The enolase superfamily: a general strategy for enzyme-catalyzed abstraction of the alpha-protons of carboxylic acids.** *Biochemistry* 1996, **35**:16489–16501.

Babbitt, P.C. *et al.*: **A functionally diverse enzyme superfamily that abstracts the alpha protons of carboxylic acids.** *Science* 1995, **267**:1159–1161.

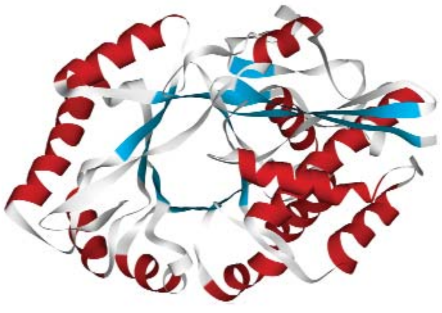


Figure 4-57 Structure of galactonate dehydratase The fold is the same as those of MR, MLE, and enolase (see Figure 4-40 for MR and MLE).

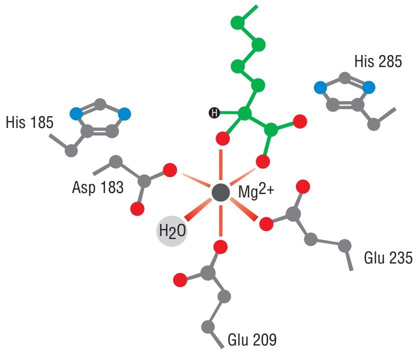


Figure 4-58 Schematic diagram of a model of the active site of galactonate dehydratase with substrate bound The metal-ion coordination and the disposition of the catalytic base histidine 285 and the electrophilic groups that interact with the substrate are similar to those found in MR, MLE and enolase, even though the overall reactions they catalyze are completely different.

4-17 Starting from Scratch: A Gene Product of Unknown Function

Function cannot always be determined from sequence, even with the aid of structural information and chemical intuition

About 30% of the 6,282 genes in the genome of the budding yeast *Saccharomyces cerevisiae* code for proteins whose function is completely unknown. One of these is gene *YBL036c*, whose sequence indicates that it encodes a protein of 257 amino acids. A comparison of this sequence against all genomic DNA sequences in the databases as of 1 June 2002 indicates about 200 other proteins whose amino-acid sequences show a greater-than-chance similarity to that of YBL036c. The putative homologs come from every kingdom of life—and none has a known function. At the time of writing, this is still true: although there are now clues to the function of the gene in yeast, the last chapter in this story is still to be written. What follows is an account of the avenues explored and where they lead.

As homologs of YBL036c are ubiquitous, the protein is more likely to function in some fundamental cellular process than to be involved in, for example, some aspect of cell–cell communication that would be confined to multicellular organisms. Thus yeast, a single-celled eukaryote whose complete genome sequence is known and whose metabolic processes can easily be studied by genetic methods, should be an ideal model organism in which to uncover the functions of this family of proteins.

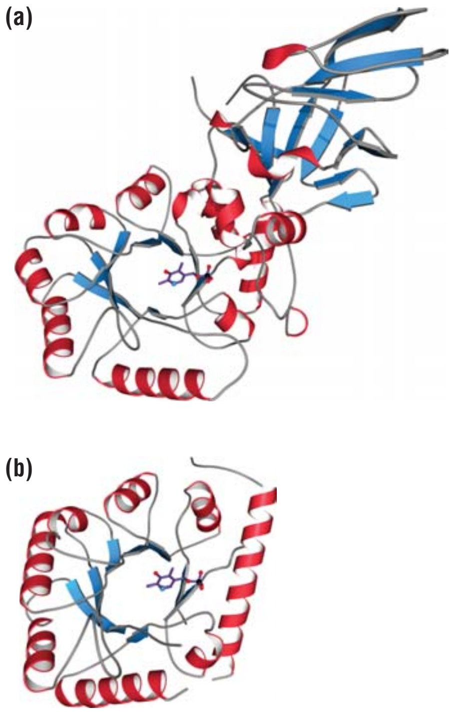


Figure 4-59 **The three-dimensional structures of bacterial alanine racemase and yeast YBL036c** The structure of the large domain of alanine racemase **(a)** is similar to the overall structure of YBL036c **(b)**. The yeast protein lacks the largely antiparallel beta-sheet domain of the racemase; however, the active sites, indicated by the presence of the bound pyridoxal phosphate cofactor (shown in ball-and-stick form), are located in the same place in both proteins. (PDB 1sft and 1ct5)

