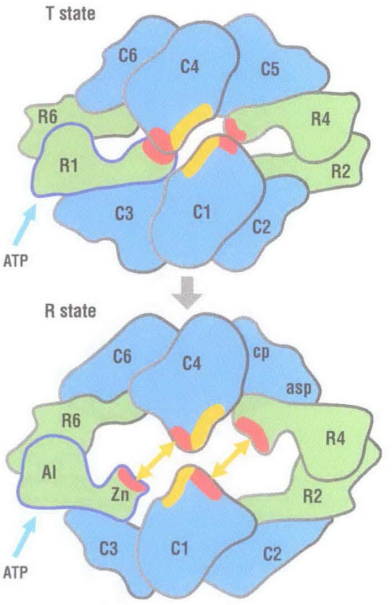


Figure 3-10 Ligand-induced conformational change activates aspartate transcarbamoylase
Binding of the allosteric activator ATP to its intersubunit binding sites on the regulatory subunits (that between R1, outlined in purple, and R6 is arrowed) of the T state of ATCase (top) causes a massive conformational change of the enzyme to the R state (bottom). In this state the structure of the enzyme is opened up, making the active sites on the catalytic subunits (C) accessible to substrate. Al and Zn in the lower diagram indicate the allosteric regions and the zinc-binding region, respectively; cp and asp indicate the binding sites for the substrates carbamoyl phosphate and aspartate, respectively. The red and yellow regions are the intersubunit interfaces that are disrupted by this allosteric transition.

Disruption of function does not necessarily mean that the active site or ligand-binding site has been disrupted

As we have seen, most proteins are like machines: they have moving parts. And like a machine, the introduction of a monkey wrench into any part of the protein can disrupt any activity that depends on the propagation of a conformational change. The monkey wrench can be a drug that binds directly to the active site or functional ligand-binding site, or a mutation at these sites. But drugs or mutations may also disrupt a protein's function by binding elsewhere and interfering with the conformational transitions necessary for function. For example, binding of anti-viral compounds to the rhinovirus coat protein blocks entry of the virus into host cells. But these drugs do not bind to the site on the virus that binds the cell-surface virus receptor; instead they bind to an unrelated site where they stabilize the structure of the coat protein, thereby preventing the structural rearrangements required for receptor-mediated virus entry into the cell.

The effects of both regulatory ligand binding and mutation are illustrated by ATCase. The T state can be stabilized by the binding of CTP to a site on the regulatory subunit. Exactly the same stabilizing effect can be achieved by mutation of tyrosine 77 to phenylalanine in the regulatory subunit. This mutation stabilizes the T state and thus shuts down catalytic activity, even though it is very far from the active site. Therefore, in the absence of structural information, one should not automatically assume that any mutation that disrupts a protein's function must be in the active site or that any molecule that inhibits function does so because it binds directly in the active site.

Binding of gene regulatory proteins to DNA is often controlled by ligand-induced conformational changes

DNA-binding proteins are not usually intended to interact with the genetic material all the time. Activators and repressors of gene expression are usually under the control of specific regulatory ligands, which may be small molecules, metal ions or proteins, whose binding determines whether or not the activator or repressor can bind to DNA. Such molecules are called **co-activators** and **co-repressors**, respectively. Expression of the gene encoding diphtheria toxin is under the control of a specific repressor, DtxR. Binding of DtxR to its operator sequence is controlled in turn by the concentration of Fe^{2+} in the bacterial cell. Iron acts as a co-repressor by binding to DtxR and inducing a conformational change that allows the helix-turn-helix DNA-recognition motif in the repressor to fit into the major groove of DNA. In the absence of bound iron, the repressor adopts a conformation in which DNA binding is sterically blocked (Figure 3-11).

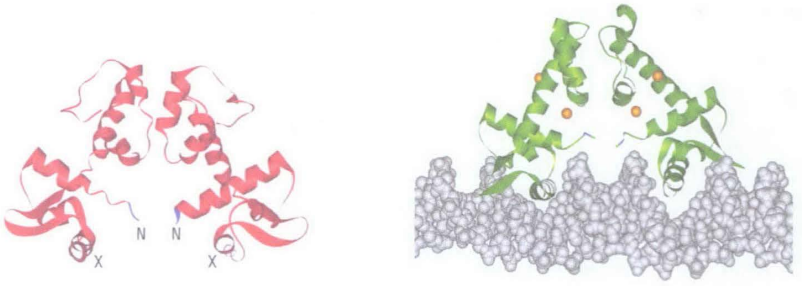


Figure 3-11 Iron binding regulates the repressor of the diphtheria toxin gene
Comparison of the structures of the aporepressor DtxR (red, left, PDB 1dpr) and the ternary complex (right) of repressor (green), metal ion (Fe^{2+}, orange) and DNA (grey) (PDB 1fst). Iron binding induces a conformational change that moves the recognition helices (X) in the DtxR dimer closer together, providing an optimal fit between these helices and the major groove of DNA. In addition, metal-ion binding changes the conformation of the amino terminus of the first turn of the amino-terminal helix (N) of each monomer. Without this conformational change, leucine 4 in this helix would clash with a phosphate group of the DNA backbone. Thus, DtxR only binds to DNA when metal ion is bound to the repressor.

©2004 New Science Press Ltd

Halozyme EX2164 (part 3 of 4)
Merck v. Halozyme
PGR2025-00039

3-6 Protein Switches Based on Nucleotide Hydrolysis

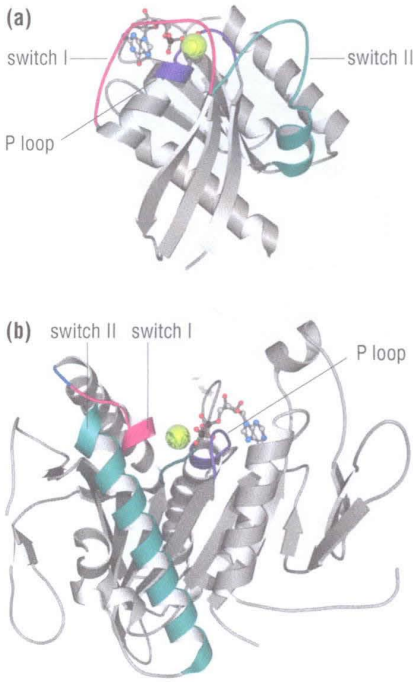


Figure 3-12 Structure of the core domains of a typical GTPase and an ATPase (a) Ribbon diagram of the minimal GTPase G domain, with the conserved sequence elements and the switch regions in different colors. The nucleoside diphosphate is shown in ball-and-stick representation and the bound Mg^{2+} ion as a green sphere. The switch I region is the pink loop that surrounds the ligand in this view. The switch II region is in turquoise and the P-loop in purple. (PDB 4q21) **(b)** Structure of the ATPase domain of the mitotic spindle kinesin Eg5. As in the GTPase, one molecule of Mg-ADP is located in the nucleotide-binding site. The equivalent switch I, switch II and P-loop regions are colored the same as for the GTPase. (PDB 1ii6)

Conformational changes driven by nucleotide binding and hydrolysis are the basis for switching and motor properties of proteins

Not every process in a living cell runs continuously. Many of them must cycle between "on" and "off" states in order to control cell growth and division and responses to extracellular signals. Signal transduction pathways such as those that operate in vision and hormone-based signaling, vesicular transport (which is often called protein or membrane trafficking), polypeptide chain elongation during protein synthesis, and actin- and tubulin-based motor functions are also examples of processes that must be switched on and off under precisely determined circumstances. This cycling is controlled by a special set of proteins that function as molecular switches. Although these proteins vary in structure and in the processes they control, they have a number of common features. The most important common element is the switching mechanism itself: most of these proteins undergo conformational changes induced by the difference between the triphosphate and diphosphate forms of a bound nucleotide. The conformational changes are such that completely different target proteins recognize the two bound states of the switch protein, providing a simple means of altering the output of a signal.

Most protein switches are enzymes that catalyze the hydrolysis of a nucleoside triphosphate to the diphosphate. Most often the nucleotide is guanosine triphosphate, GTP, and the switch protein is a GTPase that hydrolyzes it to GDP. GTPase switches (also commonly called **G proteins** or occasionally **guanine-nucleotide-binding proteins**) are one major class of switch proteins; they control the on/off states of most cellular processes, including sensory perception, intracellular transport, protein synthesis and cell growth and differentiation. The second major class of switch proteins is composed of those ATPases that are usually associated with motor protein complexes or transporters that move material into and out of cells and some organelles. Members of the third major class, the two-component response regulators, a group of switches thus far found only in microbes and plants, are composed of a histidine protein kinase and a second "response regulator" protein. They do not bind GTP or ATP in the same way as the nucleotide switch proteins discussed here, but use a covalently bound phosphate derived from the hydrolysis of ATP by the kinase to trigger a conformational change in the response regulator.

Hydrolyzable nucleotides are used to control many types of molecular switches because the energy derived from hydrolysis of the terminal phosphate of a GTP or ATP is large enough to make the conformational change in the switch effectively irreversible until another protein binds to the switch, displaces the diphosphate, and allows the triphosphate to bind again. Another reason is that using ATP or GTP can couple the switching process to the energy state of the cell and to the synthesis of DNA and RNA, both of which change the levels of nucleoside tri- and diphosphates.

All nucleotide switch proteins have some common structural and functional features

Nucleotide-dependent switch proteins are found in every kingdom of life. The ATPases and the GTPases constitute two protein families that are both characterized by a core domain that carries out the basic function of nucleotide binding and hydrolysis—called the G domain in the GTPases. The core domain structure is conserved within each family, but its secondary structure arrangement is very different in GTPases and ATPases. Remarkably, in spite of the lack of similarity between their protein folds, both the nucleotide-binding site and the switch mechanism are extremely similar for the two families. The standard fold of the G domain in

Definitions

G protein: a member of a large class of proteins with GTPase activity that act as molecular switches in many different cellular pathways, controlling processes such as sensory perception, intracellular transport, protein synthesis and cell growth and differentiation. They undergo a large conformational change when a bound GTP is hydrolyzed to GDP.

guanine-nucleotide-binding protein: see **G protein**.

P-loop: a conserved loop in GTPase- and ATPase-based nucleotide switch proteins that binds to phosphate groups in the bound nucleotide.

phosphate-binding loop: see **P-loop**.

switch I region: a conserved sequence motif in GTPase- and ATPase-based nucleotide switch proteins that, with the **switch II region**, binds the terminal gamma-phosphate in the triphosphate form of the bound nucleotide and undergoes a marked conformational change when the nucleotide is hydrolyzed.

switch II region: a conserved sequence motif in GTPase- and ATPase-based nucleotide switch proteins that, with the **switch I region**, binds the terminal gamma-phosphate in the triphosphate form of the bound nucleotide and undergoes a marked conformational change when the nucleotide is hydrolyzed.

©2004 New Science Press Ltd

the GTPases consists of a mixed six-stranded beta sheet with five helices located on both sides (Figure 3-12a). There are three conserved features in nucleotide switches: the **P-loop**, and the **switch I** and **switch II** sequence motifs. The P-loop or **phosphate-binding loop** binds the alpha- and beta-phosphates from the phosphate tail of the nucleotide. Residues from the two switch motifs coordinate the terminal gamma-phosphate in the triphosphate form of the bound nucleotide. A Mg^{2+} ion, which is complexed with the bound nucleotide, is coordinated by the nucleotide phosphate groups and, in the triphosphate form of the switch, by residues from the switch I and II regions.

These regions usually contain four to five conserved sequence elements, which are lined up along the nucleotide-binding site. Additional important contribution to binding is made by the interactions of the nucleotide base with a sequence that has the motif N/TKXD (where X is any amino acid) and confers specificity for guanine. Specificity for guanine is due to the aspartate side chain in this motif, which forms a bifurcated hydrogen bond with the base, and to a main-chain interaction of an invariant alanine (from a short SAK motif), with the guanine oxygen.

Similar conserved sequence motifs are located in equivalent spatial positions in the switch ATPases, even though the core protein fold is completely different (Figure 3-12b). The adenine base interacts with a conserved RXRP or NP motif (equivalent to the N/TKXD in the GTPases) and alpha- and beta-phosphates are bound to an equivalent P-loop with the same consensus sequence GXXXXGKS/T. Furthermore, both switch I and switch II sequence motifs of GTPases (DX_nT and DXXG respectively) have their equivalents in ATPases: NXXSSR and DXXG, respectively.

The GTPases have been most extensively studied and most of the details of the switch mechanism have been determined for them, but the basic switch mechanism is the same for the ATPases. The trigger for the conformational change is most likely universal. The triphosphate-bound state, which is usually the "on" state of the switch, can be considered as "spring-loaded" because of the terminal phosphate group of the bound nucleotide, which makes a number of interactions with the two switch regions (Figure 3-13). It is the loss of this gamma-phosphate group on hydrolysis of GTP to GDP (or ATP to ADP) that provides the trigger for the conformational change.

Changes in conformation between the triphosphate- and diphosphate-bound states are confined primarily to the two switch regions. These regions usually show an increased flexibility relative to the rest of the protein when the structure of a nucleotide-dependent switch is determined. In the triphosphate-bound form, there are two hydrogen bonds from gamma-phosphate oxygens to the main-chain NH groups of the invariant threonine (or serine in the case of the ATPases) and glycine residues in the switch I and II regions, respectively (see Figure 3-13). The glycine is part of the conserved DXXG motif; the threonine (or serine) is from the conserved sequence involved in binding the Mg^{2+} ion that coordinates to the two terminal phosphate groups of the bound nucleotide. Release of the gamma-phosphate after hydrolysis allows the two switch regions to relax into their diphosphate-specific conformations because the bonds that the threonine/serine and glycine residues make with this phosphate are disrupted when it leaves. The extent of the conformational change that results from the rearrangement of the two switch regions is different for different proteins and involves additional structural elements in some of them.

Although the fold of the core domain is conserved, there are many insertions of other domains in individual GTPases. These domains have various functions, from interacting with other proteins to magnifying the conformational change on nucleotide hydrolysis.

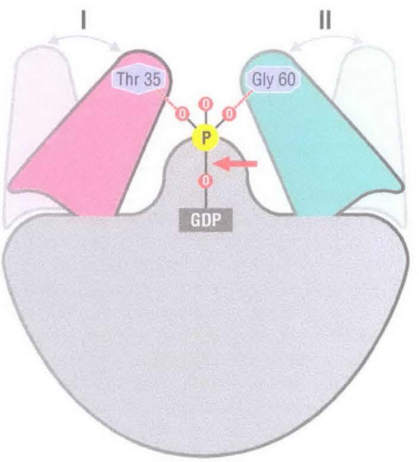


Figure 3-13 Schematic diagram of the universal switch mechanism of GTPases This picture illustrates the GTP-bound state where the switch I (pink) and switch II (turquoise) domains are bound to the gamma-phosphate via the main chain NH groups of the invariant threonine and glycine residues, in a spring-loaded mechanism. Release of the gamma-phosphate after GTP hydrolysis allows the switch regions to relax into a different conformation, as seen in Figure 3-12, which shows the diphosphate-bound forms. Adapted from Vetter, I.R. and Wittinghofer, A.: *Science* 2001, **294**:1299–1304.

References

Sablin, E.P. and Fletterick, R.J.: **Nucleotide switches in molecular motors: structural analysis of kinesins and myosins.** *Curr. Opin. Struct. Biol.* 2001, **11**:716–724.

Spudich, J.A.: **The myosin swinging cross-bridge model.** *Nat. Rev. Mol. Cell Biol.* 2001, **2**:387–392.

Vetter, I.R. and Wittinghofer, A.: **The guanine nucleotide-binding switch in three dimensions.** *Science* 2001, **294**:1299–1304.

An animation showing the conformational changes in the switch regions of Ras can be found at: http://www.mpi-dortmund.mpg.de/departments/dep1/gtpase/ras_gtp_gdp.gif

©2004 New Science Press Ltd

The switching cycle of nucleotide hydrolysis and exchange in G proteins is modulated by the binding of other proteins

All GTPase switches operate through conformational changes induced by a change in the form of the bound nucleotide, as we saw in the previous section; but this is only the core part of the switching mechanism. Which cellular process a particular G protein controls is determined by the interactions it makes with other proteins in its two conformational states. We will now consider the rest of the switching mechanism in detail in the case of the GTPases, exemplified by the small monomeric GTPase Ras, whose switch function helps to control cell growth and division.

The overall switching mechanism can be viewed as an on–off cycle in which the GTP-bound state is the "on" state and the GDP-bound state is "off". If we start with the switch in the off position, the gamma-phosphate of GTP is not present and the two switch regions are in the relaxed conformations characteristic of the diphosphate-bound state of the protein (see Figure 3-12a). GDP must now dissociate from the protein to allow GTP to bind. The normal rate of dissociation is often very slow, so most G proteins have various **guanine-nucleotide exchange factors** (**GEFs**) that bind to them and facilitate the release of GDP by inducing conformational changes that open up the binding site. A single G protein may be recognized by multiple GEFs, enabling one GTPase to serve as the focal point for integrating signals from different upstream pathways.

After GDP has been released and GTP binds, the protein conformation (that is, the arrangement of the switch I and switch II regions) changes to bind the gamma-phosphate (see Figure 3-13) and the switch is now in the on position. The altered conformation of the switch regions allows the G protein to interact with downstream effectors, activating various enzymes such as phosphatidylinositol-3-kinase and turning on various signaling and other pathways. Several different effectors may recognize the on state of the same G protein, allowing a single switch to control multiple cellular processes. The switch remains on as long as GTP remains bound to the GTPase.

Although GTPases are enzymes and catalyze the hydrolysis of GTP to GDP, they are not very efficient; their intrinsic rate of GTP hydrolysis is very slow. What determines the length of time the switch remains in the on state is the activity of various **GTPase-activating proteins** (**GAPs**) that bind to the GTP-bound conformation and stimulate the catalytic activity: in the case of Ras the binding of a GAP increases the GTPase activity by 100,000-fold (10^5). When GTP has been hydrolyzed, the switch I and switch II regions change to their relaxed conformations and the switch is back in the off state; the cycle is complete. Again, a single G protein can interact with multiple GAPs from different upstream signaling pathways. The on–off cycle is illustrated in Figure 3-14.

How the various GAP proteins facilitate GTP hydrolysis is not known in all cases, but at least some appear to function by stabilizing the transition state (see section 2-9). They insert an arginine side chain into the nucleotide-binding site of the GTPase to which they bind, and the positive charge on this side chain helps to stabilize the negative charge that builds up in the transition state for hydrolysis of the gamma-phosphate group of GTP.

The importance of proper regulation of G-protein signaling is exemplified by the Ras family of small GTPases. There are three human *RAS* genes, *H-*, *N-* and *K-RAS*, all of which code for very similar proteins of around 21 kDa molecular mass. They are post-translationally modified by the covalent attachment of lipophilic groups to the carboxy-terminal end. This modification is necessary for the Ras proteins' biological function as switches because it targets them to the

Definitions

GTPase-activating protein (GAP): a protein that accelerates the intrinsic GTPase activity of switch GTPases.

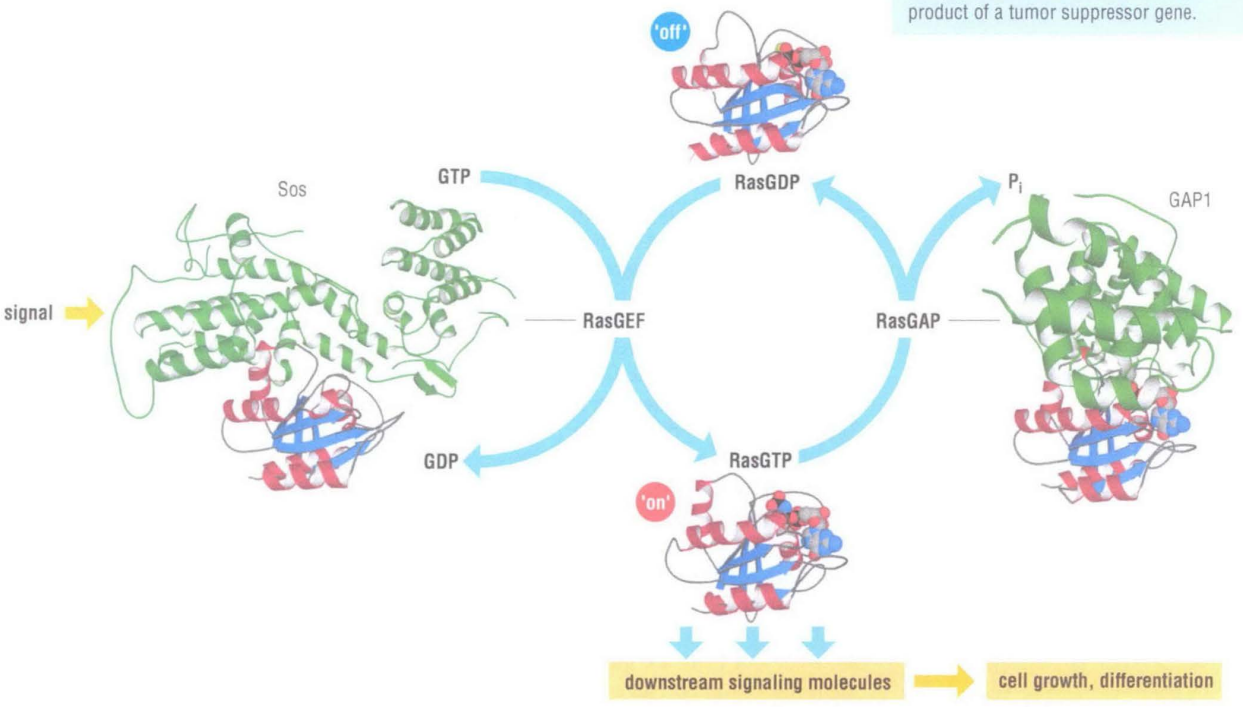
guanine-nucleotide exchange factor (GEF): a protein that facilitates exchange of GDP for GTP in switch GTPases.

©2004 New Science Press Ltd

plasma membrane, where their interacting protein partners are found (see Figure 3-4). Interestingly, the pathways controlled by Ras seem to differ in mouse, humans and yeast. Consequently, mutations in the corresponding genes lead to different phenotypes in different organisms. This is also true for some other signal transduction pathways, and it complicates the generalization of data obtained from model organisms such as the mouse.

The lifetime of the signal transduced by Ras is determined by the lifetime of the GTP-bound state. If it is artificially prolonged, the biological response may be unregulated and lead to drastic consequences in the cell. The gene coding for Ras was originally discovered as an oncogene of rodent tumor viruses, the gene responsible for their ability to cause cancer in animals. Mutant forms of the human *RAS* genes that produce a protein with a prolonged on state are found in up to 30% of human tumors. This activated Ras protein is the result of point mutations at amino acids 12, 13 or 61, the biochemical consequence of which is to reduce the rate of hydrolysis of GTP. This leaves Ras in the on state too long, leading to uncontrolled cell growth and proliferation. As many other genes involved in the Ras signal transduction pathway are also found as oncogenes in human or animal tumors, Ras itself and the Ras pathway is considered to be a prime target for anti-tumor therapy.

Figure 3-14 The switching cycle of the GTPases involves interactions with proteins that facilitate binding of GTP and stimulation of GTPase activity The cycle is illustrated here for the small G protein Ras, but the principles are the same for other GTPase switches such as protein synthesis elongation factors and the heterotrimeric G proteins, which are discussed later in this chapter. GTP-binding proteins function as molecular switches by cycling between GDP-bound "off" and GTP-bound "on" states. Exchange of the bound GDP for GTP is facilitated by guanine-nucleotide exchange factors (GEFs) whose binding to the GTPase increases the dissociation rate of the nucleotide by several orders of magnitude. Ras can be switched on by several GEFs, the most important of which are Sos (illustrated here), Ras-GRF and Ras-GRP. Sos is part of the pathway that conveys signals from an activated cell-surface receptor to Ras. In the activated GTP-bound state Ras interacts with and activates several target proteins involved in intracellular signaling pathways. Ras is switched off by hydrolysis of the bound GTP. This reaction is facilitated by the action of specific GTPase-activating proteins (GAPs), the best studied of which are GAP1 (illustrated here), p120GAP and neurofibromin, the product of a tumor suppressor gene.



