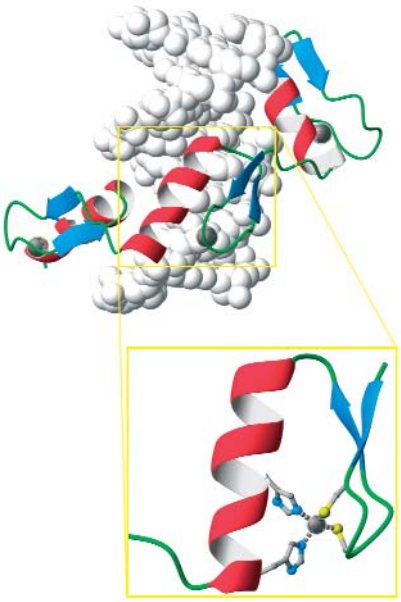
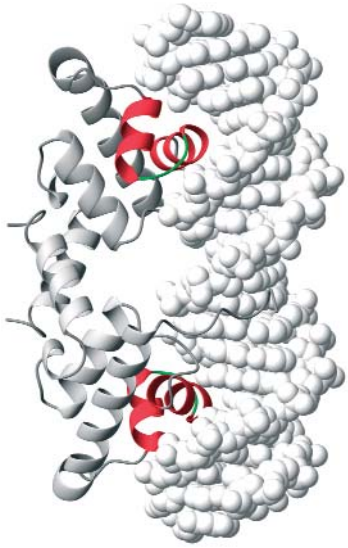


Figure 1-49 Zinc finger motif A fragment derived from a mouse gene regulatory protein is shown, with three zinc fingers bound spirally in the major groove of a DNA molecule. The inset shows the coordination of a zinc atom by characteristically spaced cysteine and histidine residues in a single zinc finger motif. The image is of Zif268. (PDB 1aay)

Protein motifs may be defined by their primary sequence or by the arrangement of secondary structure elements

The term **motif** is used in two different ways in structural biology. The first refers to a particular amino-acid sequence that is characteristic of a specific biochemical function. An example is the so-called zinc finger motif, CXX(XX)CXXXXXXXXXXXXHXXXH, which is found in a widely varying family of DNA-binding proteins (Figure 1-49). The conserved cysteine and histidine residues in this **sequence motif** form ligands to a zinc ion whose coordination is essential to stabilize the tertiary structure. Conservation is sometimes of a class of residues rather than a specific residue: for example, in the 12-residue loop between the zinc ligands, one position is preferentially hydrophobic, specifically leucine or phenylalanine. Sequence motifs can often be recognized by simple inspection of the amino-acid sequence of a protein, and when detected provide strong evidence for biochemical function. The protease from the human immunodeficiency virus was first identified as an aspartyl protease because a characteristic sequence motif for such proteases was recognized in its primary structure.

The second, equally common, use of the term motif refers to a set of contiguous secondary structure elements that either have a particular functional significance or define a portion of an independently folded domain. Along with the functional sequence motifs, the former are known generally as **functional motifs**. An example is the helix-turn-helix motif found in many DNA-binding proteins (Figure 1-50). This simple **structural motif** will not exist as a stably folded domain if expressed separately from the rest of its protein context, but when it can be detected in a protein that is already thought to bind nucleic acids, it is a likely candidate for the recognition element. Examples of structural motifs that represent a large part of a stably folded domain include the four-helix bundle (Figure 1-51), a set of four mutually antiparallel alpha helices that is found in many hormones as well as other types of proteins; the Rossmann fold, an alpha/beta twist arrangement that usually binds NAD cofactors; and the *Greek-key motif*, an all-beta-sheet arrangement found in many different proteins and which topologically resembles the design found on ancient vases. As these examples indicate, these structural motifs sometimes are suggestive of function, but more often are not: the only case here with clear functional implications is the Rossmann fold.



Identifying motifs from sequence is not straightforward

Because motifs of the first kind—sequence motifs—always have functional implications, much of the effort in bioinformatics is directed at identifying these motifs in the sequences of newly discovered genes. In practice, this is more difficult than it might seem. The zinc finger motif is always uninterrupted, and so is easy to recognize. But many other sequence motifs are discontinuous, and the spacing between their elements can vary considerably. In such cases, the term sequence motif is almost a misnomer, since not only the spacing between the residues but also the order in which they occur may be completely different. These are really functional motifs whose presence is detected from the structure rather than the sequence. For example, the "catalytic triad" of the serine proteases (Figure 1-52), which consists of an aspartic acid, a histidine and a serine, all interacting with one another, comprises residues aspartic acid 102, histidine 57

Figure 1-50 Helix-turn-helix The DNA-binding domain of the bacterial gene regulatory protein lambda repressor, with the two helix-turn-helix motifs shown in color. The two helices closest to the DNA are the reading or recognition helices, which bind in the major groove and recognize specific gene regulatory sequences in the DNA. (PDB 1lmb)

Definitions

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

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

motif: characteristic sequence or structure that in the case of a **structural motif** may comprise a whole domain or protein but usually consists of a small local arrangement of secondary structure elements which then coalesce to form domains. **Sequence motifs**, which are recognizable amino-acid sequences found in different proteins, usually indicate biochemical function. Structural motifs are less commonly associated with specific biochemical functions.