YBL036c was selected as one of the first gene products to be studied in a project aimed at determining structures for yeast proteins which, because of the absence of clear sequence similarity to other proteins, seemed likely to have novel folds. In fact when the three-dimensional structure was determined by X-ray crystallography it proved to have a familiar fold: the triosephosphate isomerase alpha/beta barrel (Figure 4-59). This is yet another clear illustration of the fact that a protein fold can be encoded by very divergent sequences. Structural comparison between this three-dimensional fold and all other folds in the structural database shows the greatest similarity with the large domain of the bacterial enzyme alanine racemase. Like alanine racemase, the structure of YBL036c also revealed a covalently bound pyridoxal phosphate cofactor, which accounts for the yellow color of the purified protein.

Comparison of the active sites of alanine racemase and YBL036c revealed both similarities and important differences. The essential lysine residue required of all pyridoxal-phosphate-dependent enzymes is present, covalently linked to the cofactor (Figure 4-60). A second interaction between the protein and the cofactor that is diagnostic for the chemical function of bacterial alanine racemase, an arginine interacting with the pyridine nitrogen of the cofactor, is also present in YBL036c. However, there were several significant differences. Alanine racemase is an obligatory dimer: residues from both subunits contribute to each other's active sites. In addition, alanine racemase has a second domain, which also contributes residues to the active site. YBL036c is a monomer and lacks the second domain entirely. Consequently, a number of residues found in the active site of alanine racemase are not present in the active site of YBL036c, raising the question of whether the biochemical function of alanine racemase has been preserved.

At this point, sequence and structure can tell us nothing further about function. Additional experimental approaches are needed (see Figure 4-1). The purified protein was first assayed to see if it had any alanine racemase activity, but it did not. More general methods of determining function must be tried. (Several of these are briefly described in section 4-4.)

Location of a protein within the cell is often informative about the cellular, if not biochemical, function. Sometimes, clues to location can be found in the sequence: for example, a carboxy-terminal KDEL sequence codes for retention of a protein in the endoplasmic reticulum. Other

References

Brent, R. and Finley, R.L. Jr.: **Understanding gene and allele function with two-hybrid methods.** *Annu. Rev. Genet.* 1997; **31**:663–704.

Eswaramoorthy, S. *et al.*: **Structure of a yeast hypothetical protein selected by a structural genomics approach.** *Acta Crystallogr. D. Biol. Crystallogr.* 2003, **59**:127–135.

Shaw, J.P. *et al.*: **Determination of the structure of alanine racemase from *Bacillus stearothermophilus* at 1.9-Å resolution.** *Biochemistry* 1997, **36**:1329–1342.

Templin, M. F. *et al.*: **Protein microarray technology.** *Trends Biotechnol.* 2002, **20**:160–166.

Tucker, C.L. *et al.*: **Towards an understanding of complex protein networks.** *Trends Cell Biol.* 2001, **11**:102–106.

van Roessel, P. and Brand, A.H.: **Imaging into the future: visualizing gene expression and protein interactions with fluorescent proteins.** *Nat. Cell Biol.* 2002, **4**:E15–E20.

von Mering, C. *et al.*: **Comparative assessment of large-scale data sets of protein–protein interactions.** *Nature* 2002, **417**:399–403.

Web site:
http://genome-www.stanford.edu/Saccharomyces/

sequence motifs specify transport into the nucleus, secretion from the cell, and so forth. YBL036c has no such motifs in its sequence, so more direct methods of determining localization must be used. The most common of these is the fusion of the protein of interest with a protein that can be visualized in the cell by antibody staining or intrinsic fluorescence. Fusion to green fluorescent protein (GFP), originally isolated from jellyfish, is a widely used strategy. Efforts are underway to apply these methods systematically to all the gene products in the yeast genome. By the GFP method, YBL036c was found distributed throughout the cell.