References

Geyer, M. and Wittinghofer A.: **GEFs, GAPs, GDIs and effectors: taking a closer (3D) look at the regulation of Ras-related GTP-binding proteins.** *Curr. Opin. Struct. Biol.* 1997, **7**:786–792.

Hamad, N.M. *et al.*: **Distinct requirements for Ras oncogenesis in human versus mouse cells.** *Genes Dev.* 2002, **16**:2045–2057.

Scheffzek, K. *et al.*: **GTPase-activating proteins: helping hands to complement an active site.** *Trends Biochem. Sci.* 1998, **23**:257–262.

Spoerner, M. *et al.*: **Dynamic properties of the Ras switch I region and its importance for binding to effectors.** *Proc. Natl Acad. Sci. USA* 2001, **98**:4944–4949.

Sprang, S.R.: **G protein mechanisms: insights from structural analysis.** *Annu. Rev. Biochem.* 1997, **66**:639–678.

Takai, Y. *et al.*: **Small GTP-binding proteins.** *Physiol. Rev.* 2001, **81**:153–208.

Vetter, I.R. and Wittinghofer, A.: **Nucleoside triphosphate-binding proteins: different scaffolds to achieve phosphoryl transfer.** *Q. Rev. Biophys.* 1999, **32**:1–56.

Wittinghofer, A.: **Signal transduction via Ras.** *Biol. Chem.* 1998, **379**:933–937.

©2004 New Science Press Ltd

3-8 GTPase Switches: Signal Relay by Heterotrimeric GTPases

Heterotrimeric G proteins relay and amplify extracellular signals from a receptor to an intracellular signaling pathway

Most GTPases are either small monomeric proteins like Ras or much larger, heterotrimeric molecules, **heterotrimeric G proteins**, which are composed of an α subunit, whose core polypeptide chain fold is the same as the canonical G-domain fold, and β and γ subunits whose folds bear no relationship to that of the GTPases (Figure 3-15). The α subunit has an extra helical domain compared with the canonical G domain, and this partially projects into the catalytic site. The β and γ subunits are tightly associated with one another, partly by means of a coiled-coil interaction. Heterotrimeric G proteins are associated with the cytoplasmic surface of the cell membrane in complexes with so-called G-protein-coupled receptors (GPCRs), integral membrane proteins with seven transmembrane alpha helices. Affinity of heterotrimeric G proteins for the membrane is aided by prenylation of the β and γ subunits. In general, GPCRs are activated as a result of the binding of specific extracellular ligands to their extracellular or transmembrane domains. Conformational changes in the receptor induced by ligand binding are relayed through the transmembrane domain to the cytoplasmic portion, allowing productive coupling of the receptor with a heterotrimeric G protein, which transduces the signal and relays it onwards by means of its switching function.

The switching cycle of the heterotrimeric G proteins resembles that of the small monomeric GTPases with additional features imparted by the presence of the β and γ subunits. Heterotrimeric G proteins bind to the cytoplasmic domains of GPCRs in their GDP-bound "off" states. When activated by binding of their extracellular ligand, these receptors act as guanine-nucleotide exchange factors (GEFs) for their partner heterotrimeric G proteins, thus triggering GDP dissociation and the GTPase switching cycle. The βγ complex acts as part of the nucleotide-exchange mechanism. In the absence of βγ, α does not bind to GPCR. In addition, βγ does not bind to α when α is in its GTP form. Consequently, βγ acts to guarantee that GPCR only catalyzes replacement of GDP by GTP and not the other way around.

When GDP is released and GTP binds, the heterotrimeric G protein dissociates from the GPCR and the β and γ subunits dissociate as a heterodimer from the α subunit. Both the free GTP-bound α subunit and the βγ heterodimer can now bind to and stimulate their own respective downstream effectors, which are generally ion channels and enzymes. Hydrolysis of GTP in the active site of the free α subunit then results in a conformational change in the switch I and switch II regions, which restores association of α to the βγ heterodimer and causes the reassembled heterotrimeric G protein to rebind to the GPCR, completing the cycle.

Just as GAP proteins switch the state of small monomeric GTPases from GTP-bound to GDP-bound by increasing the GTPase catalytic rate, so-called **regulator of G-protein signaling proteins (RGS proteins)** are responsible for the rapid turnoff of GPCR signaling pathways by functioning as activators of the heterotrimeric G protein GTPase. Structural and mutational analyses have characterized the interaction of the RGS domain of these proteins with Gα in detail. Unlike RasGAPs, the RGS proteins do not directly contribute an arginine residue or any other catalytic residue to the active site of the α subunit to assist GTP hydrolysis. In fact, the α subunit of most heterotrimeric G proteins has a "built-in" arginine residue in the extra helical domain that projects into the catalytic site. RGS proteins probably exert their GAP activity mainly through binding to the switch regions, reducing their flexibility and stabilizing the transition state for hydrolysis in that way. More than 20 different RGS proteins have been isolated, and there are indications that particular RGS proteins regulate particular GPCR signaling pathways. This specificity is probably created by a combination of cell-type-specific

Definitions

heterotrimeric G protein: a GTPase switch protein composed of three different subunits, an α subunit with GTPase activity, and associated β and γ subunits, found associated with the cytoplasmic tails of G-protein-coupled receptors, where it acts to relay signals from the receptor to downstream targets. Exchange of bound GDP for GTP on the α subunit causes dissociation of the heterotrimer into a free α subunit and a βγ heterodimer; hydrolysis of the bound GTP causes reassociation of the subunits.

RGS protein: regulator of G-protein signaling protein; protein that binds to the free GTP-bound α subunit of a **heterotrimeric G protein** and stimulates its GTPase activity.

References

Hamm, H.E. and Gilchrist, A.: **Heterotrimeric G proteins.** *Curr. Opin. Cell Biol.* 1996, **8**:189–196.

Okada, T. *et al.*: **Activation of rhodopsin: new insights from structural and biochemical studies.** *Trends Biochem. Sci.* 2001, **26**:318–324.

Rodbell, M.: **The complex regulation of receptor-coupled G-proteins.** *Adv. Enzyme Regul.* 1997, **37**:427–435.

Ross, E.M. and Wilkie, T.M.: **GTPase-activating proteins for heterotrimeric G proteins: regulators of G protein signaling (RGS) and RGS-like proteins.** *Annu. Rev. Biochem.* 2000, **69**:795–827.

©2004 New Science Press Ltd

expression, tissue distribution, intracellular localization, post-translational modifications, and domains other than the RGS domain that link these RGS proteins to other signaling pathways. Many heterotrimeric G proteins operate as complexes that include the RGS protein and the GPCR. This association causes rates of GTP hydrolysis and nucleotide exchange (the deactivation and activation rates, respectively, for the G protein) to become rapid and tightly coupled as long as the ligand for the receptor is present.

GPCRs, which are characterized by seven membrane-spanning alpha helices (Figure 3-15), are the most numerous receptors in all eukaryotic genomes (1–5% of the total number of genes). They transduce extracellular signals as varied as light, odorants, nucleotides, nucleosides, peptide hormones, lipids, and proteins. Rhodopsin, the primary photoreceptor protein in vision, is a prototypical GPCR that contains 11-*cis*-retinal as an intrinsic chromophore. The light-induced *cis–trans* isomerization of this ligand is the primary signaling event in the visual process. It is followed by slower and incompletely defined structural rearrangements leading to an "active state" intermediate of rhodopsin that signals via the heterotrimeric G protein transducin.

As well as relaying the signal onwards, heterotrimeric G proteins can also be one means by which it is amplified. It has been estimated that one rhodopsin molecule activated by a single photon of light activates around 500 molecules of transducin. Each of these activates a molecule of its target phosphodiesterase.

There are at least eight families of GPCRs that show no sequence similarities to each other and that have quite different ligand-binding domains and cytoplasmic domains, but which nevertheless activate a similar set of heterotrimeric G proteins. Homo- and heterodimerization of GPCRs seems to be the rule, and in some cases an absolute requirement, for activation. It has been estimated that there are about 100 "orphan" GPCRs in the human genome for which ligands have not yet been found. Mutations of GPCRs are responsible for a wide range of genetic diseases, including some hereditary forms of blindness. The importance of GPCRs in physiological processes is further illustrated by the fact that they are the targets of the majority of therapeutic drugs and drugs of abuse.

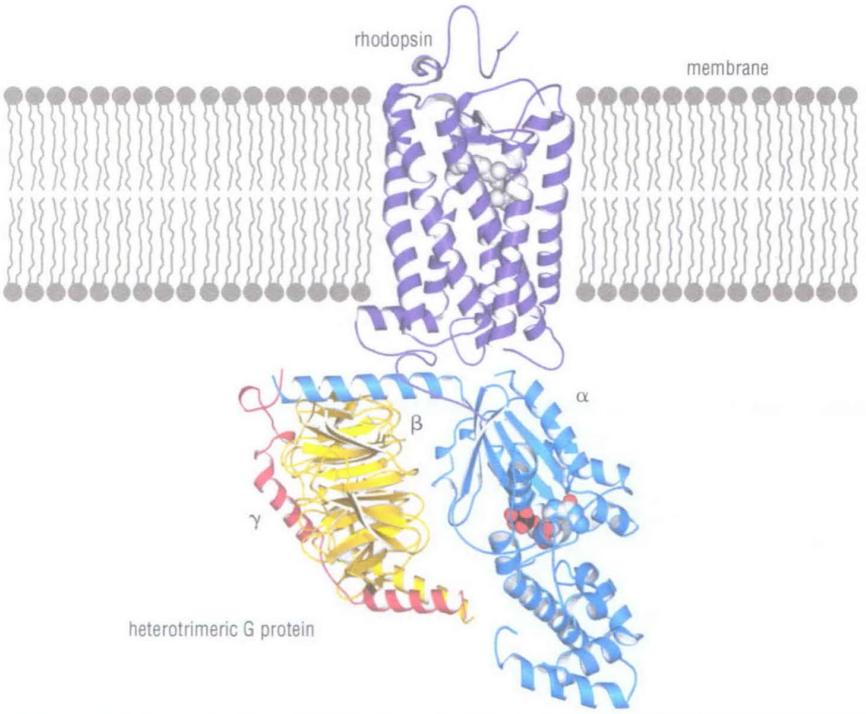


Figure 3-15 Hypothetical model of a heterotrimeric G protein in a complex with its G-protein-coupled receptor The GTPase subunit is shown in blue, the β subunit, which consists of a series of tandem WD40 domains, in yellow, and the γ subunit, which binds to the β subunit by a coiled-coil interaction, in red. The GPCR, shown here as the photoreceptor rhodopsin, is in purple embedded in a schematic of a membrane. A prenylation site on the α subunit that targets the complex to the membrane is not shown. GDP bound to the α subunit is shown in space-filling representation. Rhodopsin is activated by the photoisomerization of 11-*cis*-retinal (grey) bound to the protein. Rhodopsin has quite a short cytoplasmic tail but many other GPCRs have large cytoplasmic domains and additional extracellular domains to which extracellular ligands bind. Because the actual structure of the complex of any heterotrimeric G protein with its receptor is not known, the precise interactions between the components should not be inferred from this model. Adapted from Hamm, H.E. and Gilchrist, A.: *Curr. Opin. Cell Biol.* 1996, **8**:189–196.

©2004 New Science Press Ltd

EF-Tu is activated by binding to the ribosome, which thereby signals it to release its bound tRNA

One of the central protein components in the machinery of protein synthesis is a guanine-nucleotide-dependent switch. This molecule, called elongation factor Tu or EF-Tu in prokaryotes (the analogous factor in eukaryotes is called EF-1), consists of a core domain that has the canonical switch GTPase fold, and has two other domains that help bind a molecule of transfer RNA. Although there are many different tRNA molecules required for protein synthesis, they have the same overall structure and all can bind to the same EF-Tu. tRNA only binds to the GTP-bound state of EF-Tu; GTP hydrolysis causes dissociation of the complex.

EF-Tu performs a key function in protein synthesis: in its GTP-bound form it escorts aminoacyl-tRNAs coming into the ribosome, where its function is to facilitate codon–anticodon interactions and check their fit. If the fit is correct, this causes conformational changes in the ribosome that stabilize tRNA binding and trigger GTP hydrolysis by EF-Tu. On this, the elongation factor dissociates from the tRNA and leaves the ribosome, leaving the tRNA behind to deliver its amino acid to the growing polypeptide chain (Figure 3-16). If the pairing is incorrect, however, the codon–anticodon interaction is weaker, and this permits the aminoacyl-tRNA–EF-Tu complex to dissociate before hydrolysis of GTP and release of the tRNA can occur. Thus EF-Tu increases the ratio of correct to incorrect amino acids incorporated by providing a short delay between codon–anticodon base pairing and amino-acid incorporation.

tRNA release is driven by the same nucleotide-dependent switch mechanism that operates in the small monomeric and heterotrimeric G proteins. GTP hydrolysis causes a change in the conformations of the two switch regions, which in turn is transmitted to an alpha helix linked to one of them. The helix is part of the interface between the GTPase domain and the other two domains of EF-Tu, and when it moves the entire interface rearranges, which in turn causes a change in the conformations and relative positions of all three domains. The domains, which were closely associated with the tRNA and with each other in the GTP-bound state, now swing apart, one of them moving by almost 40 Å. This opening up of the protein structure dissociates the elongation factor from the bound tRNA.

Thus, in EF-Tu switching it is the correct pairing of the bound tRNA anticodon with the codon in the mRNA on the ribosome that acts as a GAP to promote the GTPase activity. By analogy, one expects to find the equivalent of a GEF to promote exchange of GTP for the bound GDP that remains on EF-Tu after hydrolysis. This function is provided by another protein, elongation factor Ts. When EF-Ts binds to the GDP-bound state of EF-Tu, the nucleoside diphosphate is released and GTP is able to bind. Binding of tRNA to the GTP-bound form completes the cycle (Figure 3-16). EF-Tu resembles the heterotrimeric G proteins in that GTP hydrolysis triggers a large-scale conformational change that leads to a loss of protein–protein interactions, but here the interactions are among domains in the same polypeptide chain rather than between non-identical subunits of an oligomeric protein.

The importance of EF-Tu switching in protein synthesis is underscored by the fact that dozens of antibiotics act by binding to EF-Tu and interfering with its function (although the human homolog EF-1 is similar in structure and function, it differs enough to be unaffected by these molecules). For instance, aurodox is a member of the family of kirromycin antibiotics, which inhibit protein biosynthesis by binding to EF-Tu. When aurodox is bound, the GTP-bound conformation of EF-Tu is observed, even when GDP is bound to the nucleotide-binding site.

References

Abel, K. *et al.*: **An alpha to beta conformational switch in EF-Tu.** *Structure* 1996, **4**:1153–1159.

Hogg, T. *et al.*: **Inhibitory mechanisms of antibiotics targeting elongation factor Tu.** *Curr. Protein Pept. Sci.* 2002, **3**:121–131.

Nyborg, J. and Liljas, A.: **Protein biosynthesis: structural studies of the elongation cycle.** *FEBS Lett.* 1998, **430**:95–99.

Ramakrishnan, V.: **Ribosome structure and the mechanism of translation.** *Cell* 2002, **108**:557–572.

Vogeley, L. *et al.*: **Conformational change of elongation factor Tu (EF-Tu) induced by antibiotic binding. Crystal structure of the complex between EF-Tu.GDP and aurodox.** *J. Biol. Chem.* 2001, **276**:17149–17155.

©2004 New Science Press Ltd

This suggests that aurodox fixes EF-Tu on the ribosome by locking it in its GTP form. Certain mutations in bacterial EF-Tu genes confer resistance to kirromycin by producing a protein that cannot interact with the antibiotic.

EF-Tu is not the only GTPase switch involved in translation. Another elongation factor, EF-G, is involved at a later step, in the translocation of the mRNA through the ribosome so that it is ready to accept the next aminoacyl-tRNA.

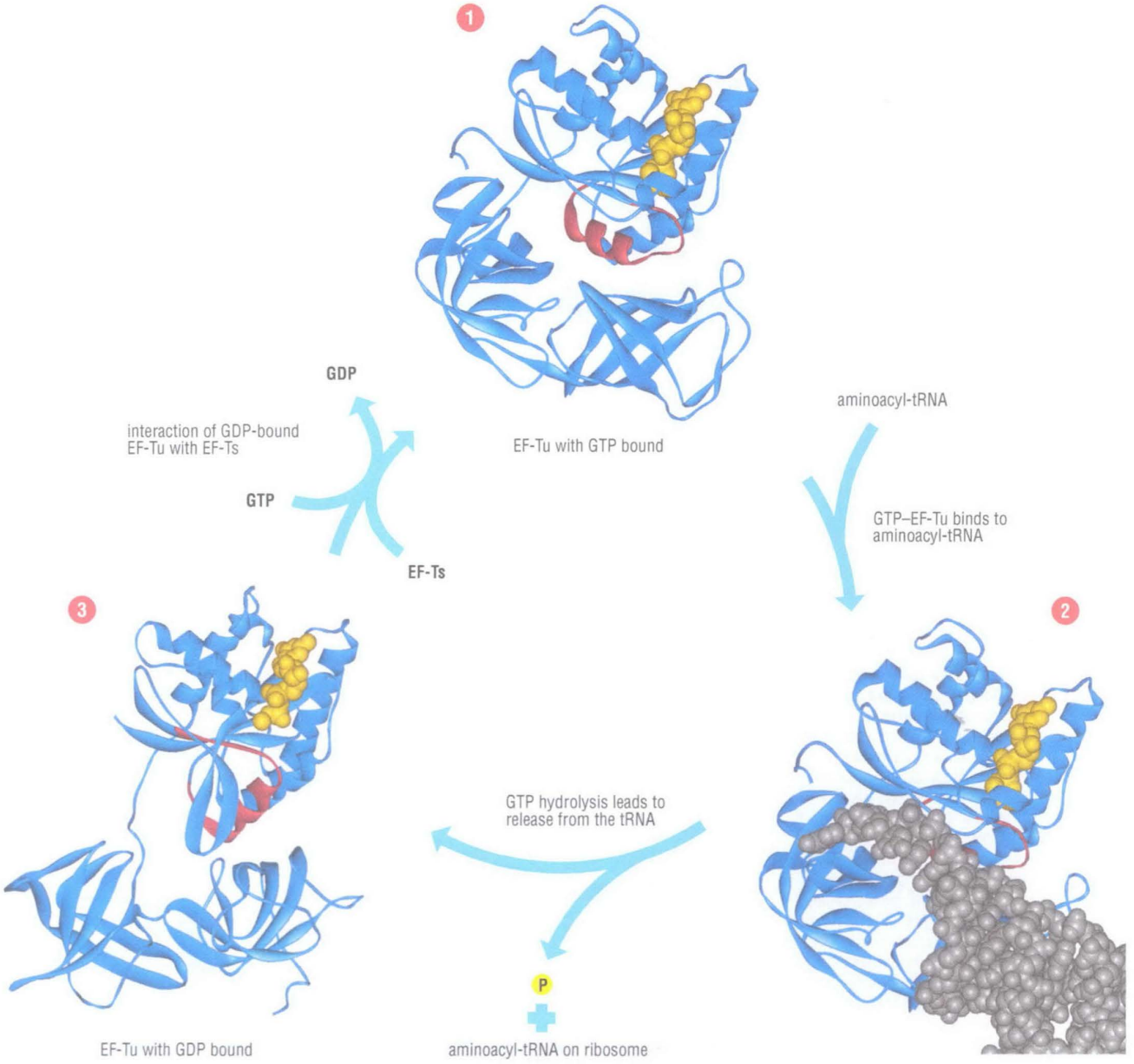


Figure 3-16 The switching cycle of the elongation factor EF-Tu delivers aminoacyl-tRNAs to the ribosome EF-Tu in its GTP-bound form (structure 1 at top; GTP shown in yellow bound to the GTPase domain) binds to an aminoacyl-tRNA (structure 2; the tRNA is in grey). This complex binds to the mRNA codon displayed in the A site on the small subunit of the ribosome through pairing of the codon with the anticodon on the tRNA (not shown). Correct pairing triggers GTP hydrolysis by EF-Tu, causing a conformational change that dissociates the factor from the tRNA molecule, which can now deliver its amino acid to the growing polypeptide chain. Generally, only those tRNAs with the correct anticodon can remain paired to the mRNA long enough to be added to the chain. The GDP form of EF-Tu (structure 3; GDP shown in yellow) released from the ribosome then interacts with EF-Ts, its guanine-nucleotide exchange factor (not shown), to facilitate exchange of GDP for GTP. The GTP-bound form of EF-Tu (1) can then restart the cycle. Shown in red is a helical segment in the GTPase domain that undergoes a significant conformational change between the GTP-bound (1) and GDP-bound (3) forms of the enzyme. The two forms also differ markedly in the relative arrangements of their domains.

©2004 New Science Press Ltd

Myosin and kinesin are ATP-dependent nucleotide switches that move along actin filaments and microtubules respectively

Many processes in living organisms and cells require movement. Directed transport of macromolecules, vesicles, organelles and chromosomes within the cytoplasm depends on motors that drive such transport. Bacterial swimming and the movement generated by muscles in higher organisms are examples of processes that also require motor power. Most biological motors move unidirectionally along protein polymers such as actin or microtubules; these polymers can be considered the rails along which protein motor "engines" move cargo; the same sort of machinery is used to slide the actin and myosin filaments past one another in muscle contraction. Defective molecular transport can result in developmental defects as well as cardiovascular and neuronal diseases.