©2004 New Science Press Ltd

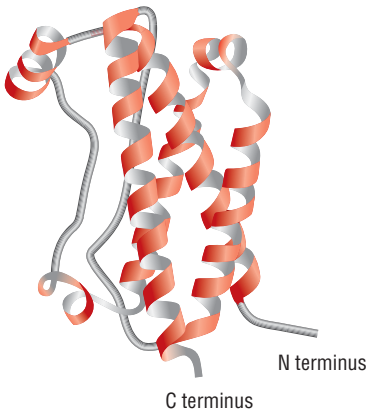


Figure 1-51 Four-helix bundle motif The four-helix bundle motif can comprise an entire protein domain, and occurs in proteins with many different biochemical functions. Shown here is human growth hormone, a signaling molecule; shown in Figure 1-28a is cytochrome b562, an electron-transport protein. In Figure 1-54 the protein myohemerythrin is shown; its function is oxygen transport.

and serine 195 in one family of serine proteases. However, in another, unrelated family of serine proteases, the same triad is made up by aspartic acid 32, histidine 64, and serine 221 (see Figure 4-35). This is a case in which both the spacing between the residues that define the motif and the order in which they occur in the primary sequence are different. Nevertheless, these residues form a catalytic unit that has exactly the same geometry in the two proteases, and that carries out an identical chemical function. This is an example of **convergent evolution** to a common biochemical solution to the problem of peptide-bond hydrolysis. One of the major tasks for functional genomics is to catalog such sequence-based motifs, and develop methods for identifying them in proteins whose overall folds may be quite unrelated.

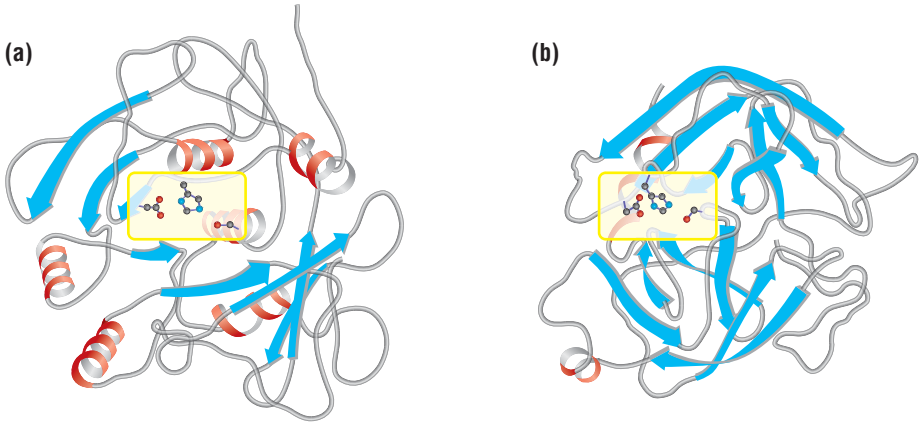


Figure 1-52 Catalytic triad The catalytic triad of aspartic acid, histidine and serine in **(a)** subtilisin, a bacterial serine protease, and **(b)** chymotrypsin, a mammalian serine protease. The two protein structures are quite different, and the elements of the catalytic triad are in different positions in the primary sequence, but the active-site arrangement of the aspartic acid, histidine and serine is similar.

Identifying structural motifs from sequence information alone presents very different challenges. First, as we have seen, many different amino-acid sequences are compatible with the same secondary structure; so there may be literally hundreds of different unrelated sequences that code for four-helix bundles. Sequence similarity alone, therefore, cannot be used for absolute identification of structural motifs. Hence, such motifs must be identified by first locating the secondary structure elements of the sequence. However, secondary structure prediction methods are not completely accurate, as pointed out earlier. Second, a number of structural motifs are so robust that large segments of additional polypeptide chain, even specifying entire different domains, can sometimes be inserted into the motif without disrupting it structurally. A common example is the so-called TIM-barrel domain, which consists of a strand of beta sheet followed by an alpha helix, repeated eight times. Protein domains are known that consist of nothing but this set of secondary structure elements; others are known in which an additional structural motif is inserted; and yet others are found in which one or more additional entire domains interrupt the pattern, but without disrupting the barrel structure (Figure 1-53).

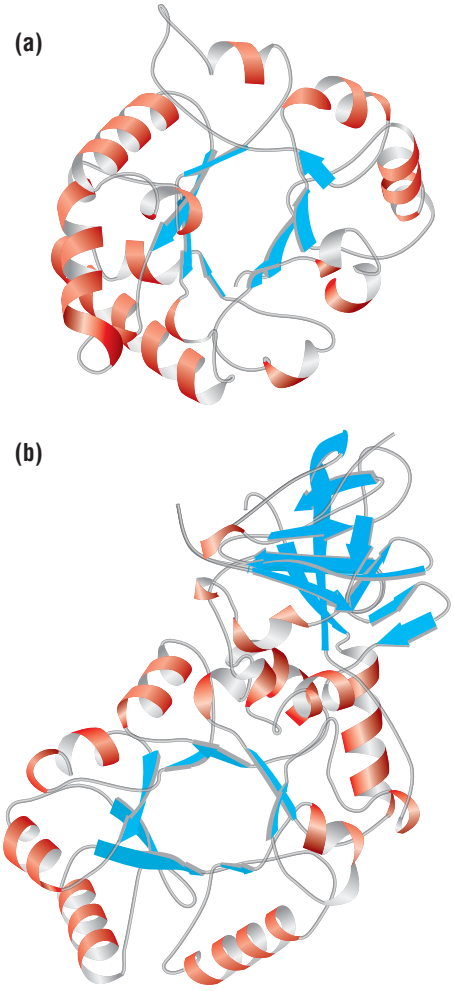


Figure 1-53 TIM-barrel proteins Triose phosphate isomerase **(a)** is shown together with alanine racemase **(b)**. In alanine racemase, the TIM-barrel domain is interrupted by an inserted domain.

References

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

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

Ponting, C.P. *et al.*: **Evolution of domain families.** *Adv. Protein Chem.* 2000, **54**:185–244.

©2004 New Science Press Ltd