In eukaryotic cells in particular, the function of every protein is likely to depend in some manner on interaction with one or more other proteins. Demonstrating a physical interaction between two proteins can thus provide a clue to cellular or biochemical function if the function of one of them is known. Several different approaches to discovering such interactions have been developed. These include co-immunoprecipitation or affinity chromotography from cell extracts followed by mass spectroscopy to identify the interacting partner(s), and cell-based methods such as the yeast two-hybrid screen (see section 4-4). Application of the two-hybrid method to YBL036c detected one interacting partner, another protein of unknown function.

Proteins function in regulatory networks inside the cell; their expression patterns change with changes in external and internal conditions and proteins that perform similar functions often display similar patterns of expression. Thus, clues to the function of a gene product can come from analysis of the pattern of its expression under different conditions and comparison with the patterns of other proteins of known function. Two widely used methods for studying expression are two-dimensional gel electrophoresis, which measures protein levels directly, and DNA microarrays, which measure levels of mRNA. Both can be applied to whole genomes or subsets of the genome. Microarray analyses of yeast gene expression have been carried out by many different laboratories under hundreds of different conditions; most contain information about YBL036c. In general, YBL036c is expressed in all stages of the cell cycle and in all growth conditions tested. Its expression is broadly the same as that of a number of genes that code for proteins involved in amino-acid metabolism. Its expression is upregulated slightly in a variety of stress conditions.

If a gene is not essential for the survival of an organism, deleting it from the genome can often give rise to a phenotype suggestive of function. Microarray analysis of such a deletion strain should show changes in the expression of genes whose function is in some manner coupled with that of the gene that has been deleted. YBL036c is not an essential gene in yeast: the deletion strain is viable and shows no growth defect under a variety of conditions. However, there is a subtle phenotype when the yeast cells form spores. Instead of being dispersed, the spores clump together. Electron microscopy of the spore shows that the ascus containing the spores and the spores themselves have an abnormal wall structure. Since sporulation requires remodeling of cell-wall structures, this phenotype implies that YBL036c is involved in this process. Microarray analysis of the deletion strain supports this conclusion: genes involved in cell-wall biosynthesis show changes in expression levels when YBL036c is absent.

Although these genome-wide methods have suggested a cellular process in which YBL036c participates, a great deal of information is still needed to fully describe the workings of this gene product within the cell. The active-site architecture suggests that the protein may be an amino-acid racemase, but the substrate has yet to be identified. If YBL036c produces a D-amino acid, as such activity would indicate, the role of this product in cell-wall structure remains to be determined. As animal cells do not have cell walls, the cellular function of YBL036c homologs in those organisms might be somewhat different. It is clear from this example that the task of determining the function(s) of a gene is one that does not end with a single organism.

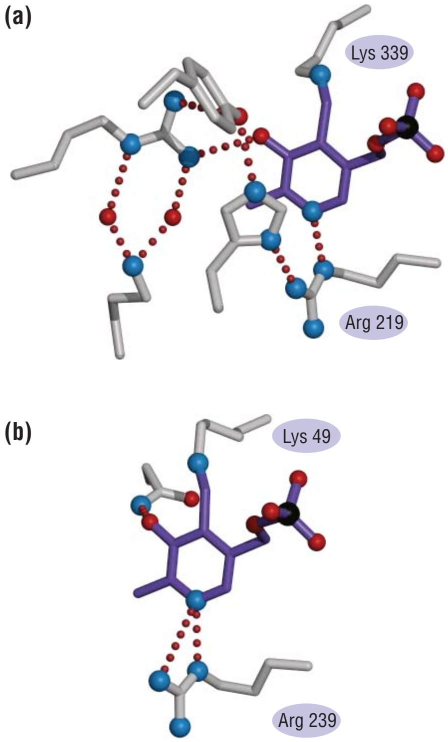


Figure 4-60 Comparison of the active sites of bacterial alanine racemase and YBL036c **(a)** Alanine racemase; **(b)** YBL036c. Although many of the interactions between the pyridoxal phosphate cofactor (shown in purple) and protein side chains are different in the two active sites, two interactions are preserved: a covalent linkage to a lysine residue, and the interaction of a nitrogen atom in the pyridine ring of the cofactor with an arginine residue. The lysine interaction is diagnostic of all pyridoxal-phosphate-dependent enzymes; the interaction with arginine is diagnostic for pyridoxal-phosphate-dependent amino-acid racemases.