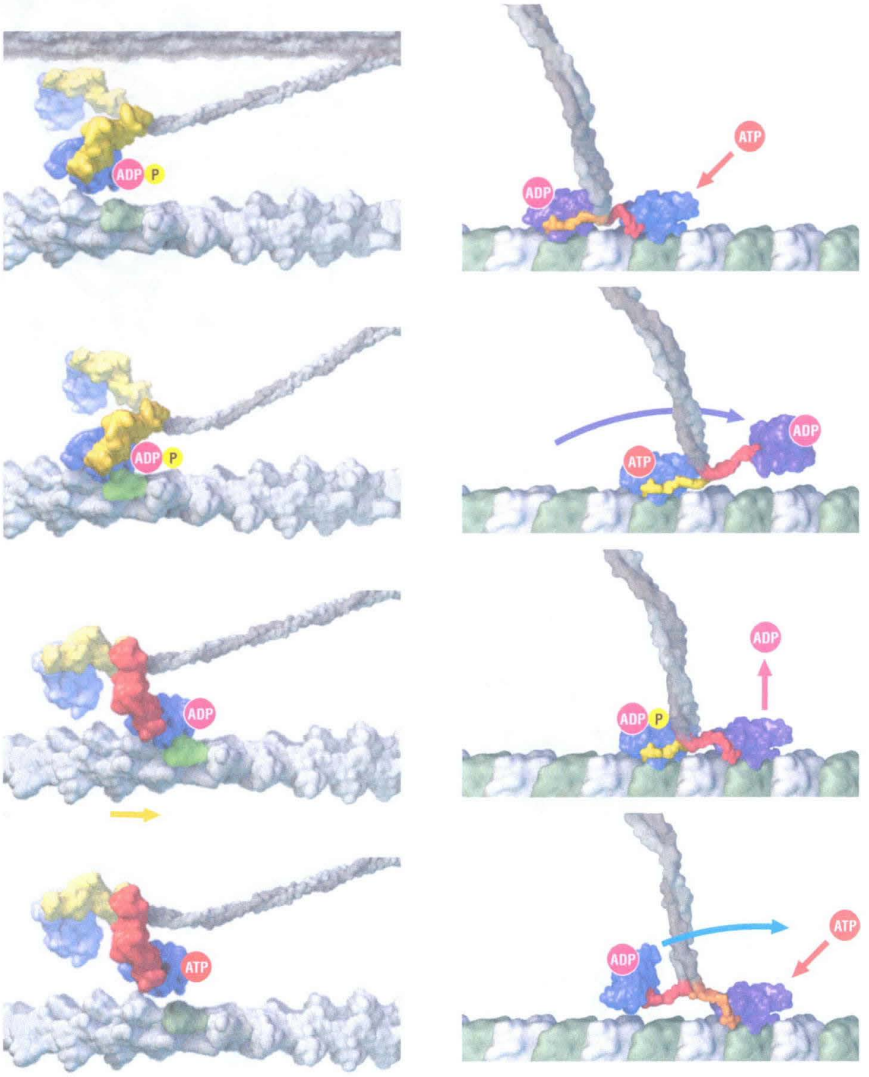


Figure 3-17 Models for the motor actions of muscle myosin and kinesin Left panel, the sequential conformational changes in muscle myosin result in a rowing "stroke" that moves an actin filament. Muscle myosin is a dimer of two identical motor heads which are anchored in the thick myosin filament (partially visible at the upper edge of the frame) by a coiled-coil tail. At the start of the cycle (top frame of the sequence) ADP and inorganic phosphate (P_i) are bound to the heads as a result of hydrolysis of a bound ATP by the intrinsic ATPase activity of the catalytic cores (blue). The lever-arms that eventually cause the movement are shown in yellow in their "prestroke" conformation. In this conformation the catalytic core binds weakly to actin (bottom of frame). The second frame shows one myosin head docking onto a specific binding site (green) on the actin thin filament (pale grey). (The two myosin heads act independently, and only one attaches to actin at a time.) As shown in the third frame, on actin docking P_i is released from the active site, and there is a conformational change in the head that causes the lever arm to swing to its "poststroke" ADP-bound position (red) while the head remains bound to the actin. This moves the actin filament by approximately 100 Å in the direction shown by the arrow. After completing the stroke, ADP dissociates and ATP binds to the active site and undergoes hydrolysis, reverting the catalytic core domain to its weak-binding actin state, and bringing the lever arm back to its prestroke state. Right panel, kinesin "walks" along a microtubule, its two head domains moving in front of each other in turn. The coiled-coil tail of the kinesin (grey) leads to the attached cargo (not shown). The catalytic heads are shown in blue and purple and the microtubule in green and white. Movement is generated by conformational changes in the linker regions that join the heads to the tail and are colored here according to their conformational state: red, not bound to microtubule; orange, partially bound to microtubule; yellow, tightly bound to microtubule. The direction in which the kinesin will move is shown by the arrows. At the beginning of the movement cycle, the "trailing" head has ADP bound and the "leading" head is empty and neither linker is docked tightly to the microtubule. As shown in the first two frames, when ATP binds to the leading head, its linker (red in the top frame and yellow in the second frame) adopts a conformation that as well as docking it firmly to the microtubule reverses its position and thus throws the trailing head forward by about 160 Å (arrow) towards the next binding site on the microtubule (second frame). This head docks onto the binding site (third frame), which moves the attached cargo forward 80 Å. Binding also accelerates the release of ADP from this head, and during this time the ATP on the other head is hydrolyzed to ADP-P_i (third frame). After ADP dissociates from the new leading head ATP binds in its turn, causing the linker to zipper onto the core (partially docked linker in orange). The new trailing head, which has released its phosphate and detached its neck linker (red) from the core, is being thrown forward. Original illustration kindly provided by Graham Johnson. From Vale, R.D. and Milligan, R.A.: *Science* 2000, **288**:88–95, with permission.

References

Capaldi, R.A. and Aggeler, R.: **Mechanism of the F(1)F(0)-type ATP synthase, a biological rotary motor.** *Trends Biochem. Sci.* 2002, **27**:154–160.

Goldstein, L.S.: **Molecular motors: from one motor many tails to one motor many tales.** *Trends Cell Biol.* 2001, **11**:477–482.

Holmes, K.C. and Geeves, M.A.: **The structural basis of muscle contraction.** *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* 2000, **355**:419–431.

Sablin, E.P. and Fletterick, R.J.: **Nucleotide switches in molecular motors: structural analysis of kinesins and myosins.** *Curr. Opin. Struct. Biol.* 2001, **11**:716–724.

Sindelar, C.V. *et al.*: **Two conformations in the human kinesin power stroke defined by X-ray crystallography and EPR spectroscopy.** *Nat. Struct. Biol.* 2002, **9**:844–848.

Spudich, J.A.: **The myosin swinging cross-bridge model.** *Nat. Rev. Mol. Cell Biol.* 2001, **2**:387–392.

Vale, R.D. and Milligan, R.A.: **The way things move:**

©2004 New Science Press Ltd

Nearly all molecular motors share a common feature: a core ATPase domain that binds and hydrolyzes ATP, and in doing so switches between different conformations in a process that is similar to the GTPase molecular switches but which is carried out with a very different protein scaffold. Attached to the core domain are smaller transmission or converter domains that read out the nucleotide-dependent conformation of the core and respond by conformational changes. In turn, these domains relay the conformational change into larger motions by altering the positions of an amplifier region, very often a lever-arm or coiled coil. Large changes in the position of the amplifier in relation to an attached protein are what finally cause very large movements of proteins (Figure 3-17).

Although the overall folds of the ATPase domains of motors and the GTPase domains of G proteins are different, there are striking similarities in the local architecture around the bound nucleotide and in the mechanism of switching. Motor ATPases have switch I and switch II regions that correspond almost exactly to those in the G proteins, and these regions are involved in binding the gamma-phosphate of ATP, just as in the GTPases (Figure 3-18). When ATP is hydrolyzed by the intrinsic ATPase activity of the motor, the switch I and II regions relax into their ADP-bound conformational states, which starts the relay of conformational changes that are ultimately propagated into large movements of attached proteins and domains.

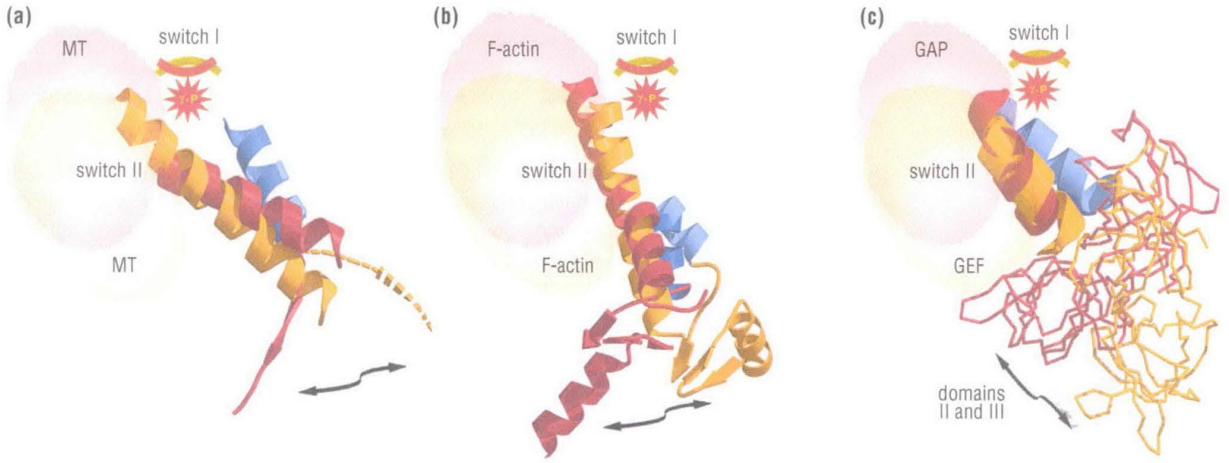


Figure 3-18 Structural and functional similarity between different families of molecular switches In all these graphics, the position of the switch II helix is shown in yellow for the nucleoside diphosphate-bound protein, and in red for the triphosphate-bound conformation. The switch I region is shown schematically with the gamma-phosphate of the nucleotide depicted as a red star. Specific macromolecular partners interact with the switch regions in these two states and are shown as surfaces of the same color. The carboxy-terminal helix of the ATPase or GTPase domain, to which the mechanical elements of molecular motors or additional domains of G proteins are attached, is shown in blue. **(a)** The switch II region of the motor protein kinesin. The nucleotide-driven conformational switching in the kinesin switch II region results in large force-generating rearrangements of the neck domain (see Figure 3-17; not shown here) and movement of kinesin along a microtubule (MT). **(b)** The same mechanism of switching in myosin converts small conformational changes in the nucleotide-binding site into larger movements of the lever-arm helix (see Figure 3-17), part of which is shown here at the bottom of the graphic (ADP state, PDB 2mys; ATP state, PDB 1br1). **(c)** An identical molecular mechanism is used by G proteins for their domain rearrangements and for controlling affinity for their specific macromolecular partners—GEFs, GAPs and effectors. As well as the switch II helix, domains II and III of EF-Tu are shown in GDP state (PDB 1tui) and the GTP state (PDB 1eft). Graphic kindly provided by E.P. Sablin. From Sablin, E.P. and Fletterick, R.J.: *Curr. Opin. Struct. Biol.* 2001, **11**:716–724, with permission.

looking under the hood of molecular motor proteins.
Science 2000, **288**:88–95.

A movie of kinesin moving may be found at:
http://valelab.ucsf.edu/images/mov-procmotconvkinrev5.mov

©2004 New Science Press Ltd

3-11 Regulation by Degradation

Protein function can be controlled by protein lifetime

Proteins not only carry signals that determine their location, they also carry signals that determine their lifetime. Different proteins can have widely differing half-lives within the cell—anywhere from a few minutes to many days—and this time depends not only on the stability of the protein's structure but also on specific cellular machinery for degradation, which in some cases recognizes specific sequence and structural features. Stability in the extracellular environment is also variable, and depends on the intrinsic stability of the protein fold and the presence or absence of both relatively nonspecific and specific proteases. The shortest-lived proteins are usually those that are important in controlling cellular processes, such as enzymes that catalyze rate-determining steps in metabolic pathways or proteins such as cyclins that regulate cell growth and division. Rapid degradation of such proteins makes it possible for their concentrations to be changed quickly in response to environmental stimuli.

Protein degradation in cells is accomplished by machinery that tags both misfolded and folded proteins for specific proteolysis and destroys them. This mechanism depends in part on intrinsic protein stability at physiological temperature. One measure of the intrinsic stability of a protein is its resistance to thermal unfolding. Although there are microorganisms that survive at very high temperatures (so-called thermophilic organisms)—and in which, presumably, all proteins are fairly thermostable—very stable proteins are also found in the cytosol of mesophilic organisms (that is, organisms that live at normal temperatures). There appears to be no correlation between the type of protein fold and the intrinsic stability of a protein; rather, it is the specific sequence that determines stability, through the presence or absence of stabilizing interactions between side chains. Comparison of the structures of thermostable proteins with their mesophilic counterparts shows that many different stabilizing factors can be employed, mostly involving electrostatic interactions of various sorts. As the net free energy of stabilization of most folded proteins is relatively small, about 21–42 kJ/mole, only a few additional interactions can make a large difference in thermal stability.

Temperature-sensitive mutants, which are widely used in genetic research, are mutant organisms that produce a mutated protein that has reduced stability at physiological temperatures. These mutant proteins are more susceptible than the normal forms to degradation by non-specific proteases, and when such mutations occur in humans they can lead to disorders caused by insufficient amounts of the active protein.

Proteins are targeted to proteasomes for degradation

The proteolytic machinery responsible for targeted protein degradation in cells is a giant multiprotein assembly called the **proteasome** (Figure 3-19). In eukaryotes, proteins carrying an appropriate signal for destruction are recognized by an enzyme that then further tags the protein for degradation in the proteasome. The tag consists of many molecules of a small

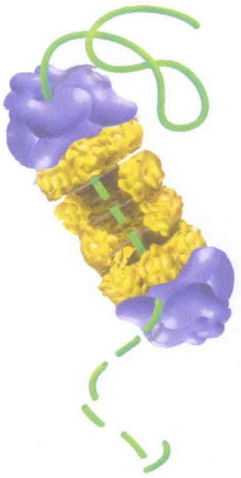


Figure 3-19 The eukaryotic proteasome Proteins targeted for destruction (green) are fed into the multiprotein complex called the proteasome. In prokaryotes, these machines of destruction consist simply of a tunnel-like enzymatic core; in eukaryotes they have an additional cap (here shown in purple) at either or both ends. The core is formed by four stacked rings surrounding a central channel that acts as a degradation chamber. The caps recognize and bind to proteins targeted by the cell for destruction. On entry into the proteasome, proteins are unfolded in a process that uses the energy released by ATP hydrolysis and injected into the central core, where they are enzymatically degraded into small fragments. Graphic kindly provided by US Department of Energy Genomes to Life Program, http://doegenomestolife.org.

Definitions

proteasome: a multiprotein complex that degrades ubiquitinated proteins into short peptides.

stress-response proteins: proteins whose synthesis is induced when cells are subjected to environmental stress, such as heat.

temperature-sensitive mutants: organisms containing a genetic mutation that makes the resulting protein sensitive to slightly elevated temperatures. The temperature at which the mutant protein unfolds is called the restrictive temperature. The term is also used for the protein itself.

ubiquitin: a small protein that when attached to other proteins (**ubiquitination**), targets them for degradation to the **proteasome**. Sometimes ubiquitin tagging targets a protein to other fates such as endocytosis.

ubiquitination: the attachment of ubiquitin to a protein.

©2004 New Science Press Ltd

protein, **ubiquitin**, which are attached to the protein to be degraded and which are recognized by the cap. **Ubiquitination** is carried out by a multi-enzyme pathway that starts by recognition of an exposed lysine near the amino terminus of the target protein. Once one ubiquitin molecule is covalently attached to this lysine, by enzymes called ubiquitin ligases, additional ubiquitins are attached to the first one, producing a polyubiquitin chain. Polyubiquitinated proteins are the primary substrates for the proteasome, which in an ATP-dependent process degrades these proteins to short peptides (Figure 3-20). In prokaryotes, a homolog of the proteasome acts without such a targeting signal.

Some specific ubiquitin-conjugating systems exist for certain proteins whose regulated destruction is crucial for cell growth, such as cyclins and other cell-cycle control proteins. Defects in ubiquitin-dependent proteolysis have been shown to result in a variety of human diseases, including cancer, neurodegenerative diseases, and metabolic disorders. Proteasome inhibitors are showing promise as treatments for certain cancers.

Proteins are targeted to the ubiquitin/proteasome system by a number of signals. The first to be discovered is remarkably simple, consisting only of the first amino acid in the polypeptide chain. Methionine, serine, threonine, alanine, valine, cysteine, glycine and proline are protective in bacteria; the remaining 12 signal proteolytic attack. Proteins are, however, more commonly targeted to the degradative pathway by a number of other more complex internal signals (for example, phosphorylation of specific residues, denaturation or oxidative damage). For example, the transcription factor NFκB is inhibited and retained in the cytoplasm by another protein IκB, which has a short peptide motif containing two serines. Phosphorylation of these serines, which is the culmination of a complex signal transduction pathway, leads to recognition by a ubiquitin ligase, the ubiquitination of IκB and its degradation, releasing NFκB. Another example of a regulated signal for ubiquitination is seen in the transcription factor hypoxia-inducible factor, HIF-1α, which is active when oxygen levels fall below about 5% O_2. In normal oxygen conditions HIF-1α is rapidly degraded because an oxygen-sensing prolyl hydroxylase modifies it and this modification is recognized by a specific ubiquitin ligase.

Unfolded and damaged proteins are preferential substrates for ubiquitination and degradation, even where the amino-terminal amino acid is protective. Some mechanisms for regulating protein degradation include modification of the protein's amino terminus. All proteins are initially synthesized with methionine as their amino-terminal amino acid. Methionine aminopeptidases and associated enzymes remove the methionine and trim the end of the polypeptide chain until certain residues are reached. However, methionine aminopeptidases will not remove this initiator methionine if it would expose an amino acid that signals destruction.

A class of proteins exists in all cells in all organisms that bind to selected proteins to help solubilize and refold aggregates of misfolded and unfolded forms before they can be degraded. These are known as **stress-response proteins** and their synthesis is upregulated in response to heat shock and other stresses that increase the amount of protein unfolding, misfolding and aggregation in the cell.

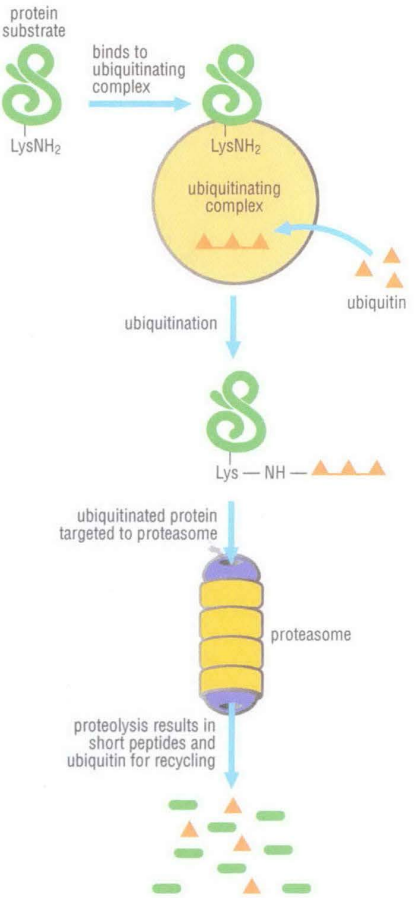


Figure 3-20 Pathway for degradation of ubiquitinated proteins A substrate protein with an exposed lysine side chain near the amino terminus is targeted by binding of a multienzyme ubiquitinating complex which, in this example, recognizes the amino-terminal amino acid of the substrate. The complex attaches polyubiquitin chains to the substrate in an ATP-dependent reaction. The polyubiquitinated substrate is then targeted to the proteasome, whose cap recognizes the ubiquitin tag. After the substrate is chopped up into peptide fragments (which may then be degraded further by other proteases), the ubiquitin is recycled.

References

Benaroudj, N. *et al.*: **The unfolding of substrates and ubiquitin-independent protein degradation by proteasomes.** *Biochimie* 2001, **83**:311–318.

Goldberg, A.L. *et al.*: **The cellular chamber of doom.** *Sci. Am.* 2001, **284**:68–73.

Kisselev, A.F. and Goldberg, A.L.: **Proteasome inhibitors: from research tools to drug candidates.** *Chem. Biol.* 2001, **8**:739–758.

Laney, J.D. and Hochstrasser, M.: **Substrate targeting in the ubiquitin system.** *Cell* 1999, **97**:427–430.

Petsko GA.: **Structural basis of thermostability in hyperthermophilic proteins, or "there's more than one way to skin a cat".** *Methods Enzymol.* 2001, **334**:469–478.

Sherman, M.Y. and Goldberg, A.L.: **Cellular defenses against unfolded proteins: a cell biologist thinks about neurodegenerative diseases.** *Neuron* 2001, **29**:15–32.

Stock D, *et al.*: **Proteasome: from structure to function.** *Curr. Opin. Biotechnol.* 1996, **7**:376–385.

Web site:
http://www.chembio.uoguelph.ca/educmat/chm736/degradat.htm

©2004 New Science Press Ltd

3-12 Control of Protein Function by Phosphorylation

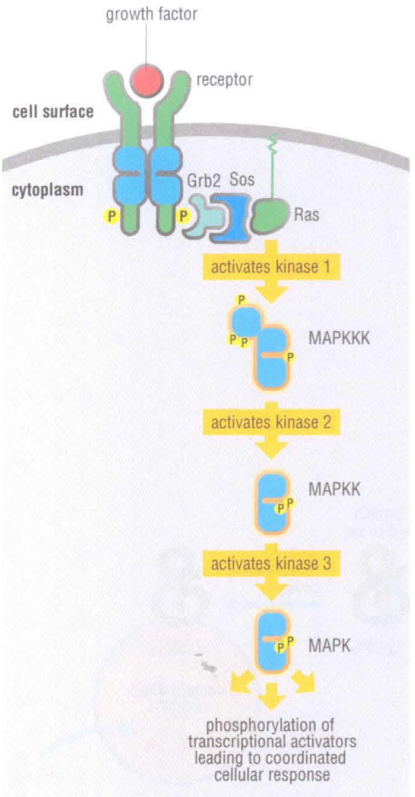


Figure 3-21 A kinase activation cascade in an intracellular signaling pathway that regulates cell growth A growth factor, or mitogen, initiates a kinase cascade by inducing dimerization of its receptor, whose cytoplasmic regions contain tyrosine kinase domains. When these domains are brought together they phosphorylate and activate one another, leading to phosphorylation of tyrosines on other parts of the receptor's cytoplasmic domains, and creating binding sites for an adaptor molecule (such as Grb2), which in turn binds a guanine-nucleotide exchange factor (Sos), bringing this into proximity with the membrane-associated G protein Ras, which is thereby activated (see Figure 3-14). Ras then activates a kinase cascade of three so-called mitogen-activated protein kinases (MAPKs): a mitogen-activated protein kinase kinase kinase (MAPKKK), which phosphorylates and activates a mitogen-activated protein kinase kinase (MAPKK), which in turn phosphorylates and activates a mitogen-activated protein kinase (MAPK). This last kinase phosphorylates a number of different gene regulatory proteins.

Protein function can be controlled by covalent modification

It is estimated that 50–90% of the proteins in the human body are post-translationally modified. Post-translational covalent modification allows the cell to expand its protein structural and functional repertoire beyond the constraints imposed by the 20 naturally encoded amino acids. More than 40 different post-translational covalent modifications have been identified in eukaryotic cells. While some are widespread, others have been observed on only a few proteins. Phosphorylation, glycosylation, lipidation, and limited proteolysis are the most common. Of the remainder, some of the most important are methylation, *N*-acetylation, attachment of the protein SUMO and nitrosylation. In the remaining sections of this chapter, we shall describe the part played by these modifications in controlling protein function.

Most covalent modifications can change the location of the protein, or its activity, or its interactions with other proteins and macromolecules. Limited proteolysis can also be deployed to amplify low-concentration or transient signals, through proteolytic cascades in which the initial stimulus activates a proteolytic enzyme which in turn activates many molecules of the next enzyme in the cascade, and so on. The complement cascade, which is activated in response to microbial cell surfaces and promotes immune defenses, is one example; the blood clotting cascade, discussed later, is another.

The commonest form of covalent modification on proteins is reversible phosphorylation of serine, threonine or tyrosine side chains and we shall discuss this first. Phosphorylation, like limited proteolysis, can produce a very large response to a relatively small signal input. A single molecule of an enzyme that is catalytically activated by phosphorylation can process many thousands of substrate molecules. If the substrate is itself another enzyme, the amplification effects can be magnified further. When phosphorylation of one enzyme causes it to become active so that it can, in turn, covalently modify another enzyme and so on, a regulatory cascade can be set up which produces a large final response very rapidly.

Unlike limited proteolysis, however, phosphorylation is reversible, and this makes it well suited as a regulatory mechanism for intracellular signaling pathways, where it is important that responses can be turned off rapidly as well as turned on (Figure 3-21). In signaling pathways, an important feature of the regulatory cascade may be the opportunity it offers for independent regulation of different downstream targets and thus for a flexible and coordinated response by the cell to the initial signal.

Phosphorylation is the most important covalent switch mechanism for the control of protein function

Post-translational modification by phosphorylation has been found in all living organisms from bacteria to humans. Target proteins are phosphorylated by the action of protein kinases and dephosphorylated by **protein phosphatases**, and this reversible modification provides a switch that controls many diverse cellular processes including metabolic pathways, signaling cascades, intracellular membrane traffic, gene transcription, and movement. The phosphoryl group derives from the terminal phosphate of nucleoside triphosphates, usually ATP. Use of separate enzymes for phosphorylation (kinases) and dephosphorylation (phosphatases) enables independent control of these events by different stimuli. Kinases that phosphorylate proteins on serine, threonine or tyrosine residues constitute the third most common domain encoded in the human genome sequence, with 575 such kinases (about 2% of the genome) identified to date. In prokaryotes, a different type of kinase domain phosphorylates proteins on histidine and aspartate

Definitions

protein phosphatase: enzyme that specifically removes phosphate groups from phosphorylated serines, threonines or tyrosines on proteins.

References

Barford, D. *et al.*: **Structural mechanism for glycogen phosphorylase control by phosphorylation and AMP.** *J. Mol. Biol.* 1991, **218**:233–260.

Cozzone, A.J.: **Regulation of acetate metabolism by protein phosphorylation in enteric bacteria.** *Annu. Rev. Microbiol.* 1998, **52**:127–164.

Dean, A.M. and Koshland, D.E. Jr.: **Electrostatic and steric contributions to regulation at the active site of isocitrate dehydrogenase.** *Science* 1990, **249**:1044–1046.

Hunter, T.: **Signaling—2000 and beyond.** *Cell* 2000, **100**:113–127.

Hurley, J.H. *et al.*: **Regulation of an enzyme by phosphorylation at the active site.** *Science* 1990, **249**:1012–1016.

Johnson, L.N. and O'Reilly, M.: **Control by phosphorylation.** *Curr. Opin. Struct. Biol.* 1996, **6**:762–769.

Lee, M. and Goodbourn, S.: **Signalling from the cell surface to the nucleus.** *Essays Biochem.* 2001, **37**:71–85.

Manning, G. *et al.*: **The protein kinase complement of the human genome.** *Science* 2002, **298**:1912–1934.

Wilkinson, M.G. and Millar, J.B.: **Control of the eukaryotic cell cycle by MAP kinase signaling pathways.** *FASEB J.* 2000, **14**:2147–2157.

Zolnierowicz, S. and Bollen, M.: **Protein phosphorylation and protein phosphatases.** *EMBO J.* 2000, **19**:483–488.

©2004 New Science Press Ltd

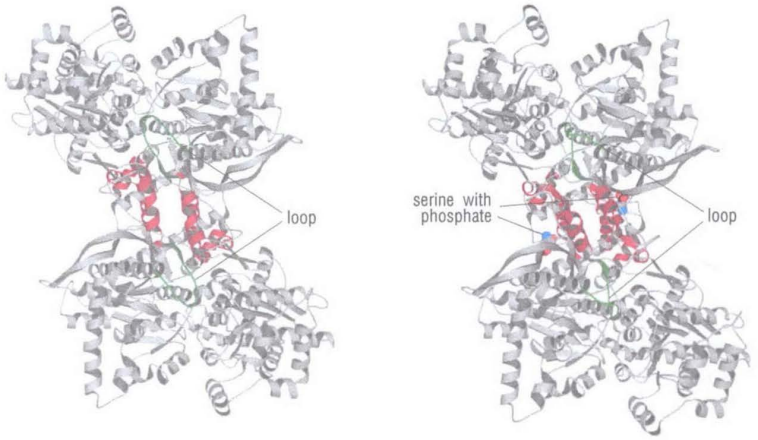


Figure 3-22 Conformational change induced by phosphorylation in glycogen phosphorylase The enzyme glycogen phosphorylase (GP) cleaves glucose units from the nonreducing ends of glycogen, the storage polymer of glucose, by phosphorolysis to produce glucose 6-phosphate. Muscle GP is a dimer of identical subunits. In response to hormonal signals the enzyme is covalently converted from the less active phosphorylase b form (left) to the more active phosphorylase a form (right) through phosphorylation of a single serine residue, serine 14, catalyzed by the enzyme phosphorylase kinase. On phosphorylation, a loop (shown in green) that sterically restricts access to the active site in the b form moves out, making the substrate-binding site more accessible. The large changes in conformation that occur at the interface between the two monomers are shown in red. (PDB 1gpa and 1gpb)

residues (although some eukaryotic-like serine/threonine and tyrosine kinases have been found in some bacteria), and the percentage of the genome devoted to regulation by phosphorylation is strikingly similar: about 1.5% of the *E. coli* genome, for example, encodes such proteins.

Covalent addition of a phosphoryl group to the side chain of serine, threonine, tyrosine, histidine or aspartic acid can have profound effects on the function of the protein. Phosphorylated residues have acquired a group that carries a double negative charge and is capable of multiple hydrogen-bonding interactions. Structural studies of phosphorylated proteins have shown that two types of interaction predominate: hydrogen bonding to main-chain amide groups at the positively polarized amino-terminal end of an alpha helix, and salt-bridging to one or more arginine residues. Other residues may also be involved in the recognition and binding of phosphoryl groups in specific cases.

Phosphorylation can affect the target protein in two ways, which are not mutually exclusive. One effect is to change the activity of the target protein, either considerably or subtly. This change in activity may come about solely from the added bulk and charge properties of the phosphoryl group, or may result from a large conformational change in the protein, or both. The second effect of phosphorylation is to provide a new recognition site for another protein to bind (see, for example, the binding of Grb2 to the phosphorylated receptor tail in Figure 3-21). Such protein–protein interactions usually involve specialized interaction domains on the second protein that recognize the phosphorylated peptide segment; the most common domain recognizing phosphotyrosines, for example, is the SH2 domain (see section 3-1).

An instance of phosphorylation activating an enzyme by inducing a large conformational change is seen in muscle glycogen phosphorylase. Covalent attachment of a phosphoryl group to serine 14 results in a rearrangement of the amino-terminal residues in the enzyme such that the serine side chain shifts 50 Å, leading to a change in the subunit–subunit contacts of this dimeric protein, with a concomitant rearrangement of the active-site residues that activates the enzyme (Figure 3-22). The phosphoserine is stabilized in its new position primarily by salt bridges to two arginine side chains.

In contrast, the inactivation of the TCA-cycle enzyme isocitrate dehydrogenase by serine phosphorylation involves no conformational changes. Serine 113, the residue phosphorylated, is located in the active site (Figure 3-23), and attachment of a phosphoryl group inhibits binding of the negatively charged substrate isocitrate by steric exclusion and electrostatic repulsion. In isocitrate dehydrogenase the phosphoserine is stabilized by helix dipole and main-chain hydrogen-bond interactions.

One of the principal functions of reversible phosphorylation carried out by the receptor tyrosine kinases that initiate many eukaryotic signal transduction pathways is the temporary provision of new protein-interaction sites (see Figure 3-21, where phosphorylation of the receptor's own cytoplasmic domains provides a recognition site for an adaptor molecule that relays the signal onward). Another is to phosphorylate and directly activate other downstream signaling and effector molecules. The regulation of these kinases is therefore of central importance in biology, and in the next section we describe the conserved features of their activation mechanism and discuss its regulation in two important kinase families: the Src family and the cyclin-dependent kinases.

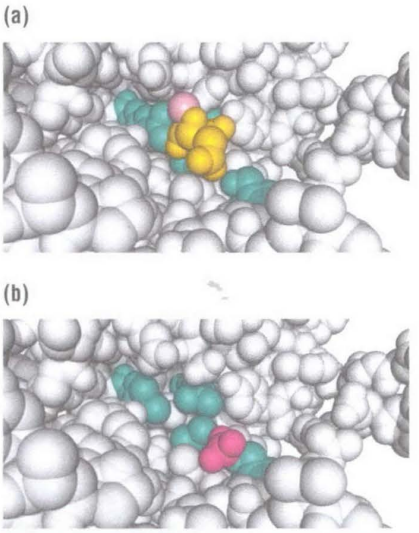


Figure 3-23 Inactivation of the active site of *E. coli* isocitrate dehydrogenase by phosphorylation The substrate-bound **(a)** and phosphorylated **(b)** states of the active site of isocitrate dehydrogenase are shown as space-filling models. The active-site residues include the serine 113 side chain, which is shown in green, the negatively charged portion of the isocitrate substrate in yellow and the phosphoryl group on serine 113 in red (in b). The phosphoryl group occupies almost the same location in the active site as the negatively charged portion of isocitrate. The phosphorylated form is inactive because isocitrate binding is both sterically blocked and electrostatically repelled by the negatively charged phosphoryl group. (PDB 5icd)

©2004 New Science Press Ltd

3-13 Regulation of Signaling Protein Kinases: Activation Mechanism

Protein kinases are themselves controlled by phosphorylation

Phosphorylation of proteins on serine/threonine or tyrosine residues is probably the single most important regulatory mechanism in eukaryotic signal transduction, and tyrosine phosphorylation lies at the heart of control of the eukaryotic cell cycle. To a first approximation, the protein kinases responsible for phosphorylating proteins on serine, threonine and tyrosine residues all have the same fold for the catalytic domain (Figure 3-24a), although many of them also have other subunits or other domains that serve regulatory functions or target their kinase activity to specific protein substrates (Figure 3-24b). Most, but not all, of these kinases are normally inactive, and before they can phosphorylate other proteins they must themselves be activated by phosphorylation of a threonine or tyrosine residue that is located in a region termed the **activation segment**, also known as the **activation loop**. The conserved mechanism of activation of these kinases is illustrated in Figure 3-25. Kinases controlled by activation-loop phosphorylation represent some of the most important enzymes in signal transduction cascades. Two of the best understood are the Src family of tyrosine protein kinases, which are activated early in eukaryotic cell signaling pathways, and the cyclin-dependent kinases (Cdks) which coordinate the eukaryotic cell cycle. We discuss the part played by regulatory and targeting domains in the activation of Src kinases below, and in the next section we describe the activation and targeting of the cyclin-dependent kinase Cdk2.

Src kinases both activate and inhibit themselves

Src-family kinases are activated early in many signaling pathways and once activated sustain their own activated states by **autophosphorylation**, providing for a large amplification of the signal. The original Src kinase was discovered as an oncogenic variant in a tumor virus that causes sarcomas in chickens (Src is short for sarcoma). Its tumorigenic action is due to a mutation that causes unregulated autophosphorylation, leading to a sustained growth signal, and we now understand the structural basis both for the normal regulation of the kinase and for its oncogenic activation.

Src kinases recognize their target proteins through SH2 and SH3 domains that are joined to one another and to the catalytic domain by flexible linker regions (Figure 3-24b). In the absence of activating signals, these domains bind to the kinase domain, holding it in an inactive conformation (Figure 3-26): the SH2 domain binds to an inhibitory phosphate on a tyrosine

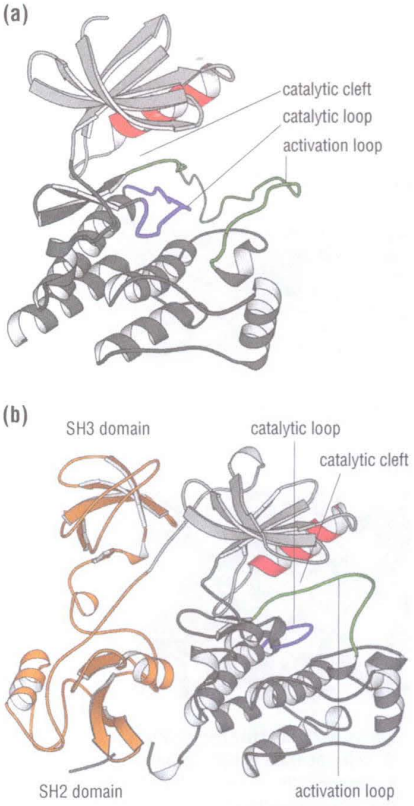


Figure 3-24 The conserved protein kinase catalytic domain (a) The catalytic domain of Lck, a signaling tyrosine kinase expressed in the cells of the immune system, is typical of the highly conserved catalytic domain of all protein kinases. This is structurally divided into two lobes, the amino-terminal (upper) lobe (light grey) being formed almost entirely from beta strands but for a single alpha helix (the C helix, red), while the carboxy-terminal (lower) lobe (dark grey) is formed almost entirely from alpha helices. Between the two lobes, which are joined by a flexible hinge, is the catalytic cleft containing the catalytic loop (purple). The activation loop, which is repositioned by phosphorylation and thereby regulates the catalytic action of the kinase, is shown in green. The flexible hinge between the two lobes is important in allowing major conformational changes that accompany activation and inhibition in some kinases. Movement of the single helical element in the amino-terminal lobe is critical to the establishment of the active conformation of the catalytic site. Kindly provided by Ming Lei and Stephen Harrison. **(b)** Structure of the Src-family tyrosine protein kinase Hck in its inactive form. The catalytic domain is attached to two small domains (gold), an SH3 domain and an SH2 domain, that target this kinase to specific protein substrates. (PDB 1ad5)

Definitions

activation loop: a stretch of polypeptide chain that changes conformation when a kinase is activated by phosphorylation and/or protein binding. This segment may or may not be the one containing the residue that is phosphorylated to activate the kinase. Usually, in the inactive state, the activation loop blocks access to the active site.

activation segment: see **activation loop**.

autophosphorylation: phosphorylation of a protein kinase by itself. Autophosphorylation may occur when the active site of the protein molecule to be phosphorylated catalyzes this reaction (*cis* autophosphorylation) or when another molecule of the same kinase provides the active site that carries out the chemistry (*trans* autophosphorylation). Autophosphorylation *in trans* often occurs when kinase molecules dimerize, a process that can be driven by ligand binding as in the receptor tyrosine kinases.

©2004 New Science Press Ltd

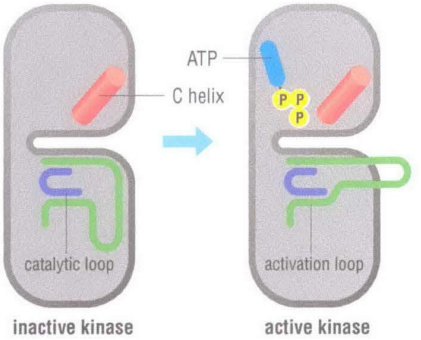


Figure 3-25 Conserved mechanism of kinase activation In all protein kinases, the activation loop (green) plays a central part in regulating catalytic activity. A conserved aspartate in the activation loop is critical to the catalytic action of the kinase, and changes in the position of the loop on phosphorylation lead to repositioning of this and other critical catalytic residues, and in some cases also regulate access of the substrate to the active-site cleft. While the active configuration of the active site (right) is essentially the same for all kinases, the inactive configuration may vary considerably. In the example schematically represented here, the activation loop swings down in the inactive conformation (left), but in some kinases it instead swings up and occludes the catalytic cleft so that substrate cannot gain access. The purple loop is the conserved catalytic loop, found in all kinases, which contains residues that participate in the chemical step of protein phosphorylation.

residue close to the carboxyl terminus of the protein, and when it is bound, the linker region joining the SH2 domain to the catalytic domain forms a polyproline helix to which the SH3 domain can bind. This clamps the catalytic domain in an inactive state from which it is released either by dephosphorylation of the tyrosine on the carboxyl tail of the protein, or by binding of the SH2 domain by a phosphotyrosine on the cytoplasmic tail of an activated receptor tyrosine kinase: this causes conformational changes that activate phosphorylation by the Src kinase by rearranging the activation loop in the active site, and at the same time releases the SH3 domain to bind to the target (Figure 3-26). The oncogenic properties of the viral Src kinase are now known to be due to a mutation of the carboxyl tail of the molecule that eliminates the tyrosine residue that is the target of the inhibitory phosphorylation.

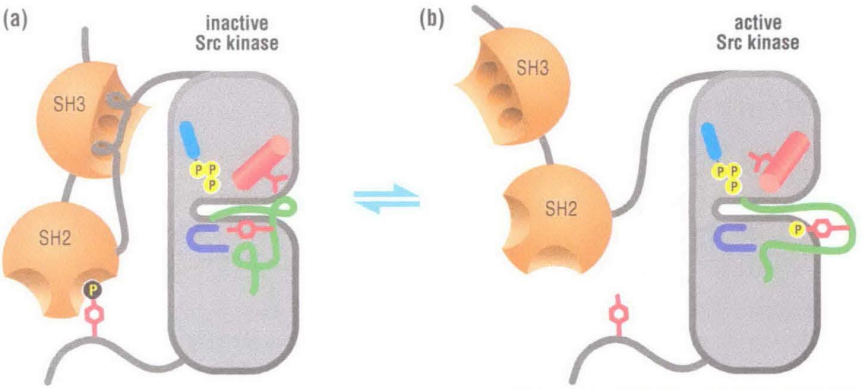


Figure 3-26 Regulation of a Src-family protein kinase In the absence of activating signals the SH2 and SH3 domains hold the kinase in an inactive conformation **(a)** in which the SH2 domain binds to an inhibitory phosphate on a tyrosine residue close to the carboxyl terminus of the protein, while the SH3 domain binds to a polyproline helix in the linker region joining the SH2 domain to the catalytic domain. **(b)** When the SH2 domain releases the carboxyl tail of the protein, either because the inhibitory phosphate is removed, or on binding a target phosphotyrosine, the polyproline helix rearranges, releasing the SH3 domain, and initiating a series of structural changes that propagate to the C helix (red) in the upper catalytic domain and the activation loop (green), which assume new conformations appropriate for substrate binding and autophosphorylation of a tyrosine in the activation loop.

References

Huse, M. and Kuriyan, J.: **The conformational plasticity of protein kinases.** *Cell* 2002, **109**:275–282.

Theodosiou, A. and Ashworth, A.: **MAP kinase phosphatases.** *Genome Biol.* 2002, **3**:reviews3009.

Xu, W. *et al.*: **Crystal structures of c-Src reveal features of its autoinhibitory mechanism.** *Mol. Cell* 1999, **3**:629–638.

Young, M.A. *et al.*: **Dynamic coupling between the SH2 and SH3 domains of c-Src and Hck underlies their inactivation by C-terminal tyrosine phosphorylation.** *Cell* 2001, **105**:115–126.

©2004 New Science Press Ltd

3-14 Regulation of Signaling Protein Kinases: Cdk Activation

Cyclin acts as an effector ligand for cyclin-dependent kinases

Cyclin-dependent kinases (Cdks) are the enzymes that drive the cell cycle. Activated periodically during the cycle, they phosphorylate proteins that, for example, move the cell onward from growth phase to DNA replication phase or from DNA replication phase into mitosis. The activation of a Cdk is a two-step process that requires binding of a regulatory protein, the cyclin, and phosphorylation on a threonine in the activation loop of the Cdk by the enzyme Cdk-activating kinase (CAK). In eukaryotes, a Cdk known as Cdk2 regulates passage from the G1 state of the cell cycle, in which cells are growing and preparing to divide, to the S phase in which their chromosomes replicate. Cdk2 is activated by cyclin A. The cyclin subunit of the cyclin–Cdk 2 complex also targets Cdk2 to the downstream targets it phosphorylates.

In its inactive, unphosphorylated state, Cdk2 is autoinhibited by the activation loop (here called the T-loop), which partially blocks the ATP-binding site (Figure 3-27a). Structural studies of phosphorylated Cdk2 in the absence of bound cyclin, a form that has 0.3% of

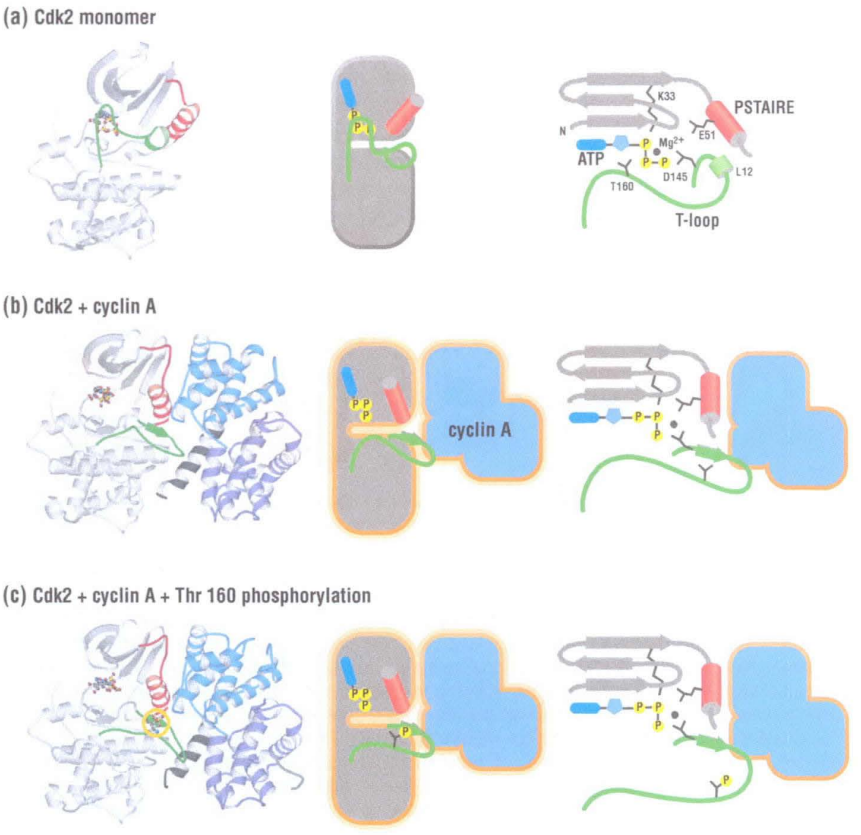


Figure 3-27 Regulation of Cdk2 activation A structure graphic representation of Cdk in various states is shown on the left, with the equivalent schematic representation in the middle and a close-up of the active site on the right. The activation loop is shown in green, with the critical helix in the upper lobe (here called PSTAIRE) in red. **(a)** In the inactive conformation of the kinase, in the absence of either cyclin or phosphorylation, a small helix (L12) in the activation loop (here called the T loop) displaces the PSTAIRE helix which contains a glutamate residue (E51) critical to catalysis. **(b)** When cyclin binds, the PSTAIRE helix moves inward and the L12 helix melts to form a beta strand, allowing E51 to interact with lysine 33 (K33) and leading to the correct orientation of the ATP phosphates. Cyclin binding also shifts the activation loop out of the active-site cleft. **(c)** Phosphorylation of threonine 160 in the activation loop flattens the activation loop, increasing its interaction with cyclin A, which allows it to interact more effectively with substrates. Taken from David O. Morgan: *The Cell Cycle: Principles of Control* (New Science Press, in the press).

©2004 New Science Press Ltd

maximal enzymatic activity, show that attachment of a phosphoryl group to threonine 160 causes the activation loop to become disordered, which presumably allows ATP to bind sometimes. No other significant conformational changes are observed. The structure of the complex of unphosphorylated Cdk2 with cyclin A (Figure 3-27b), a form that is also only 0.3% active, shows some significant conformational changes, including reorganization of the activation loop. Cyclin binding and phosphorylation therefore both induce conformational changes in the protein, but these changes are different and neither alone is sufficient to fully activate the enzyme.

The structure of the phospho-Cdk2–cyclin A complex (Figure 3-27c) shows why both modifications are needed for full activity (Figure 3-28). The conformational changes seen on cyclin A binding alone all occur, but in addition the phosphothreonine side chain makes interactions that are not observed without cyclin binding. The phosphothreonine turns into the protein to interact with three arginine residues, leading to a further reorganization of the activation loop beyond that which occurs on either phosphorylation or cyclin binding alone.

This cumulative change in conformation of the activation segment is crucial for proper recognition of the substrate peptides on which this kinase acts. Specific kinases usually phosphorylate specific sequence motifs that incorporate the serine, threonine or tyrosine at which they react. Cdk2 recognizes the sequence SPXR/K. This peptide segment binds in an extended conformation across the catalytic cleft and primarily makes contact with the activation loop. In the unphosphorylated Cdk2–cyclin complex, the conformation of the activation segment does not allow the proline residue to fit properly into the active site because it would lead to a steric clash with valine 163 on the kinase. The phospho-Cdk2–cyclin A structure has a different conformation for the activation loop that repositions valine 163 out of the way. At the same time, the adjacent residue, valine 164, rotates its backbone carbonyl group away from the peptide substrate (Figure 3-28). It is this rotation that gives the kinase absolute specificity for proline in the substrate sequence motif. Proline is the only amino acid that does not have a backbone –N–H group (see Figure 1-3 in Chapter 1). Any other amino acid would have its –N–H pointing into the place where the valine 164 carbonyl group used to be, leading to an unsatisfied hydrogen bond donor in the enzyme–substrate complex, a very unfavorable situation energetically. Proline-containing substrates do not have this problem. Thus, the conformational changes in Cdk2 on phosphorylation and cyclin binding combine not only to activate the enzyme but also to allow specific recognition of the "correct" substrate.

Cdk2 provides a striking example of how protein–protein interactions combine with covalent modification by phosphorylation in regulation. This principle of multiple regulatory mechanisms operating on the same target is widely seen, especially in eukaryotic cells, and provides exquisite control of protein function.

Dephosphorylation by protein phosphatases reverses the effects of phosphorylation and it is assumed that the protein structure simply returns to its original conformation. There are many fewer phosphatases than kinases, implying that they are less specific, but some of them are known to be regulated by phosphorylation or by binding of effector molecules. Since phosphatases are in general able to recognize many phosphorylated protein substrates, it is likely that they are targeted to specific substrates as needed by mechanisms that change protein localization within the cell, as described in other sections of this chapter.

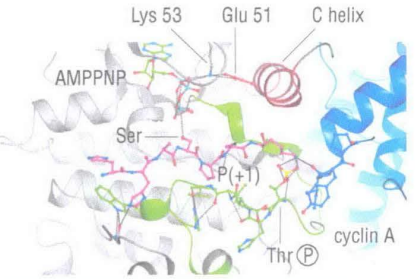


Figure 3-28 The substrate-binding site of Cdk2 The substrate to be phosphorylated (pink) binds to Cdk2 (the activation loop is shown in green) in an extended conformation. The serine to be phosphorylated (Ser) has its –OH group pointed at the gamma phosphate that will be transferred from ATP (here, the nonhydrolyzable analog AMPPNP). Activation of Cdk2 includes a change in the position of the C helix (sequence PSTAIRE, the red helix in Figure 3-27) so that glutamate 51 can interact with lysine 53 to position the ATP properly. This conformational change is induced only by the binding of cyclin A. Sequence specificity for the peptide substrate is provided by the proline residue P(+1) in the substrate. This proline comes close to the backbone of the dipeptide Val 163–Val 164 in the activation loop, which has undergone a conformational change on phosphorylation of Cdk2 to allow the substrate to fit. Graphic kindly provided by Jane Endicott and Martin Noble.

References

Brown, N.R. *et al.*: **The structural basis for specificity of substrate and recruitment peptides for cyclin-dependent kinases.** *Nat. Cell Biol.* 1999, **1**:438–443.

Brown, N.R. *et al.*: **Effects of phosphorylation of threonine 160 on cyclin-dependent kinase 2 structure and activity.** *J. Biol. Chem.* 1999, **274**:8746–8756.

Cook, A. *et al.*: **Structural studies on phospho-CDK2/cyclin A bound to nitrate, a transition state analogue: implications for the protein kinase mechanism.** *Biochemistry* 2002, **41**:7301–7311.

Song, H. *et al.*: **Phosphoprotein-protein interactions revealed by the crystal structure of kinase-associated phosphatase in complex with phosphoCDK2.** *Mol. Cell* 2001, **7**:615–626.

©2004 New Science Press Ltd

3-15 Two-Component Signaling Systems in Bacteria

Two-component signal carriers employ a small conformational change that is driven by covalent attachment of a phosphate group

G proteins and motor ATPases are generally absent from prokaryotes, which use a different class of molecular switches. Their signaling pathways, referred to as **two-component systems**, are structured around two families of proteins. The first component is an ATP-dependent histidine protein kinase (HK). This is typically a transmembrane protein composed of a periplasmic sensor domain that detects stimuli and cytoplasmic histidine kinase domains that catalyze ATP-dependent autophosphorylation. The second component of the system, a cytoplasmic response regulator protein (RR), is activated by the histidine kinase. Signals transmitted through the membrane from the sensor domain modulate the activities of the cytoplasmic kinase domains, thus regulating the level of phosphorylation of the RR. Phosphorylation of the RR results in its activation and generation of the output response of the signaling pathway. Thus, although ATP hydrolysis is used to drive this switching process, it does so indirectly: a phosphoryl group originally derived from ATP covalently modifies the RR and serves as the trigger.

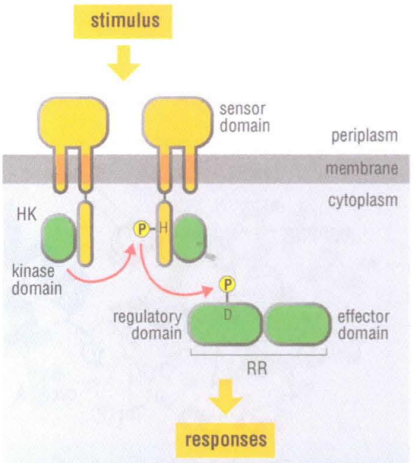


Figure 3-29 Two-component signaling mechanisms The basic two-component phosphotransfer system, found in bacteria, consists of a dimeric transmembrane receptor histidine kinase (HK) and a cytoplasmic response regulator (RR). Information flows between the two proteins in the form of a phosphoryl group (PO_3) that is transferred from the HK to the RR. HKs catalyze ATP-dependent autophosphorylation of a specific conserved His residue (H). The activities of HKs are modulated by environmental signals such as nutrients or osmotic stress. The phosphoryl group (P) is then transferred to a specific aspartic acid residue (D) located within the conserved regulatory domain of an RR. Phosphorylation of the RR typically activates an associated (or downstream) effector domain, which ultimately elicits a specific cellular response.

Two-component system proteins are abundant in most eubacterial genomes, in which they typically constitute ~1% of encoded proteins. For example, the *E. coli* genome encodes 62 two-component proteins, which are involved in regulating processes as diverse as chemotaxis, osmoregulation, metabolism and transport. In eukaryotes, two-component pathways constitute a very small number of all signaling systems. In fungi, they mediate environmental stress responses and hyphal development. In the slime mold *Dictyostelium* they are involved in osmoregulation and development, while in plants they are involved in responses to hormones and light, leading to changes in cell growth and differentiation. To date, two-component proteins have not been identified in animals and do not seem to be encoded by the human, fly or worm genomes. In most prokaryotic systems, the output response is generated directly by the RR; in many systems it functions as a transcription factor whose transcription-activating or repressing activity depends on its state of phosphorylation. Fungi and plants also contain RRs that function as transcription factors. In addition, eukaryotic two-component proteins are found at the beginning of signaling pathways where they interface with more conventional eukaryotic signaling strategies such as mitogen-activated protein (MAP) kinase and cyclic nucleotide cascades.

Similarly to most proteins in signaling pathways, two-component systems are modular in architecture. Different arrangements of conserved domains within proteins and different integration of proteins into pathways provide adaptations of the basic scheme to meet the specific regulatory needs of many different signaling systems. The prototypic prokaryotic two-component pathway (Figure 3-29) illustrates the fundamental phosphotransfer switching mechanism of both simple and more elaborate systems. Stimuli detected by the sensor domain of the histidine kinase regulate the kinase's activities. The kinase catalyzes ATP-dependent autophosphorylation of a specific histidine residue. The RR then catalyzes transfer of the phosphoryl group from this phosphorylated histidine to one of its own aspartate residues, located on the regulatory domain. Phosphorylation of the regulatory domain of the RR activates an effector domain on the same protein (or, on rare occasions, a separate effector protein) that produces the specific output response.

The regulatory domains of RRs have three activities. First, they interact with phosphorylated histidine kinase and catalyze transfer of a phosphoryl group to one of their own aspartate residues. Second, they are phosphatases that catalyze their own dephosphorylation—the counterpart to the GTPase activity of the G proteins. The phosphatase activity varies greatly among different RRs, with half-lives for the phosphorylated state ranging from seconds to hours, a span of four orders

Definitions

two-component systems: signal transduction systems found in bacteria and some eukaryotes involving a membrane-bound histidine kinase and a cytoplasmic response regulator protein that is activated by phosphorylation.

References

Parkinson, J.S. and Kofoid, E.C.: **Communication modules in bacterial signaling proteins.** *Annu. Rev. Genet.* 1992, **26**:71–112.

Stock A.M. *et al.*: **Two-component signal transduction.** *Annu. Rev. Biochem.* 2000, **69**:183–215.

Stock A.M. *et al.*: **Structure of the Mg^{2+}-bound form of CheY and mechanism of phosphoryl transfer in bacterial chemotaxis.** *Biochemistry* 1993, **32**:13375–13380.

West, A.H. and Stock, A.M.: **Histidine kinases and response regulator proteins in two-component signaling systems.** *Trends Biochem. Sci.* 2001, **26**:369–376.

Wurgler-Murphy, S.M. and Saito, H.: **Two-component signal transducers and MAPK cascades.** *Trends Biochem. Sci.* 1997, **22**:172–176.

©2004 New Science Press Ltd

of magnitude. And third, RRs regulate the activities of their associated effector domains (or effector proteins) in a phosphorylation-dependent manner. The different lifetimes of different regulators allow two-component signal systems to regulate a wide variety of cellular processes.

All RRs have the same general fold and share a set of conserved residues (Figure 3-30a). Phosphorylation of the active-site aspartate is associated with an altered conformation of the regulatory domain. A common mechanism appears to be involved in the structural changes that propagate from the active site. The phosphorylated aspartate is positioned by interaction of two of the phosphate oxygens with both a divalent cation (usually a magnesium ion) coordinated to three active-site carboxylate side chains and with the side chain of a conserved lysine residue. Two other highly conserved side chains have distinctly different orientations in the phosphorylated form of the regulatory domain compared to the unphosphorylated form: in the phosphorylated state a serine/threonine side chain is repositioned to form a hydrogen bond with the third phosphate oxygen and a phenylalanine/tyrosine side chain is reoriented towards the interior of the domain, filling the space that is normally occupied by the serine/threonine in the unphosphorylated protein.

Unlike the GTPases, which have large conformational changes in their switching mechanisms, the regulatory domains of two-component signaling pathways undergo less dramatic structural rearrangements on phosphorylation. Structural differences between the unphosphorylated and phosphorylated regulatory domains map to a relatively large surface involving several beta strands, alpha helices and adjacent loops, with backbone displacements ranging from 1 to 7 Å (Figure 3-30b).

In other words, like the G proteins, RR regulatory domains function as generic on–off switch modules, which can exist in two distinct structural states with phosphorylation modulating the equilibrium between them, providing a very simple and versatile mechanism for regulation. Surfaces of the regulatory domain that have altered structures in the two different conformations are exploited for protein–protein interactions that regulate effector domain function.

Not all effector domains are DNA-binding transcriptional regulators. Some function, for example, as regulators of bacterial flagellar rotation. There are many different strategies for regulation of effector domains by RR regulatory domains, including inhibition of effector domains by unphosphorylated regulatory domains, allosteric activation of effector domains by phosphorylated regulatory domains, dimerization of effector domains mediated by domain dimerization, and interaction of RRs with heterologous target effector proteins. Different RRs use different subsets of the regulatory domain surface for phosphorylation-dependent regulatory interactions. Thus, although RRs are fundamentally similar in the design of their phosphorylation-activated switch domains, there is significant versatility in the way these domains are used to regulate effector activity.

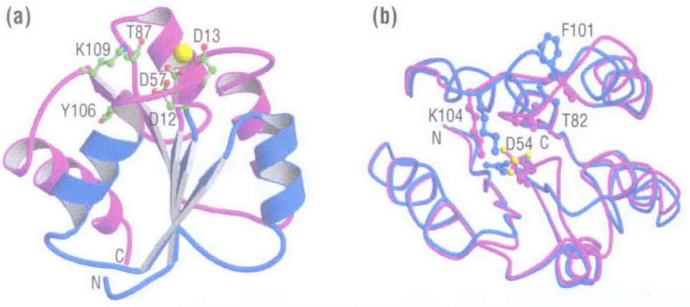


Figure 3-30 Conserved features of RR regulatory domains (a) A ribbon diagram of the bacterial RR CheY, whose protein fold is just a bare regulatory domain, is shown with ball-and-stick representations of the side chains of highly conserved residues. Residues that are highly conserved in all RR regulatory domains are clustered in two regions: an active site formed by loops that extend from the carboxy-terminal ends of strands 1, 3 and 5, and a pair of residues that form a diagonal path extending across the molecule from the active site. Three aspartic acid residues (D12, D13 and D57) position a Mg^{2+} ion (yellow) that is required for catalysis of phosphoryl transfer from an HK to aspartate 57. Three additional residues (K109, T87 and Y106) are important in propagation of the conformational change that occurs on phosphorylation. The regions of the backbone that have been observed to differ in unphosphorylated and phosphorylated regulatory domains are shown in magenta. **(b)** The conserved mechanism involved in the phosphorylation-induced conformational change is illustrated by the structures of the unphosphorylated (blue) and phosphorylated (magenta) regulatory domains of the RR protein FixJ. When D54 is phosphorylated, K104 forms an ion pair with the phosphate (yellow). T82 also forms a hydrogen bond with the phosphate and F101 has an inward orientation, positioned in the space occupied by T82 in the unphosphorylated structure. These changes trigger other rearrangements of secondary structure elements, creating a different protein-binding surface in the two states. Graphic kindly provided by Ann Stock. From West, A.H. and Stock, A.M.: *Trends Biochem. Sci.* 2001, **26**:369–376, with permission.

©2004 New Science Press Ltd

3-16 Control by Proteolysis: Activation of Precursors

Limited proteolysis can activate enzymes

Both the ubiquitin-dependent proteasome pathway discussed in section 3-11 and digestion by many other nonspecific proteases degrade polypeptide chains to small peptides or individual amino acids. But proteolysis does not have to lead to destruction and inactivation. Many proteins are post-translationally modified by limited proteolytic digestion that produces an active form from their inactive or marginally active precursors.

Limited proteolysis involves the cleavage of a target protein at no more than a few specific sites—commonly just one—usually by a specific protease. The resulting cleaved protein can have one of two fates: either the fragments remain associated, covalently (if they are disulfide-linked) or noncovalently, or they may dissociate to give two or more different polypeptides, each of which may have a completely separate fate and function. Both these outcomes are in fact observed in the maturation of the inactive precursor chymotrypsinogen to active alpha-chymotrypsin, a digestive protease (Figure 3-31). Chymotrypsinogen has the overall fold of an active serine protease with the exception that the active-site region lacks the proper configuration of main chain and catalytic side chains for both catalysis and substrate recognition. The protein is, however, "spring-loaded"; it is only the existence of a covalent bond between arginine 15 and isoleucine 16 and a set of noncovalent interactions between these residues and their neighbors that prevent it from rearranging into the correct conformation for catalysis. Specific proteolytic cleavage between arginine 15 and isoleucine 16 by the protease trypsin releases it from this constraint. After this cleavage, the amino-terminal peptide remains attached through a disulfide linkage, while the newly formed amino terminus (old isoleucine 16) swaps into a new conformation in which it makes a new set of interactions with residues of the active site, causing the active site to assume its correct catalytic configuration. The product is fully active enzymatically, although two more autocatalytic cleavages remove residues 14 and 15 and residues 147 and 148, to yield the mature form of alpha-chymotrypsin that is found in the digestive tract. The function of these latter two modifications is unknown. The activation of most serine proteases of this class from their inactive proenzyme forms occurs in a similar fashion (Figure 3-32).

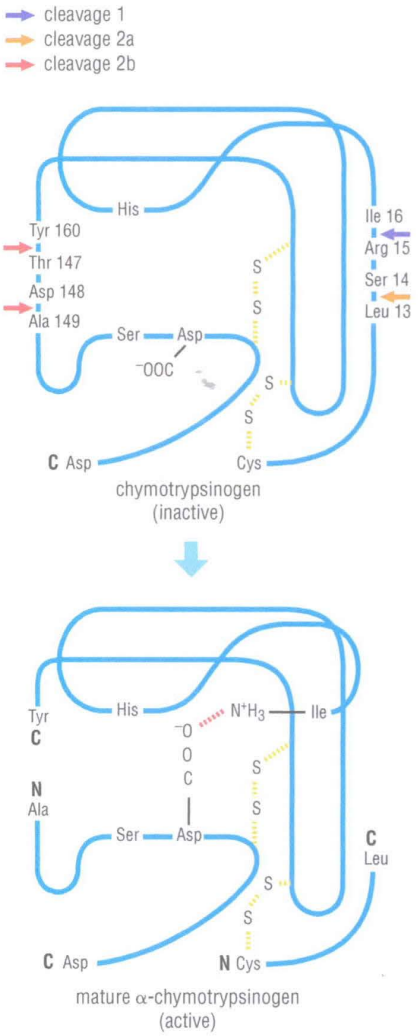


Figure 3-31 Activation of chymotrypsinogen
A schematic view of the activation of chymotrypsinogen. The polypeptide chain (blue) is held together by disulfide (–S–S–) bridges (for clarity only two of the five are shown). Cleavage between residues 15 and 16 (top) results in a rearrangement of part of the polypeptide chain (bottom) and the formation of a new interaction between isoleucine 16 and an aspartate elsewhere in the chain, which helps form the functional active site. After the first cleavage, the first 15 residues remain attached to the rest of the protein, in part by a disulfide linkage. Two more cleavage events remove residues 14–15 and residues 147–148 to produce the mature alpha-chymotrypsin.

Polypeptide hormones are produced by limited proteolysis

Limited proteolysis can also produce polypeptides with new functions, as in the production of short polypeptide hormones from long precursor proteins (Figure 3-33). All known polypeptide hormones are synthesized in "prepro" form, with a signal (pre) sequence and additional (pro) sequences that are cleaved out during maturation. Often, a single precursor sequence may contain two or more distinct hormones, each of which is released by additional cleavages. A striking example is the pituitary multihormone precursor pro-opiomelanocortin, which contains sequences for the hormones beta-lipotropin, melanocyte-stimulating hormone (MSH), endorphin, enkephalin, and adrenocorticotropic hormone (ACTH, corticotropin). ACTH is a 39-residue polypeptide that stimulates the synthesis of mineralocorticoids and glucocorticoids in the adrenal cortex, which leads to activation of steroid hormone synthesis. The other hormone products have very different target tissues and different physiological activities. Pro-opiomelanocortin is cleaved at different sites in different cell types so that they produce different spectra of hormones derived from the single precursor. Each of the cleaving enzymes is specific for a particular sequence in the precursor protein, by virtue either of its active-site structure or of the compartment in which it encounters its substrate. Hormone-processing proteases are attractive targets for drugs designed to modulate hormone-sensitive processes such as control of blood pressure and inflammation.

Definitions

proteolytic cascade: a series of protein cleavages by proteases, each cleavage activating the next protease in the cascade.

©2004 New Science Press Ltd

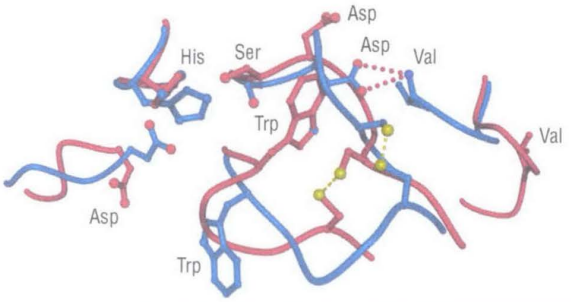


Figure 3-32 Comparison of the active sites of plasminogen and plasmin Plasminogen (red) is the inactive precursor of plasmin (blue), the primary enzyme that dissolves blood clots. Like chymotrypsinogen and many other inactive serine protease precursors, plasminogen is activated by a proteolytic cleavage. The active-site residues in plasminogen must rearrange to form an active enzyme. In chymotrypsinogen, most of the catalytic residues are already in the active configuration, but in plasminogen the entire active site is distorted. The primary cause of this distortion is a tryptophan residue that blocks the substrate-binding pocket and pushes the catalytic residues away from their correct orientations. As in all proteases of this class, the oxyanion hole involved in catalysis (see Figure 2-40) has not yet formed. Cleavage of the arginine–valine bond (not shown) in plasminogen creates a new positively charged amino-terminal amino group that interacts with the central aspartic acid residue of the oxyanion peptide, causing the peptide to rearrange into the correct configuration of the oxyanion hole. At the same time, the tryptophan swings out of the active site, unblocking the substrate-binding pocket and allowing the catalytic residues (His, Ser and Asp at the left of the diagram) to switch to the correct configuration for catalysis.

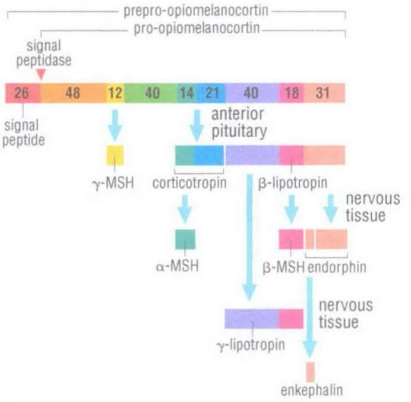


Figure 3-33 Schematic diagram of prepro-opiomelanocortin and its processing Shown are the domains that comprise the large pre-pro hormone precursor opiomelanocortin, with the number of amino acids in each domain indicated. This primary gene translation product is cleaved into many different smaller hormones by a series of proteolytic steps. First, the signal peptide that directed its secretion is cleaved off by a signal peptidase, generating pro-opiomelanocortin. Then, a number of other cleavages, which are tissue-specific, occur in different cell types to produce different ensembles of hormones. For instance, in the pituitary, cleavage occurs to generate ACTH and beta-lipotropin, while processing in the central nervous system gives endorphin and enkephalin, among other products.

Limited proteolysis can provide for enormous, extremely rapid amplification of a signal if the product of one cleavage is an active protease that can go on to activate other proteins by cleavage. Several such activations may follow one another to generate a **proteolytic cascade** in which the initial activation of a single molecule of an inactive proenzyme produces a huge final output. This is how the blood coagulation pathway works. Regardless of whether the pathway is activated internally, as in thrombosis, or at an open wound, the first step is the production of an active, specific protease from an inactive precursor. A single molecule of this activated protease is capable of activating thousands of molecules of its specific substrate, another inactive protease; each newly activated protease molecule can then activate thousands of molecules of their specific proenzyme substrate, and so on. The final product is millions of activated thrombin molecules, which cleave the soluble blood protein fibrinogen to form insoluble fibrin, which polymerizes to form the clot (Figure 3-34).

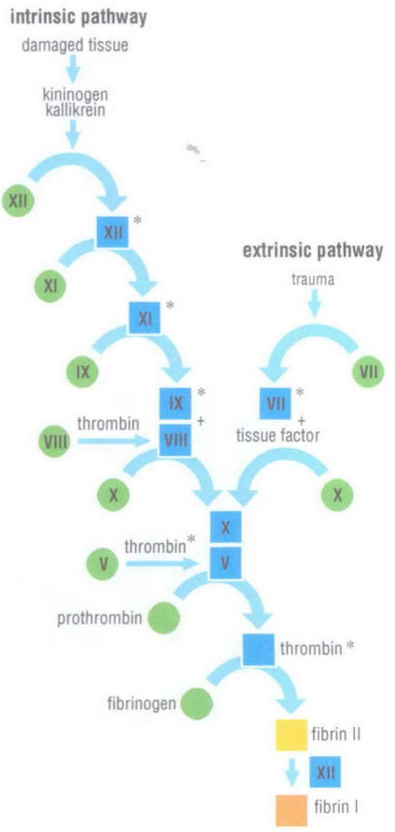


Figure 3-34 The blood coagulation cascade Each factor in the pathway is a protein that can exist in an inactive form (green circles) and an active form (blue squares), and many of these are serine proteases (denoted by an asterisk). The cascade of proteolytic activations that ends with the activation of fibrinogen to clotting fibrin is initiated by exposure of blood at damaged tissue surfaces (intrinsic pathway) or from internal trauma to blood vessels (extrinsic pathway). Although the specific initiating factors differ in the two pathways, in both cases the process begins with the conversion of an inactive serine protease to an active one. As little as one molecule of this active protease may be all that is needed to activate thousands of molecules of the next precursor protein in the cascade. Each of these can activate many more of the next, and so on down the cascade. Each protease is highly specific for a cleavage sequence in the next factor in the pathway, preventing unwanted activation. Roman numerals refer to specific proteins called factors.

References

Hook, V.Y. *et al.*: **Proteases and the emerging role of protease inhibitors in prohormone processing.** *FASEB J.* 1994, **8**:1269–1278.

Kalafatis, M. *et al.*: **The regulation of clotting factors.** *Crit. Rev. Eukaryot. Gene Expr.* 1997, **7**:241–280.

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

Rockwell, N.C. *et al.*: **Precursor processing by kex2/furin proteases.** *Chem. Rev.* 2002, **102**: 4525–4548.

Steiner, D.F.: **The proprotein convertases.** *Curr. Opin. Chem. Biol.* 1998, **2**:31–39.

©2004 New Science Press Ltd

3-17 Protein Splicing: Autoproteolysis by Inteins

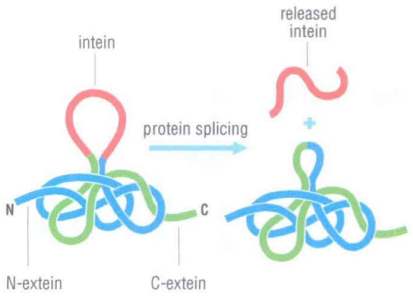
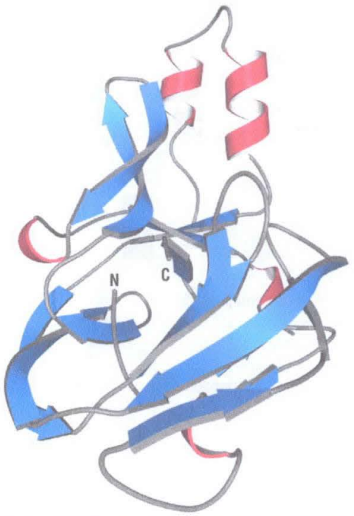


Figure 3-35 Protein splicing Schematic diagram of the process by which an internal segment of a polypeptide chain (intein) excises itself from the protein in which it is embedded, leading to the ligation of the flanking domains to yield a single polypeptide chain. Red indicates sequence to be excised (intein); blue and green are the flanking sequences. Protein splicing does not involve peptide bond breakage and religation but instead occurs entirely by peptide bond rearrangements. This allows it to proceed without a source of metabolic energy such as ATP, which would be required if the precursor protein were cleaved and religated.

Figure 3-36 Schematic of the organization of intein-containing proteins The intein is represented by the peptide segment labeled A to G. This segment contains two splicing domains separated by a linker region that may be a homing endonuclease. Sites A, B and G represent conserved sequences important for self-splicing. Red indicates sequence to be excised (intein); blue and green are the flanking sequences.



Some proteins contain self-excising inteins

Several genes from prokaryotes and lower eukaryotes contain an in-frame open reading frame encoding an internal domain that is subsequently cleaved out from the protein to form an independent protein known as an **intein**. This process is analogous to the splicing out of introns in messenger RNA, but occurs post-translationally at the level of the protein itself. In protein splicing, internal segments in the translation product (inteins) excise themselves from the protein while ligating the flanking polypeptides (**exteins**) to form the final protein. This process generates two functional proteins from one polypeptide chain, one derived from an internal sequence in the other (Figure 3-35). The novelty of this protein splicing process is the fact that it is self-catalyzed. To date, more than 100 inteins have been discovered, approximately 70% of which reside in proteins involved in DNA replication and repair. Inteins have been found that contain as few as 134 amino acids and as many as 608. Inteins are usually composed of two distinct regions, a protein-splicing domain, which is split into two segments, and an intervening endonuclease domain (Figure 3-36). (In smaller inteins this endonuclease domain may be replaced by a short peptide linker.) The protein-splicing domain catalyzes a series of peptide bond rearrangements that lead to the excision of the entire intein from the protein as well as religation of the flanking exteins. This protein-splicing activity is autoproteolytic and requires no accessory host proteins or cofactors.

As far as is known, inteins are mobile genetic elements with no function other than their own propagation. This function depends on the endonuclease domain. It belongs to a family of enzymes called homing endonucleases. These cleave the DNA sequence corresponding to their own amino-acid sequence from the genome and then mediate its insertion into genes that lack the intein sequence. Why these elements should prefer to exist as inteins in DNA replication and repair proteins in particular is unclear.

Mutant inteins that actually cleave the protein sequence in which they occur, a reaction that does not occur in protein splicing under physiological conditions, are proving useful in protein chemistry and protein engineering. For example, tobacco plants have been genetically engineered to express a mammalian antimicrobial peptide as a fusion protein with a modified intein from vacuolar membrane ATPase. The peptide can then be purified from the plant tissue by taking advantage of the intein-mediated self-cleaving mechanism. Inteins can also be used in the semisynthesis of entire proteins for industrial and medical applications: part of the protein can be synthesized chemically, allowing for the introduction of unnatural amino acids and chemical labels, and part of it can be produced genetically. Intein sequences can then be used to couple these parts together.

The mechanism of autocatalysis is similar for inteins from unicellular organisms and metazoan Hedgehog protein

Both inteins and each of the different types of homing endonucleases are characterized by a few short signature sequence motifs and can be recognized in gene sequences. Certain amino acids are highly conserved within the splicing domain. For instance, the first amino acid on

Figure 3-37 Structure of an intein Ribbon diagram of the polypeptide chain fold of the intein from the gyrase A subunit of *Mycobacterium xenopi*. It contains an unusual beta fold in which the catalytic splice junctions are located at the ends of two adjacent antiparallel beta strands at the amino and carboxyl termini of the intein. (PDB 1am2)

Definitions

intein: a protein intron (intervening sequence). An internal portion of a protein sequence that is post-translationally excised in an autocatalytic reaction while the flanking regions are spliced together, making an additional protein product.

extein: the sequences flanking an **intein** and which are religated after intein excision to form the functional protein.

References

Chevalier, B.S. and Stoddard, B.L.: **Homing endonucleases: structural and functional insight into the catalysts of intron/intein mobility.** *Nucleic Acids Res.* 2001, **29**:3757–3774.

Hall, T.M., *et al.*: **Crystal structure of a Hedgehog autoprocessing domain: homology between Hedgehog and self-splicing proteins.** *Cell* 1997, **91**:85–97.

Jeong, J. and McMahon, A.P.: **Cholesterol modification**

©2004 New Science Press Ltd

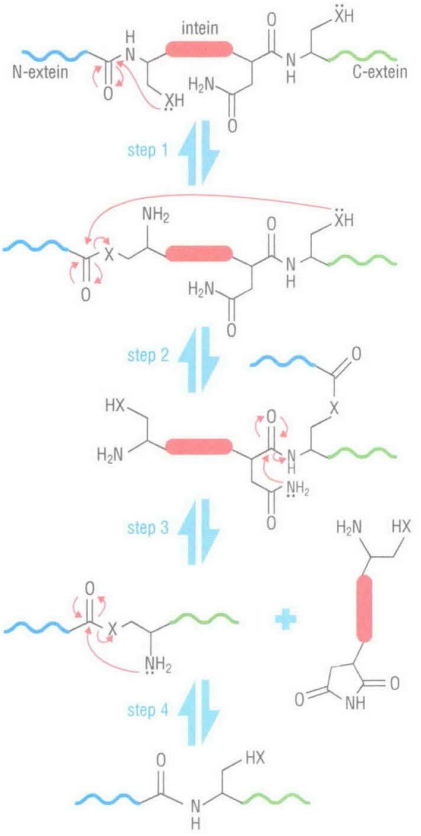


Figure 3-38 Four-step mechanism for protein splicing In step 1, a side-chain oxygen or sulfur (X) of the first intein residue (serine, cysteine or threonine; A in Figure 3-36) attacks the carbonyl group of the peptide bond that it makes with the preceding amino acid. In step 2, this carbonyl, now in the form of an ester or thioester, is attacked by the first residue of the carboxy-terminal extein segment, which is also serine, cysteine or threonine. The last residue of the intein, which is most commonly an asparagine (corresponding to position G in Figure 3-36) then cyclizes internally through its own peptide carbonyl group (step 3), thereby releasing both the intein (with a cyclic asparagine at its carboxyl terminus) and the extein, in which the amino-terminal and carboxy-terminal segments are connected via the side chain of the first residue of the carboxy-terminal segment. The spontaneous rearrangement of this ester or thioester to a normal peptide bond (step 4) completes the splicing process.

both the amino terminus of the intein domain (indicated by an A in Figure 3-36) and the carboxy-terminal extein fragment is either serine, cysteine or threonine, although an alanine has sometimes been observed. The amino acid on the carboxyl terminus of the intein (indicated by a G in Figure 3-36) is most commonly asparagine, often preceded by histidine. Finally, a ThrXXHis sequence within the amino-terminal splicing domain of the intein (indicated by a B in Figure 3-36) is usually observed. These residues have specific roles in the structure of the intein and the mechanism of protein splicing. The three-dimensional structure of an intein has now been solved for a mycobacterial protein and contains an unusual beta fold with the catalytic splice junctions at the ends of two adjacent beta strands (Figure 3-37). The arrangement of the active-site residues Ser 1 (corresponding to A in Figure 3-36), Thr 72-X-X-His 75 (corresponding to B in Figure 3-36), His 197 and Asn 198 (corresponding to G in Figure 3-36) is consistent with a proposed four-step mechanism for the protein splicing reaction (Figure 3-38).

The reaction carried out by inteins is similar to that performed by the eukaryotic Hedgehog proteins (HH), which play a central part in developmental patterning in vertebrates and invertebrates and are implicated in some important human tumors. Depending on the context, HH signals can promote cell proliferation, prevent programmed cell death, or induce particular cell fates. HH family members can exert their effects not only on cells neighboring the source of the protein signal, but also over considerable distances (up to 30 cell diameters), acting in at least some cases as classic morphogens. Such morphogens are signaling molecules that diffuse from a source to form a concentration gradient over an extended area of the target field and elicit different responses from cells according to their position within the gradient, which reflects the dosage of the ligand they are exposed to. In many cases, the precise biochemical function of the HH protein is unknown.

Like inteins, which autocatalyze their removal from inside other proteins with the concomitant joining of their flanks by a peptide bond, the carboxy-terminal domain of HH proteins autocatalyzes their cleavage from the amino-terminal domain of precursor proteins with a concomitant covalent attachment of a cholesterol molecule to the cleavage point on the amino-terminal domain. This targets the HH protein to the membrane, a process that is essential to its signaling function. The autocatalytic reactions are similar, with the cleaved peptide bond (the amino-terminal one in inteins) first changed to an ester/thioester bond that is cleaved by a nucleophilic attack from the carboxy-terminal flanking sequence in inteins or the cholesterol molecule in HH. Although these two types of self-splicing proteins are found in different types of organisms (single-celled organisms for inteins and metazoans for HH proteins), they have significant sequence similarity and their core three-dimensional structures are very similar (compare Figures 3-37 and 3-39).

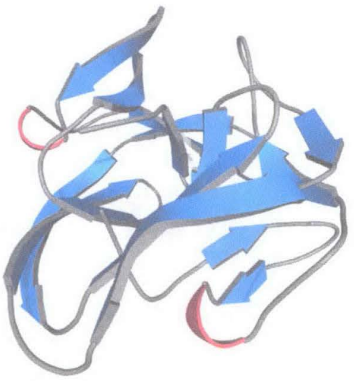


Figure 3-39 Structure of part of the Hedgehog carboxy-terminal autoprocessing domain The carboxy-terminal domain of *Drosophila* Hedgehog protein (HH-C) possesses an autoprocessing activity that results in an intramolecular cleavage of full-length Hedgehog protein and covalent attachment of a cholesterol molecule to the newly generated amino-terminal fragment. The 17 kDa fragment of HH-C (HH-C17) shown here is active in the initiation of autoprocessing. The HH-C17 structure comprises two homologous subdomains that appear to have arisen from tandem duplication of a primordial gene. Both the protein structure and the chemical features of the reaction mechanism are conserved between HH-C17 and the self-splicing regions of inteins (see Figure 3-37). (PDB 1at0)

of Hedgehog family proteins. *J. Clin. Invest.* 2002, **110**:591–596.

Klabunde, T. *et al.*: **Crystal structure of GyrA intein from *Mycobacterium xenopi* reveals structural basis of protein splicing.** *Nat. Struct. Biol.* 1998, **5**:31–36.

Paulus, H.: **Protein splicing and related forms of protein autoprocessing.** *Annu. Rev. Biochem.* 2000, **69**:447–496.

Perler, F.B.: **Protein splicing of inteins and hedgehog autoproteolysis: structure, function, and evolution.** *Cell* 1998, **92**:1–4.

©2004 New Science Press Ltd

3-18 Glycosylation

Figure 3-40 Immunoglobulin A protects mucosal surfaces from pathogenic organisms Secreted immunoglobulin A is the major antibody protecting mucosal surfaces of the gastrointestinal and respiratory tracts from pathogenic microorganisms. *N*- and *O*-linked sugars (yellow and orange respectively) protect vulnerable regions of the molecule from proteolysis. The *O*-glycosylated hinge region of IgA also provides binding sites for bacteria, protecting the mucosal surface from infection. Graphic kindly provided by Pauline Rudd. From Royle, L.M. *et al.*, unpublished.

Glycosylation can change the properties of a protein and provide recognition sites

Perhaps the most complex and diverse post-translational modifications to proteins are those that attach carbohydrate chains (oligosaccharides) to asparagine, serine and threonine residues on protein surfaces in the process known as **glycosylation**. Almost all secreted and membrane-associated proteins of eukaryotic cells are glycosylated, with oligosaccharides attached to the polypeptide chain at one or more positions by either *N*- or *O*-glycosidic bonds. There are two major types of functions for sugars attached to proteins. First, specific oligosaccharides provide recognition sites that tag glycoproteins for recognition by other proteins both inside and outside the cell and thus direct their participation in various biological processes. One important example of this is the targeting of leukocytes to sites of inflammation. Second, since oligosaccharides can be as large as the protein domains to which they are attached, they shield large areas of the protein surface, providing protection from proteases (Figure 3-40) and nonspecific protein–protein interactions. Inside the cell, oligosaccharides increase the solubility of nascent glycoproteins and prevent their aggregation.

The nature of the modification on any given protein varies from species to species and often from one tissue to the next. Glycoproteins exist as populations of glycoforms in which the same amino-acid sequence is modified by a glycosidic core with a range of different derivatives at each glycosylation site. Protein glycosylation is almost exclusively a eukaryotic property and the complexity of the modification increases as one proceeds up the evolutionary tree. Simple eukaryotes like yeast attach only a simple set of sugars; mammals modify their proteins with highly branched oligosaccharides composed of a wider range of carbohydrates. The commonest modifications are shown in Figure 3-41. The most common *N*-linked attachment consists of two *N*-acetylglucosamines with three mannoses attached; the most common *O*-linked attachment involves an *N*-acetylgalactosamine, beta-galactose core attached at serine or threonine.

For *N*-linked sugars, the core in most eukaryotes is the same, and derives from a larger preformed precursor, Glc3Man9GlcNAc2. As the protein is being synthesized in the endoplasmic reticulum (ER) the precursor is transferred to asparagine residues (N) in the glycosylation signal sequence NXS (Figure 3-42). To produce the mature oligosaccharides, the *N*-glycosylation core is first trimmed in the ER by glucosidases and mannosidases, and additional processing of the oligomannose structures can then occur in the Golgi complex (Figure 3-42), giving rise to diverse complex sugars.

In contrast to *N*-linked sugars, all of which have the same core structure, *O*-linked sugars can have at least eight different core structures, attached through *N*-acetylgalactosamine to the side chains of serine or threonine residues; these are extended by the addition of monosaccharide units.

Some viruses and bacteria use carbohydrates on cell-surface glycoproteins as Trojan horses to gain entry into the cell: they have proteins that bind specifically to these oligosaccharides, a process that leads to internalization. Antibiotics that change the structure of these carbohydrates prevent this binding, and a number of antibiotics act by blocking enzymes in glycoprotein biosynthesis, including 1-deoxymannojirimycin, which inhibits mannosidases in both the Golgi complex and the ER.

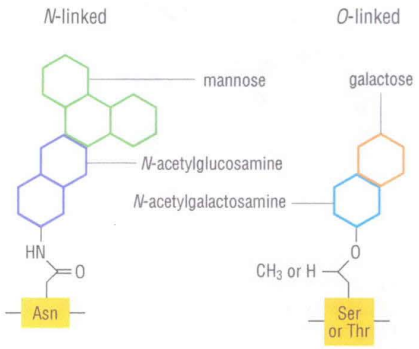


Figure 3-41 Schematic representation of the core *N*-linked oligosaccharide and a representative *O*-linked core oligosaccharide

Definitions

glycosylation: the post-translational covalent addition of sugar molecules to asparagine, serine or threonine residues on a protein molecule. Glycosylation can add a single sugar or a chain of sugars at any given site and is usually enzymatically catalyzed.

References

Bruckner, K. *et al.*: **Glycosyltransferase activity of Fringe modulates Notch-Delta interactions.** *Nature* 2000, **406**:411–415.

Imperiali, B. and O'Connor, S.E.: **Effect of *N*-linked glycosylation on glycopeptide and glycoprotein structure.** *Curr. Opin. Chem. Biol.* 1999, **3**:643–649.

Parekh, R.B. and Rohlff, C.: **Post-translational modification of proteins and the discovery of new medicine.** *Curr. Opin. Biotechnol.* 1997, **8**:718–723.

Petrescu, T. *et al.*: **The solution NMR structure of glycosylated *N*-glycans involved in the early stages of glycoprotein biosynthesis and folding.** *EMBO J.* 1997, **16**:4302–4310.

Rudd, P.M. and Dwek, R.A.: **Glycosylation: heterogeneity and the 3D structure of proteins.** *Crit. Rev. Biochem. Mol. Biol.* 1997, **32**:1–100.

Wells, L. *et al.*: **Glycosylation of nucleoplasmic proteins: signal transduction and *O*-GlcNac.** *Science* 2001, **291**:2376–2378.

©2004 New Science Press Ltd

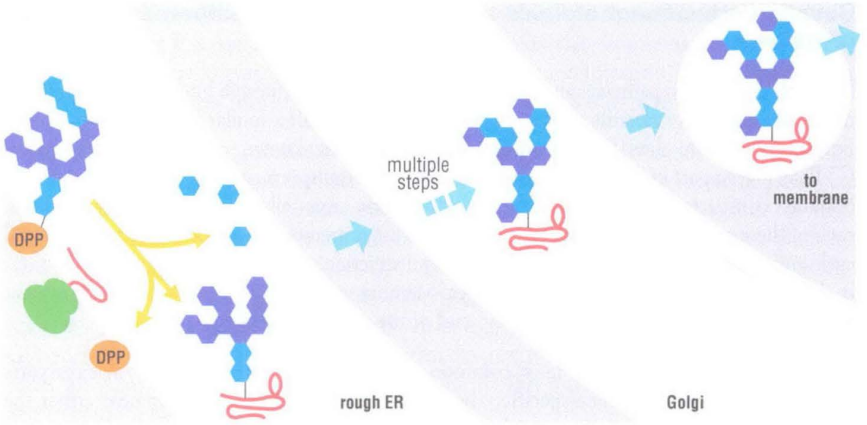


Figure 3-42 Oligosaccharide processing Schematic pathway of oligosaccharide processing on newly synthesized glycoproteins, The oligosaccharide is first transferred from its membrane-bound dolichol pyrophosphate (DPP) carrier to a polypeptide chain while the latter is still being synthesized on the ribosome. As the polypeptide chain grows within the endoplasmic reticulum (ER), monosaccharides may be cleaved from the nonreducing ends of the oligosaccharide. After completion of the synthesis of the polypeptide chain, the immature glycoprotein is carried in a transport vesicle to the Golgi apparatus where further modification occurs. New monosaccharides may be added and others removed in a multi-step process involving many different enzymes. The completed glycoprotein is finally transported to its ultimate destination.

The importance of the proper glycosylation of glycoproteins can be seen in mice lacking *N*-acetylglucosaminyltransferase I (the initial enzyme in the processing of hybrid or complex *N*-linked sugars), which die at mid-gestation, and in patients with congenital deficiencies in glycosylation, who have many pathologies ranging from neurological disorders to motor and ocular problems. Oligosaccharides attached to proteins have many roles. They may modify the activity of an enzyme, mediate the proper folding and assembly of a protein complex or facilitate the correct localization of glycoproteins; and they can be immunogenic. Glycosylation serves as an identity badge for proteins, especially those on the cell surface, indicating what type of cell or organism they come from, distinguishing mammals from their pathogens. So characteristic is this badge that many viruses bind to specific oligosaccharides as a mechanism for targeting the type of cell they infect. Because asparagine, serine and threonine-linked carbohydrates are nearly always on the outside of a folded protein they contribute to protein–protein recognition in many processes, including cell adhesion during development, immune surveillance, inflammatory reactions, and the metastasis of cancer cells. Yet there are many examples of glycoproteins that can be enzymatically stripped of their attached sugars with no effect on their biochemical function *in vitro*; in other cases, mutational loss of the oligosaccharide attachment sites has no apparent cellular consequence, and many eukaryotic proteins can be expressed in bacteria, which perform no glycosylation, and yet retain full activity *in vitro*. While this may suggest that not all glycosylations are important, it should be borne in mind that a protein enters into many interactions during its lifetime and usually only one functional assay is used to determine whether or not the sugars are important. One physical property that is dramatically affected by glycosylation is solubility: many glycoproteins become very insoluble when their sugar residues are cleaved off.

The structures of some of the most important oligosaccharide complexes have been determined by X-ray diffraction or, more commonly, by NMR. For instance, the structure of Glc3Man9GlcNac2, the precursor oligosaccharide in protein glycosylation, is shown in Figure 3-43. No obvious rules for the structures of these attached carbohydrates have yet been deduced, and in many glycoproteins whose structures have been determined, the oligosaccharides are disordered.

Another, simpler modification deserves special mention because its function has recently been elucidated. This modification is the direct enzymatic attachment of a single *N*-acetyl-glucosamine residue to a serine or threonine residue after the protein has already folded. This glycosylation is reversible by enzymatic cleavage. The sites of monoglycosylation are often identical to the sites of phosphorylation, but the effect on activity of the modified protein is usually completely different. Thus, monoglycosylation appears to transiently block phosphorylation, and is likely to be important in regulating that process. It is important to recognize that mutations that abolish phosphorylation sites may often also remove a site of monoglycosylation, and if the effects of these modifications are different then the observed phenotype will result from blocking both of them. In many breast cancer cells, monoglycosylation levels are significantly reduced owing to an increase in the synthesis of the cleaving enzyme. Less monoglycosylation would lead to more protein phosphorylation, which is also observed in most tumors and is correlated with uncontrolled growth.

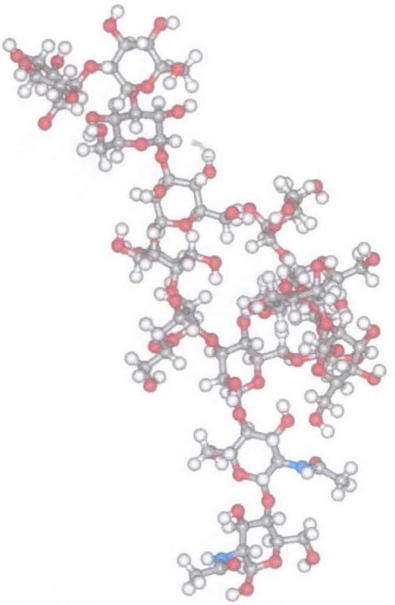


Figure 3-43 The structure of Glc3Man9GlcNac2 Average solution structure determined by NMR of the oligosaccharide involved in the early stage of glycoprotein biosynthesis. This is the 14-mer first transferred from dolichol pyrophosphate to the nascent polypeptide chain. Graphic kindly provided by Mark R. Wormald.

©2004 New Science Press Ltd

3-19 Protein Targeting by Lipid Modifications

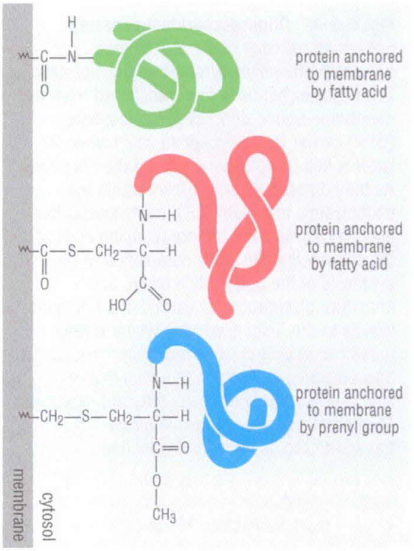


Figure 3-44 Membrane targeting by lipidation Covalent attachment of a lipid molecule to either the amino-terminal or carboxy-terminal end of a protein provides a hydrophobic anchor that localizes the protein to a membrane. The linkage is an amide (top), a thioester (middle) or a thioether (bottom) bond.

Covalent attachment of lipids targets proteins to membranes and other proteins

In signal transduction pathways and the control of protein traffic through intracellular membranes there are often many variants of one type of protein, each with a similar biochemical function. For example, more than 50 mammalian Rab proteins are known, and each of these small GTPases plays a part in directing the vesicular traffic that transports proteins between membrane-bounded compartments of the cell. To prevent chaotic cross-talk from their similar activities, each of these proteins is targeted to specific intracellular membranes at specific times in the cell cycle and during processes such as metabolism and secretion, as the need for switching functions in different compartments of the cell changes. Membrane targeting is mediated by the covalent attachment of a specific lipid group to an end of the polypeptide chain.

Lipid attachment is one of the most common post-translational modifications in eukaryotic cells. The process is sequence-specific, always involves residues either at or near either the carboxyl terminus or the amino terminus of the protein, and may require several enzymatic steps. There are four broad types of lipid modifications with distinct functional properties, classified according to the identity of the attached lipid.

These are: **myristoylation**, in which a 14-carbon fatty-acid chain is attached via a stable amide linkage to an amino-terminal glycine residue (Figure 3-44, top); **palmitoylation**, in which a 16-carbon fatty-acid chain is attached via a labile thioester linkage to cysteine residues (other fatty-acid chains can sometimes substitute for the palmitoyl group, so this modification is often called ***S*-acylation**) (Figure 3-44, middle); **prenylation**, in which a prenyl group (either a farnesyl or geranylgeranyl group) is attached via a thioether linkage to a cysteine residue initially four positions from the carboxyl terminus that becomes carboxy-terminal after proteolytic trimming and methylation of the new carboxyl terminus (Figure 3-44, bottom); and modification by a **glycosylphosphatidylinositol (GPI) anchor**, which is attached through a carbohydrate moiety (Figure 3-45).

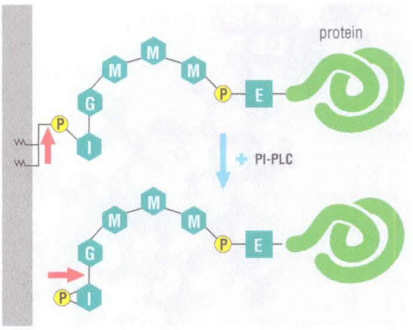


Figure 3-45 Glycosylphosphatidylinositol anchoring Schematic diagram of the reversible modification of a protein by a GPI anchor. The anchor consists of an oligosaccharide chain. The protein is connected through an amide linkage to a phosphoethanolamine molecule (E-P) that is attached to a core tetrasaccharide composed of three mannose sugars (M) and a single glucosamine sugar (G). The tetrasaccharide is in turn attached to phosphatidylinositol (I). It is the fatty-acid residues of this phosphatidylinositol group that anchor the protein to the membrane. These fatty acids may vary somewhat with the protein being modified, providing membrane specificity.

N-myristoylated proteins include select alpha subunits of heterotrimeric G proteins, a number of non-receptor tyrosine kinases, and a few monomeric G proteins, among others. The myristoyl moiety is attached to the protein co-translationally. *S*-acylated proteins include most alpha subunits of heterotrimeric G proteins, members of the Ras superfamily of monomeric G proteins, and a number of G-protein-coupled receptors, among others. *S*-acylation is post-translational and reversible, a property that allows the cell to control the modification state, and hence localization and biological activity, of the lipidated protein. Lymphoma proprotein convertase, an enzyme that activates proproteins in the secretory pathway, occurs in both the palmitoylated and unmodified form. The palmitoylated form is degraded much faster than the unmodified one, allowing the lifetime of this enzyme to be regulated by **lipidation**. There is some overlap between *N*-myristoylated and *S*-acylated proteins such that many contain both lipid modifications. Such dual modification can have important consequences, most notably targeting of the dually modified species to distinct membrane subdomains termed lipid rafts or to caveolae; this provides a way of sublocalizing proteins to microdomains that also contain specific protein–protein interaction partners.

S-prenylation occurs in two classes, those with a single prenyl group on a cysteine residue at or near the carboxyl terminus, and those that are modified on two cysteine residues at or near the carboxyl terminus. In both cases, attachment is stable and post-translational. The difference between the two types of attachment is based on the recognition motifs recognized by the

Definitions

glycosylphosphatidylinositol anchor: a complex structure involving both lipids and carbohydrate molecules that is reversibly attached to some proteins to target them to the cell membrane.

lipidation: covalent attachment of a fatty-acid group to a protein.

myristoylation: irreversible attachment of a myristoyl group to a protein via an amide linkage.

palmitoylation: reversible attachment of a palmitoyl group to a protein via a thioester linkage.

prenylation: irreversible attachment of either a farnesyl or geranylgeranyl group to a protein via a thioether linkage.

***S*-acylation:** reversible attachment of a fatty-acid group to a protein via a thioester linkage; **palmitoylation** is an example of *S*-acylation.

©2004 New Science Press Ltd

prenylating enzyme. One important protein whose activities are controlled by prenylation is the small GTPase Ras (see section 3-7). A number of *S*-prenylated proteins are also subject to *S*-acylation at a nearby cysteine residue. This type of dual modification does not apparently target the protein to the same type of membrane subdomain as dual acylation does.

A number of mammalian proteins are modified by attachment of GPI to the amino terminus. This anchor is an alternative to the lipid tails described above. GPI-anchored proteins participate in such processes as nutrient uptake, cell adhesion and membrane signaling events. All GPI-anchored proteins are destined for the cell surface via transport through the secretory pathway, where they acquire the preassembled GPI moiety. GPI modification is reversible, as the anchored protein can be released from the membrane by the action of phospholipases. In some cases, this enzyme-catalyzed release can activate an enzyme. A number of human parasites have GPI-anchored enzymes on their cell surfaces that are inactive in this state but become activated when the parasite encounters host phospholipases. By this mechanism the parasite can sense and respond to the host environment.

The enzymes involved in lipid attachment and hydrolysis are potential targets for anti-cancer, anti-fungal and other types of drugs. Several genetic diseases have been identified that involve protein lipidation. Mutations in genes encoding proteins essential for the geranyl-geranyl modification of Rab proteins are responsible for the diseases choroideremia (an incurable X-linked progressive retinal degeneration leading to blindness) and Hermansky-Pudlak syndrome (a rare disorder characterized by oculocutaneous albinism, a bleeding tendency and eventual death from pulmonary fibrosis). These diseases result from a failure to localize Rab properly.

The GTPases that direct intracellular membrane traffic are reversibly associated with internal membranes of the cell

We have already mentioned the Rab GTPases that direct transport vesicles between the membrane-bounded compartments of the cell. Two other small GTPases—Ser1 and ADP ribosylation factor (ARF)—also play an essential part in the control of vesicular traffic by recruiting the specialized coat proteins that are required for vesicle budding from the donor membrane. These small GTPases associate reversibly with their target membranes through the exposure of a covalently attached myristoyl group which occurs on exchange of GDP for GTP. We can illustrate the mechanism of vesicle formation with the ARF protein, which participates in the recruitment of vesicle coat proteins to the membrane of the Golgi complex.

ARF is myristoylated at its amino terminus and when GDP is bound this hydrophobic tail is sequestered within the protein, which therefore exists in soluble form in the cytoplasm. ARF is itself recruited to the Golgi membrane by a GTP-exchange protein, and on binding GTP, ARF undergoes a conformational change that releases the amino-terminal tail so that the protein becomes anchored to its target membrane (Figure 3-46). It is then thought to recruit preassembled coatomers—complexes of coat proteins—to the membrane, causing the coated membrane to bud off from the Golgi complex, capturing specific membrane and soluble proteins for delivery to another membrane-bounded compartment. One of the complex of coat proteins is thought to serve as a GTPase-activating protein, causing ARF to hydrolyze GTP back to GDP, withdraw its tail from the vesicle membrane, and return to the cytoplasm, with the dispersal of the coat components of the vesicle.

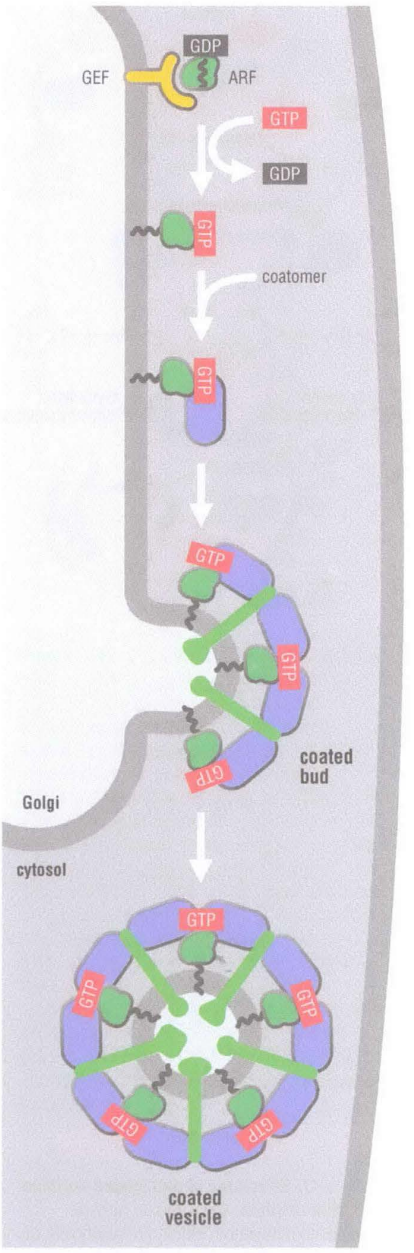


Figure 3-46 Working model for vesicular transport between Golgi compartments A key role in the process is played by the small GTPase, ARF. In its GDP-bound form, ARF is a soluble protein with a hydrophobic myristoylated tail that is tucked into the protein. When ARF is recruited to the Golgi membrane by a GTP-exchange factor, GDP is exchanged for GTP and the amino-terminal region of ARF rearranges, releasing the tail which targets ARF to the membrane. Membrane-bound ARF then recruits coat proteins necessary for vesicle budding and transport, trapping membrane and soluble proteins (not shown) in the vesicle as it pinches off.

References

Butikofer, P. *et al.*: **GPI-anchored proteins: now you see 'em, now you don't.** *FASEB J.* 2001, **15**:545–548.

Casey, P.J.: **Lipid modifications** in *Encyclopedic Reference of Molecular Pharmacology.* Offermanns, S. and Rosenthal, W. eds (Springer-Verlag, Heidelberg, 2003) in the press.

Casey, P.J.: **Protein lipidation in cell signaling.** *Science* 1995, **268**:221–225.

Chatterjee, S. and Mayor, S.: **The GPI-anchor and protein sorting.** *Cell. Mol. Life Sci.* 2001, **58**:1969–1987.

Park, H.W. and Beese, L.S.: **Protein farnesyltransferase.** *Curr. Opin. Struct. Biol.* 1997, **7**:873–880.

Pereira-Leal, J.B. *et al.*: **Prenylation of Rab GTPases: molecular mechanisms and involvement in genetic disease.** *FEBS Lett.* 2001, **498**:197–200.

Spiro, R.G.: **Protein glycosylation: nature, distribution, enzymatic formation, and disease implications of glycopeptide bonds.** *Glycobiology* 2002, **12**:43R–56R.

©2004 New Science Press Ltd

3-20 Methylation, *N*-acetylation, Sumoylation and Nitrosylation

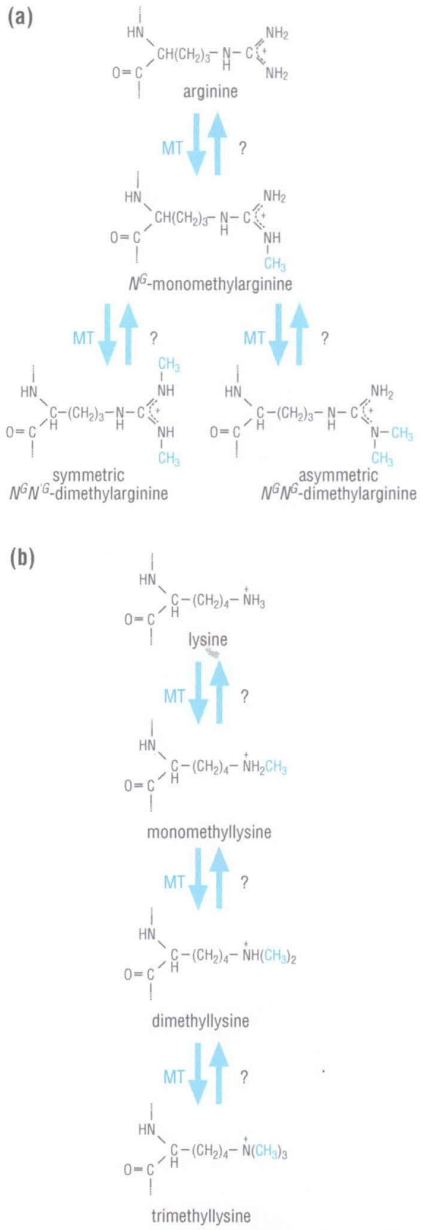


Figure 3-47 Structures of methylated arginine and lysine residues (a) Arginine may be mono- or dimethylated, either symmetrically or asymmetrically; **(b)** lysine may be mono-, di- or trimethylated. MT, methyltransferase.

Fundamental biological processes can also be regulated by other post-translational modifications of proteins

Phosphorylation, glycosylation, lipidation, and limited proteolysis are the commonest post-translational covalent modifications of proteins. However, important regulatory functions are also performed by methylation, *N*-acetylation, attachment of SUMO and nitrosylation.

Methylation occurs at arginine or lysine residues and is a particularly common modification of proteins in the nuclei of eukaryotic cells. Methylation of eukaryotic proteins is performed by a variety of methyltransferases, which use *S*-adenosylmethionine as the methyl donor. Three main forms of methylarginine have been found in eukaryotes: N^G-monomethylarginine, N^G,N^G-asymmetric dimethylarginine, and N^G,N'^G-symmetric dimethylarginine (Figure 3-47a); lysine may be mono-, di-, or trimethylated (Figure 3-47b). In contrast to phosphorylation, methylation appears to be irreversible: methylated lysine and arginine groups are chemically stable, and no demethylases have been found in eukaryotic cells. Thus regulation of this modification must occur through regulation of methyltransferase activity, and removal of the methylated proteins themselves. Arginine methylation most commonly occurs at an RGG sequence, and less often at other sites such as RXR and GRG. Although methylation does not change the overall charge on an arginine residue, it greatly alters the steric interactions this group can make and eliminates possible hydrogen-bond donors. It is not surprising, therefore, that methylation has been shown to alter protein–protein interactions. For example, asymmetric methylation of Sam68, a component of signaling pathways that recognizes proline-rich domains, decreases its binding to SH3-domain-containing but not to WW-domain-containing proteins. Arginine methylation is particularly prevalent in heterogeneous nuclear ribonucleoproteins (hnRNPs), which have roles in pre-mRNA processing and nucleocytoplasmic RNA transport. A second centrally important group of nuclear proteins to undergo methylation are the histones that package chromosomal DNA in the DNA–protein complex known as **chromatin**. Histone methylation on lysine by histone methyltransferases changes the functional state of chromatin in the region of the modification, with important effects on gene expression and DNA replication and repair. Some of these are thought to be due to effects on the compaction of the chromatin, but it is clear that others depend on the recruitment to the DNA of "silencing" proteins that recognize specific modifications to specific lysines through chromodomains characteristic of these proteins (see Figure 3-2) and suppress gene expression. Histones can also undergo phosphorylation, ubiquitination and acetylation (see below), which also affect the functions directed by chromosomal DNA.

***N*-acetylation**, which is catalyzed by one or more sequence-specific *N*-acetyltransferases, usually modifies the amino terminus of the protein backbone with an acetyl group derived from acetyl-CoA. It has been estimated that more than one-third of all yeast proteins may be so modified. This modification has a number of roles, including blocking the action of aminopeptidases and otherwise altering the lifetime of a protein in the cell. *N*-acetylation of

Figure 3-48 *N*-acetylation Acetylation by histone acetyltransferase (HAT) of amino-terminal lysine residues is an important regulatory modification of histone proteins. Deacetylation is catalyzed by histone deacetylases (HDs).

Definitions

chromatin: the complex of DNA and protein that comprises eukaryotic nuclear chromosomes. The DNA is wound around the outside of highly conserved histone proteins, and decorated with other DNA-binding proteins.

methylation: modification, usually of a nitrogen or oxygen atom of an amino-acid side chain, by addition of a methyl group. Some bases on DNA and RNA can also be methylated.

***N*-acetylation:** covalent addition of an acetyl group from acetyl-CoA to a nitrogen atom at either the amino terminus of a polypeptide chain or in a lysine side chain. The reaction is catalyzed by *N*-acetyltransferase.

nitrosylation: modification of the –SH group of a cysteine residue by addition of nitric oxide (NO) produced by nitric oxide synthase.

sumoylation: modification of the side chain of a lysine residue by addition of a small ubiquitin-like protein (SUMO). The covalent attachment is an amide bond between the carboxy-terminal carboxylate of SUMO and the NH_2 on the lysine side chain of the targeted protein.

References

Hochstrasser, M.: **SP-RING for SUMO: New functions bloom for a ubiquitin-like protein.** *Cell* 2001, **107**:5–8.

Jenuwein, T. and Allis, C.D.: **Translating the histone code.** *Science* 2001, **293**:1074–1080.

Kouzarides, T.: **Histone methylation in transcriptional control.** *Curr. Opin. Genet. Dev.* 2002, **12**:198–209.

©2004 New Science Press Ltd

the amino terminus is usually irreversible, but the epsilon amino group on the side chain of lysine residues can also be acetylated by other, specific acetyltransferases, and this reaction can be reversed by deacetylases (Figure 3-48). In contrast to methylation, which maintains the charge on the amino group, acetylation does not. Reversible *N*-acetylation at lysine in histone proteins has a major role in the control of gene expression and other chromosomal functions: while histone methylation may induce either an active or an inactive state of chromatin, depending on the position and the nature of the methyl group, histone acetylation is always associated with an active state of chromatin: this is promoted by chromatin-remodeling enzymes recruited to the DNA by proteins containing bromodomains (see Figure 3-2) that specifically recognize acetylated lysines. Thus, in contrast to methylation, which regulates chromatin by creating non-binding surfaces for regulatory proteins on core histones, acetylation may influence genome function in part through affecting higher-order protein structure. Deregulation of chromatin modification pathways is widely observed in cancer.

Covalent attachment of one protein to another is a very common post-translational modification. In addition to ubiquitination (see section 3-11), another such modification is the attachment of the ubiquitin-like protein SUMO (small ubiquitin-related modifier), called **sumoylation** (Figure 3-49). The consensus sequence for sumoylation is ψKXE (where ψ is a hydrophobic amino acid and X is any amino acid); this is in marked contrast to ubiquitination, where no consensus sequence has ever been found. In yeast, the gene coding for the sole SUMO-like protein, Smt3, is essential for progression through the cell cycle. Septin, a GTP-binding protein that is essential for cell separation, is sumoylated during the G2/M phase of the cell cycle. Like ubiquitin, SUMO is attached to the amino group of lysine residues by specific SUMO-activating and -conjugating enzymes. Attachment of SUMO to proteins has been shown to change their subcellular localization, transcriptional activity and stability: for example, the SUMO conjugate of the RanGAP1 protein binds preferentially to the nuclear pore complex and so sumoylation appears to localize this protein. Other functions of SUMO attachment are uncertain at present.

Nitrosylation is one of only two post-translational modifications conserved throughout evolution—phosphorylation being the other. Yet it has been much less studied, in part because it has been difficult to understand how specificity of action is achieved for the reversible modification of proteins by NO groups. In general, NO modifies the –SH moiety of cysteine residues (Figure 3-50) and reacts with transition metals in enzyme active sites. Over 100 proteins are known to be regulated in this way. The majority of these are regulated by reversible *S*-nitrosylation of a single critical cysteine residue flanked by an acidic and a basic amino acid or by a cysteine in a hydrophobic environment. Cysteine residues are important for metal coordination, catalysis and protein structure by forming disulfide bonds. Cysteine residues can also be involved in modulation of protein activity and signaling events via other reactions of their –SH groups. These reactions can take several forms, such as redox events (chemical reduction or oxidation), chelation of transition metals, or *S*-nitrosylation. In several cases, these reactions can compete with one another for the same thiol group on a single cysteine residue, forming a molecular switch composed of redox, NO or metal ion modifications to control protein function. For example, the JAK/STAT signaling pathway is regulated by NO at multiple loci. NO is a diffusible gas, which modifies reactive groups in the vicinity of its production, which is catalyzed by the enzyme nitric oxide synthase. Thus, the localization of this enzyme is a key factor in which proteins will be susceptible to modification by NO. The mechanism by which this modification is reversed is unknown in many cases; some redox enzymes have been implicated; free glutathione may also be involved.

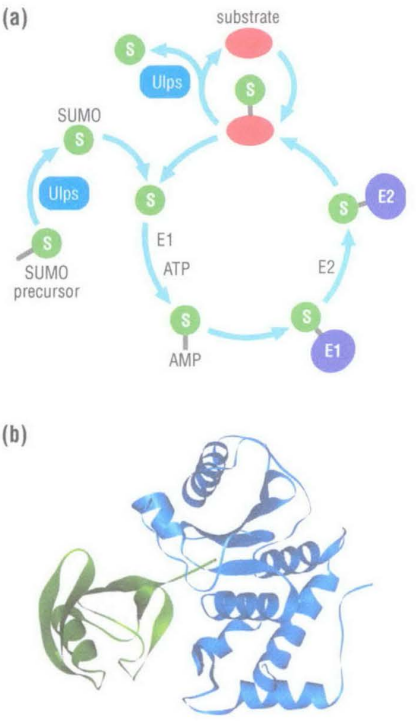


Figure 3-49 Sumoylation (a) In the SUMO cycle, a SUMO precursor is processed to SUMO by Ulp proteins. SUMO is then derivatized with AMP by E1, the SUMO-activating enzyme, before transfer of SUMO to a cysteine of E1 to form an E1–SUMO thioester intermediate. SUMO is passed to the SUMO-conjugating enzyme, E2, to form an E2–SUMO thioester intermediate. This latter complex is the proximal donor of SUMO to a substrate lysine in the ψKXE target sequence in the final substrate protein. SUMO can also be cleaved from sumoylated proteins by Ulp proteins. **(b)** The structure of the complex of the SUMO-binding domain of a SUMO-cleaving enzyme, Ulp1, with the yeast SUMO protein Smt3. SUMO is the small domain on the left. The size of the active-site cleft of Ulp1 allows even large SUMO–protein conjugates to bind and be cleaved.

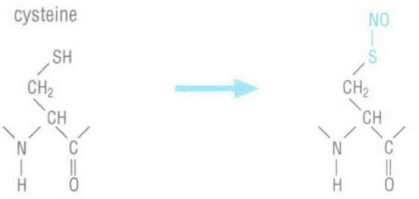


Figure 3-50 Cysteine nitrosylation

McBride, A.E. and Silver, P.A.: **State of the Arg: protein methylation at arginine comes of age.** *Cell* 2001, **106**:5–8.

Stamler, J.S. *et al.*: **Nitrosylation: the prototypic redox-based signaling mechanism.** *Cell* 2001, **106**:675–683.

Turner, M.: **Cellular memory and the histone code.** *Cell* 2002, **111**:285–291.

Workman, J.L. and Kingston, R.E.: **Alteration of nucleosome structure as a mechanism of transcriptional regulation.** *Annu. Rev. Biochem.* 1998, **67**:545–579.

Web resource on nitrosylation:

http://www.cell.com/cgi/content/full/106/6/675/DC1

©2004 New Science Press Ltd

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
4-1 Sequence Alignment and Comparison
4-2 Protein Profiling
4-3 Deriving Function from Sequence
4-4 Experimental Tools for Probing Protein Function
4-5 Divergent and Convergent Evolution
4-6 Structure from Sequence: Homology Modeling
4-7 Structure From Sequence: Profile-Based Threading and "Rosetta"
4-8 Deducing Function From Structure: Protein Superfamilies
4-9 Strategies for Identifying Binding Sites
4-10 Strategies for Identifying Catalytic Residues
4-11 TIM Barrels: One Structure with Diverse Functions
4-12 PLP Enzymes: Diverse Structures with One Function
4-13 Moonlighting: Proteins with More than One Function
4-14 Chameleon Sequences: One Sequence with More than One Fold
4-15 Prions, Amyloids and Serpins: Metastable Protein Folds
4-16 Functions for Uncharacterized Genes: Galactonate Dehydratase
4-17 Starting From Scratch: A Gene Product of Unknown Function

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.

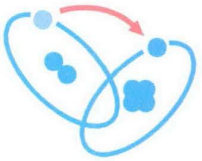
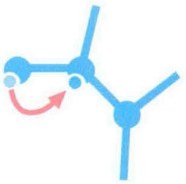
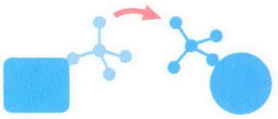
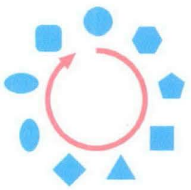
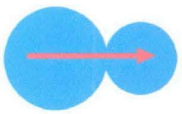
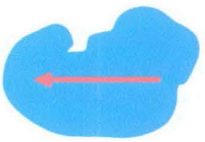
Figure 4-1 Time and distance scales in functional genomics The various levels of function of proteins encompass an enormous range of time (scale on the left) and distance (scale on the right). Depending on the time and distance regime involved, different experimental approaches are required to probe 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.

©2004 New Science Press Ltd

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.

©2004 New Science Press Ltd

```
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.

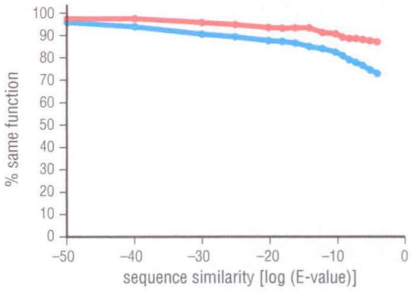


Figure 4-3 Plot of percentage of protein pairs having the same biochemical function as sequence changes When a series of sequences of homologous proteins are compared, it is observed that as sequence similarity (measured by the E-value from a sequence comparison) decreases, the probability that homologs 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.

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.

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.

©2004 New Science Press Ltd

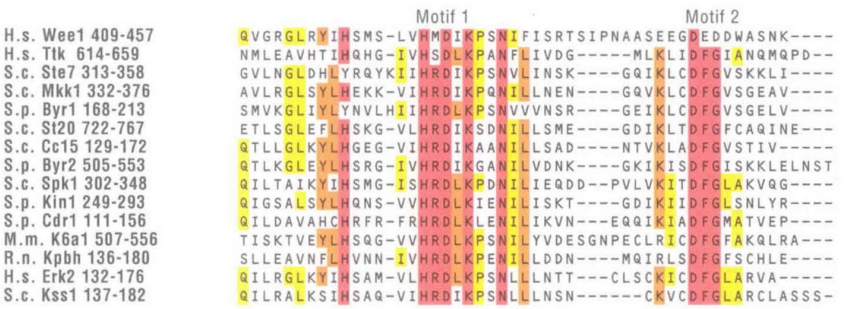


Figure 4-4 Multiple alignment A small part of a large multiple alignment of more than 6,000 protein kinase domains in the Pfam database (http://pfam.wustl.edu), displaying part of the region shown in Figure 4-2. Residues identical in all or almost all sequences in the complete alignment are highlighted 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.

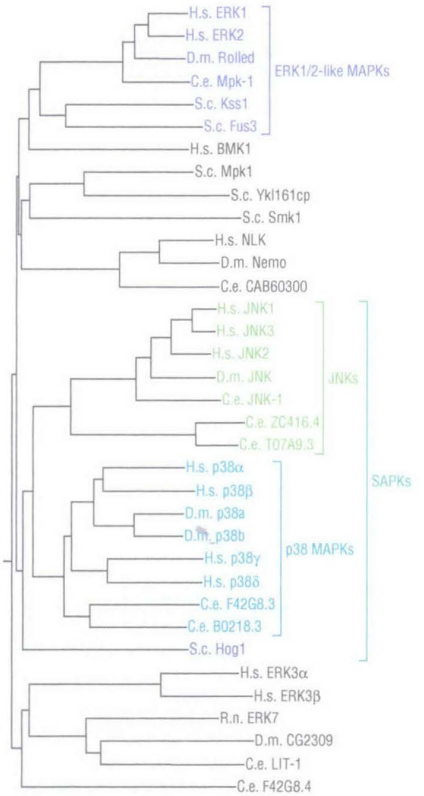


Figure 4-5 Phylogenetic tree comparing the three major MAP kinase subgroups The three major subgroups of MAP kinases (ERKs, JNKs and p38) are well conserved throughout evolution. This dendrogram shows the evolutionary relationships between the ERKs, JNKs and p38 in the budding yeast *S. cerevisiae* (S.c.), 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

©2004 New Science Press Ltd

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

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.

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

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**

©2004 New Science Press Ltd

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 RXFXYXD[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 $0.8 \times 0.6 \times 0.2 \times 0.4 \times 0.2 = 0.0077$, comparable to the score of the aligned sequence CGHSV ($0.8 \times 0.4 \times 0.2 \times 0.6 \times 0.2$). The sequence CLHTG would have a score of zero ($0.8 \times 0.0 \times 0.2 \times 0.4 \times 0.0$).

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

©2004 New Science Press Ltd

4-3 Deriving Function from Sequence

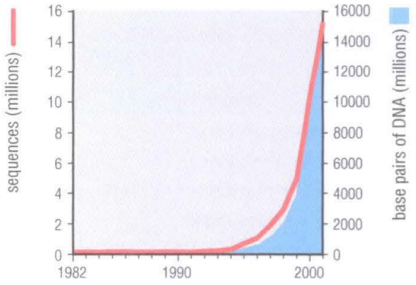


Figure 4-9 The growth of DNA and protein sequence information collected by GenBank over 20 years There has been an exponential increase in both base pairs of DNA sequence and coding sequences, especially since 1994 when various genomics projects were initiated. (Information from http://www.ncbi.nlm.nih.gov/Genbank/genbankstats.html)

Sequence information is increasing exponentially

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

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

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

In some cases function can be inferred from sequence

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

Genome Sizes of Representative Organisms

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

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

©2004 New Science Press Ltd

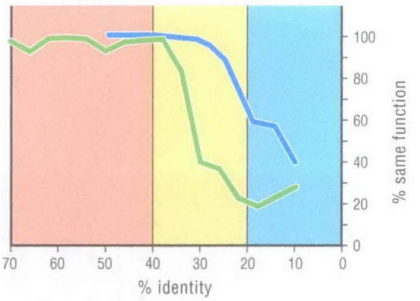


Figure 4-11 Relationship of sequence similarity to similarity of function The percentage of protein pairs with the same precise biochemical function is plotted against the sequence identity (enzymes, blue curve; non-enzymes, green curve). The orange area represents proteins whose fold and function can 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

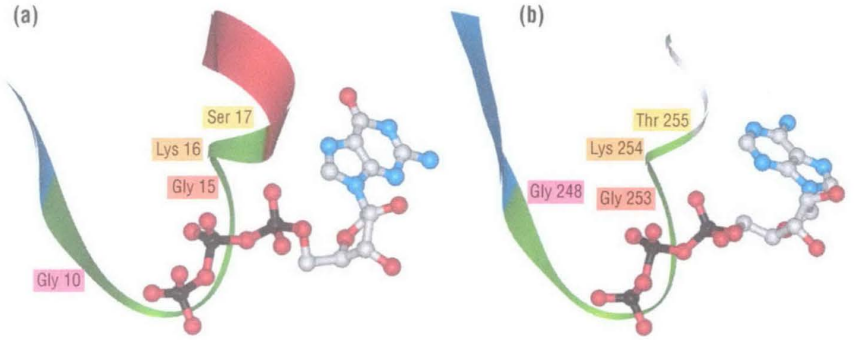


Figure 4-12 The P loop of the Walker motif A contiguous sequence block, the so-called Walker A block or P loop, is a stretch of sequence with a consensus pattern of precisely spaced phosphate-binding residues; this is found in a number of ATP- or GTP-binding proteins, for example ATP synthase, myosin 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).

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

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

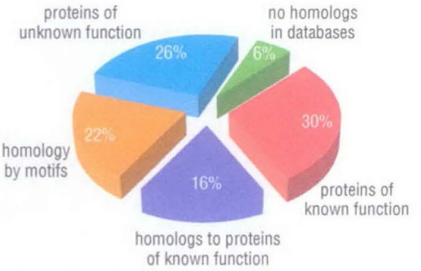


Figure 4-13 Analysis of the functions of the protein-coding sequences in the yeast genome Some are known experimentally, some are surmised from sequence comparison with proteins of known function in other organisms, and some are deduced from motifs that are characteristic of a particular function. Some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

©2004 New Science Press Ltd

4-4 Experimental Tools for Probing Protein Function

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

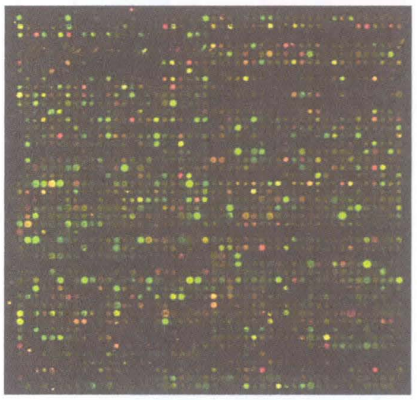


Figure 4-14 DNA microarray Part of a microarray chip showing changes in gene expression when yeast cells are treated with a drug. Genes whose expression increases on drug treatment appear as red spots; those that decrease are green; those that do not change are yellow. Some genes do not appear because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

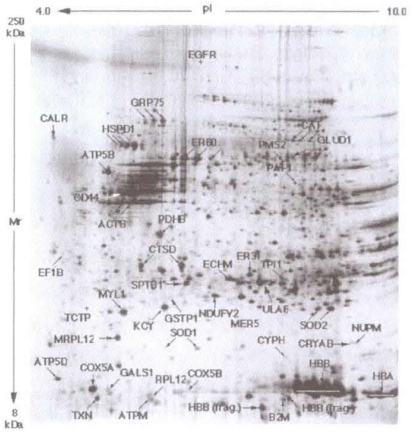


Figure 4-15 2-D protein gel Some spots have been identified and labeled.

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.

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.

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.

©2004 New Science Press Ltd

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.

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.)

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.

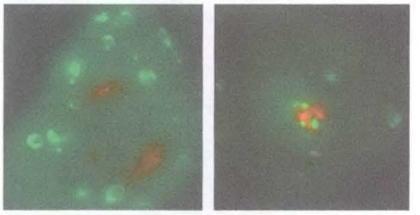


Figure 4-17 Protein localization in the cell The protein has been fused to GFP (green); the nucleus is stained red. In different stages of the cell cycle the protein is either cytoplasmic (left) or localized to the nucleus (right). Photographs kindly provided by Daniel Moore and Terry Orr-Weaver. (From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

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.

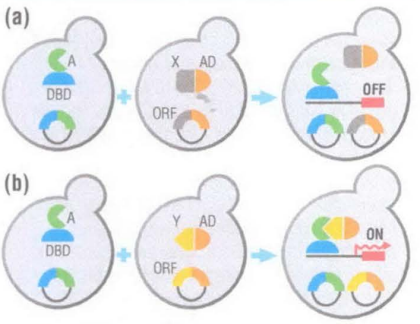


Figure 4-18 Two-hybrid system for finding interacting proteins The "bait" vector expresses a transcription factor DNA-binding domain (DBD, blue) fused to the test protein (protein A, green). The "prey" expression vectors each contain an individual open reading frame (ORF) of interest placed adjacent 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.

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.

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

©2004 New Science Press Ltd

4-5 Divergent and Convergent Evolution

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.

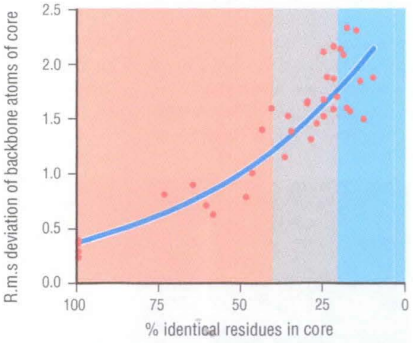


Figure 4-19 Relationship between sequence and structural divergence of proteins The percent identity of the protein cores of 32 pairs of proteins from eight different structural families was plotted against their structural divergence as measured by the root-mean-square difference in spatial positions 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.

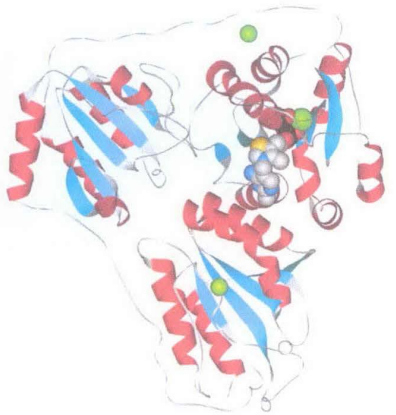
Irving, J.A. *et al.*: **Protein structural alignments and functional genomics.** *Proteins* 2001, **42**:378–382.

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.

©2004 New Science Press Ltd



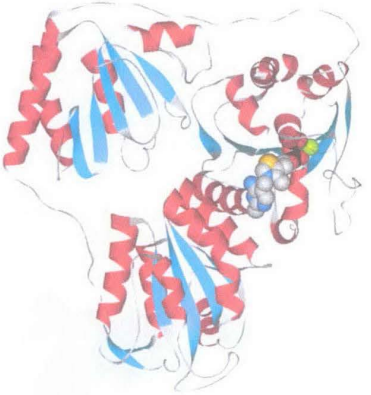


Figure 4-20 Ribbon diagram of the structure of a monomer of benzoylformate decarboxylase (BFD) and pyruvate decarboxylase (PDC) BFD (top) and PDC (bottom) share a common fold and overall biochemical function, but they recognize different substrates and have low (21%) sequence identity. The bound thiamine 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).

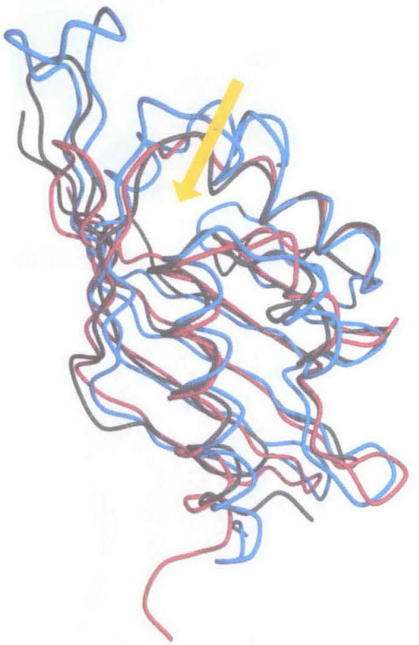


Figure 4-21 Superposition of the three-dimensional structures of steroid-delta-isomerase, nuclear transport factor-2 and scytalone dehydratase The active site is indicated by an arrow. (PDB 8cho, 1oun, 1std)

©2004 New Science Press Ltd

4-6 Structure from Sequence: Homology Modeling

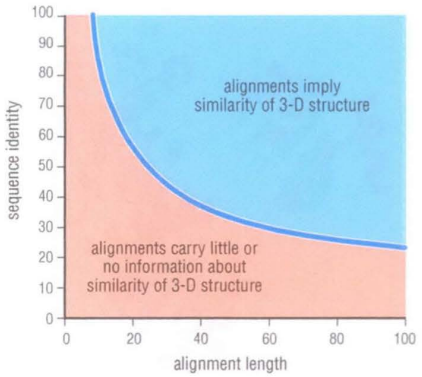
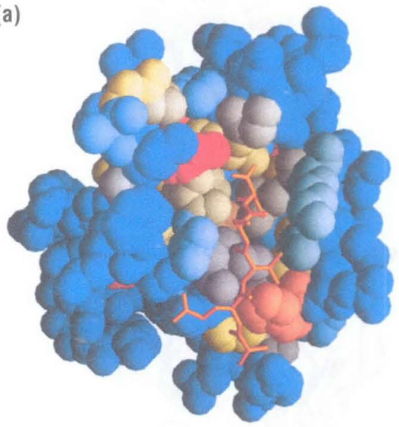


Figure 4-22 The threshold for structural homology Sequence space, plotted as a function of length of the segment being aligned and the percent identity between the two sequences, can be divided into two regions. The upper region (above the curve) shows where sequence similarity is likely to yield enough 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.

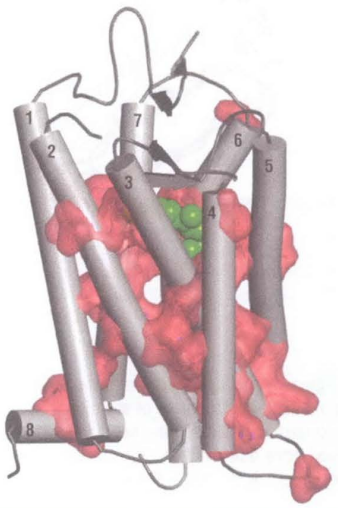
(a)



Gstat

0.000 1.500 3.000 4.500 6.000

(b)



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-**

©2004 New Science Press Ltd

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.

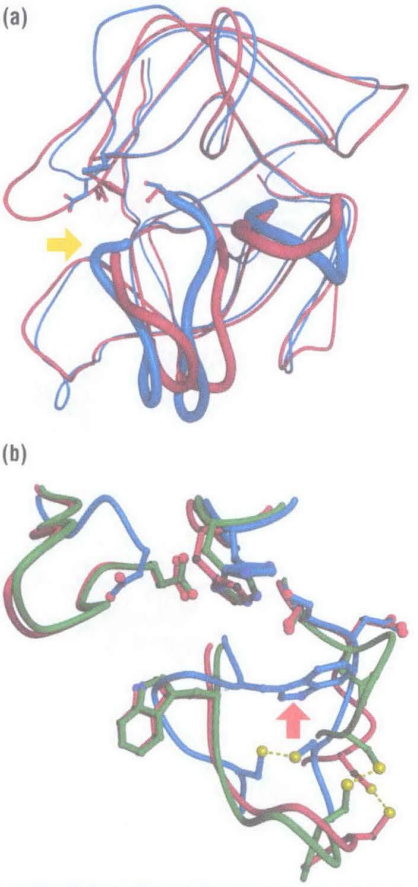


Figure 4-24 Structural changes in closely related proteins (a) The structures of plasminogen (blue) and chymotrypsinogen (red) are very similar, as befits their high sequence identity. Yet the small differences in the positions of loops have important functional consequences, as seen in (b). **(b)** 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

©2004 New Science Press Ltd

4-7 Structure from Sequence: Profile-Based Threading and "Rosetta"

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

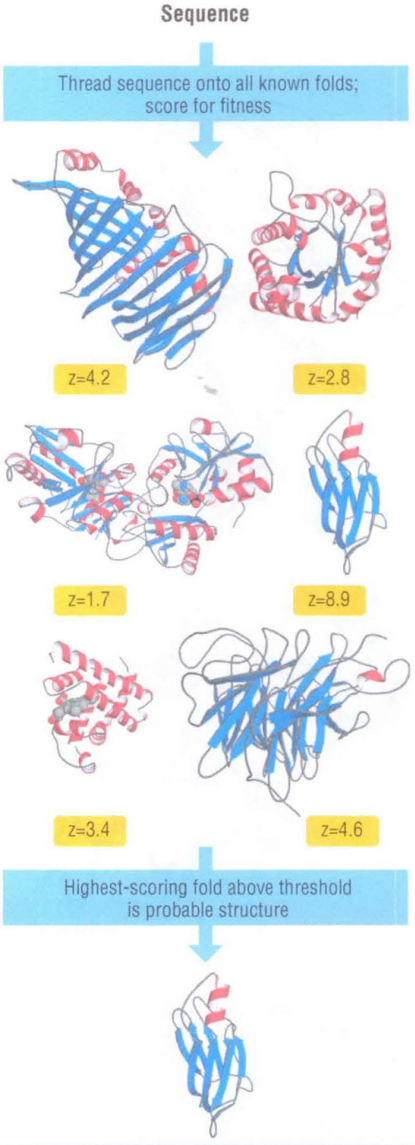


Figure 4-25 The method of profile-based threading A sequence of unknown structure is forced to adopt all known protein domain folds, and scored for its suitability for each fold. The z-value relates the score for the query sequence to the average score for a set of random sequences with the same amino-acid 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.

©2004 New Science Press Ltd

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.

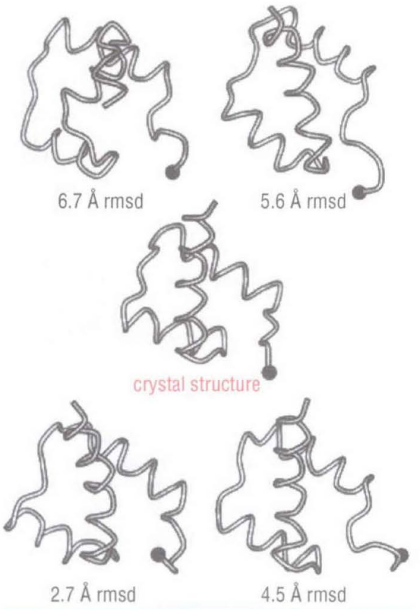


Figure 4-26 Some decoy structures produced by the Rosetta method The structure at the center is the target, the experimentally determined structure of a homeodomain. The other structures are generated by the Monte Carlo approach in Rosetta, using only the sequence of the protein. Although some of the 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.)

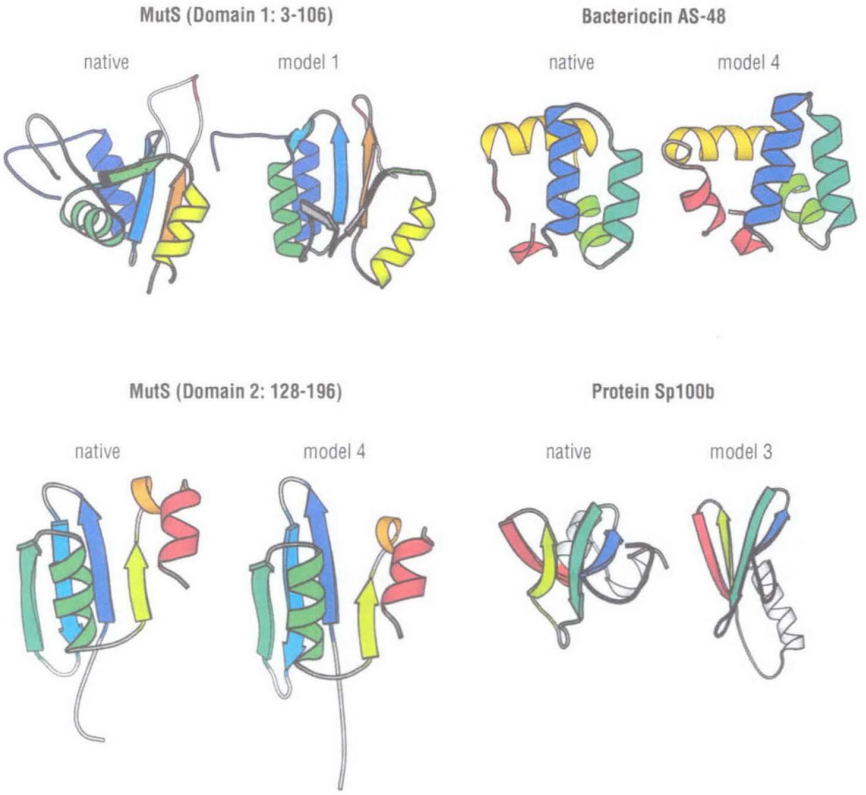


Figure 4-27 Examples of the best-center cluster found by Rosetta for a number of different test proteins The level of agreement with the known native structure varies, but in many cases the overall fold is predicted well enough to be recognizable. Note, however, that the relative positions of the secondary 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

©2004 New Science Press Ltd