pression subtractive hybridization: a method for generating differentially regulated or tissue-specific cDNA probes and libraries, *Proc. Natl. Acad. Sci. USA* 93(12), 6025–6030.

Ding, C., Cantor, C.R. **2004**, Quantitative analysis of nucleic acids: the last few years of progress, *J. Biochem. Mol. Biol.* 37(1), 1–10.

Dykxhoorn, D.M., Novina, C.D., Sharp, P.A. **2003**, Killing the messenger: short RNAs that silence gene expression, *Nat. Rev. Mol. Cell Biol.* 4(6), 457–467.

Elahi, E., Kumm, J., Ronaghi, M. **2004**, Global genetic analysis, *J. Biochem. Mol. Biol.* 37(1), 11–27.

Goldsmith, Z.G., Dhanasekaran, N. **2004**, The microrevolution: applications and impacts of microarray technology on molecular biology and medicine, *Int. J. Mol. Med.* 13(4), 483–495.

Green, C.D., Simons, J.F., Taillon, B.E., Lewin, D.A. **2001**, Open systems: panoramic views of gene expression, *J. Immunol. Methods* 250(1–2), 67–79.

Horak, C.E., Snyder, M. **2002**, Global analysis of gene expression in yeast, *Funct. Integr. Genomics* 2(4–5), 171–180.

Hubank, M., Schatz, D.G. **1994**, Identifying differences in mRNA expression by representational difference analysis of cDNA, *Nucleic Acids Res.* 22(25), 5640–5648.

Kurian, K.M., Watson, C.J., Wyllie, A.H. **1999**, DNA chip technology, *J. Pathol.* 187(3), 267–271.

Lichter, P. **1997**, Multicolor FISHing: what's the catch? *Trends Genet.* 13, 475–479.

Lorkowski, S., Cullen, P. **2002**, *Analyzing Gene Expression: A Handbook of Methods. Possibilities and Pitfalls*, vol. 2, Wiley-VCH, Weinheim.

Matz, M.V., Lukyanov, S.A. **1998**, Different strategies of differential display: areas of application, *Nucleic Acids Res.* 26(24), 5537–5543.

McClelland, M., Mathieu-Daude, F., Welsh, J. **1995**, RNA fingerprinting and differential display using arbitrarily primed PCR, *Trends Genet.* 11(6), 242–246.

McGall, G.H., Christians, F.C. **2002**, High-density genechip oligonucleotide probe arrays, *Adv. Biochem. Eng. Biotechnol.* 77, 21–42.

Pevsner, J. **2003**, *Bioinformatics and Functional Genomics*, Wiley, New York.

Pollok, B.A., Heim, R. **1999**, Using GFP in FRET-based applications, *Trends Cell Biol.* 9, 57–60.

Rapley, R., Harbron, S. **2004**, *Molecular Analysis and Genome Discovery*, Wiley, New York.

Reece, R.J. **2004**, *Analysis of Genes and Genomes*, Wiley, New York.

Reits, E.A., Neefjes, J.J. **2001**, From fixed to FRAP: measuring protein mobility and activity in living cells, *Nat. Cell Biol.* 3, E145–147.

Sensen, C.W. **2005**, *Handbook of Genome Research: Genomics, Proteomics, Metabolomics, Bioinformatics, Ethical and Legal Issues*, vol. 2, Wiley, New York.

Shi, H., Maier, S., Nimmrich, I., Yan, P.S., Caldwell, C.W., Huang, T.H. **2003**, Oligonucleotide-based microarray for DNA methylation analysis: principles and applications, *J. Cell Biochem.* 88(1), 138–143.

Simon, R., Mirlacher, M., Sauter, G. **2004**, Tissue microarrays, *Biotechniques* 36(1), 98–105.

Southern, E.M. **2000**, Blotting at 25, *Trends Biochem. Sci.* 25, 585–588.

Sung, Y.H., Song, J., Lee, H.W. **2004**, Functional genomics approach using mice, *J. Biochem. Mol. Biol.* 37(1), 122–132.

Taylor, G.R., Day, I.N., Human Genome Organization (HUGO) **2005**, *Guide to Mutation Detection*, Wiley, New York.

Tucker, C.L. **2002**, High-throughput cell-based assays in yeast, *Drug Discov. Today* 7, S125–130.

Velculescu, V.E., Vogelstein, B., Kinzler, K.W. **2000**, Analyzing uncharted transcriptomes with SAGE, *Trends Genet.* 16, 423–425.

Westermeier, R., Naven, T. **2002**, *Proteomics in Practice: A Laboratory Manual of Proteome Analysis*, Wiley-VCH, Weinheim.

23
Protein–Protein and Protein–DNA Interaction

Learning Objectives ***Virtually all biological processes rely heavily on protein–protein interactions. Most of these interactions are mediated by protein domains, of which the human genome alone codes for about 750. While all domains are assumed to have defined interaction partners, these are known for only a small number of domains. Protein interactions are often organized in stable complexes. On the average, a single interaction involves 20 amino acids; this corresponds to a surface area of about 800 Å. The most important forces involved are hydrophobic interactions, hydrogen bonds, and ionic bonds. The mass action law can describe protein interactions as bimolecular reactions; an average energy in the magnitude of 10 kcal is needed to break 1 mol of dimers. Essential techniques for the examination of interactions are protein purification and characterization of complexes with mass spectroscopy, but the two-hybrid method, FRET and in vitro binding assays are equally important. Regulatory mechanisms use the expression of proteins, but also their localization, stability, and possible covalent modifications (and noncovalent modifications like bound ligands). Protein interactions can sometimes be predicted theoretically by, e.g., the Rosetta Stone method, molecular docking, or phylogenetic profiles.***

Protein–DNA interactions play essential roles in all aspects of gene regulation, specific recognition of gene sequences being of central importance. So far, it is impossible to predict the sequence specificity of a given DNA-binding protein. However, protein–DNA interactions can be examined with a number of methods stemming from biophysics and molecular biology. Just like with protein–protein interactions, x-ray structure analysis is the only method capable of exploring interactions in atomic detail. Nature uses only a limited number of domains or motifs for DNA binding, and many of these domains are understood down to their three-dimensional structure. Based on their structure and function, DNA-binding proteins can be classified into groups and families.

An Introduction to Molecular Biotechnology. Edited by M. Wink

Protein interactions and protein–DNA interactions are objects of manifold efforts in medicine and biotechnology, e.g., when developing cancer therapeutics, which block protein interactions, or ligands that prevent certain protein–DNA interactions.

23.1 Protein–Protein Interactions

Almost all cellular processes feature protein–protein interactions in prominent roles. For instance, all structuring elements like actin filaments or microtubules consist of protein complexes held together by protein interactions. Many enzymes are also composed of subunits that develop their full activity only when compounded. RNA polymerases are an arbitrary example for this, taken from hundreds of protein complexes within a cell. Their subunits need to enter numerous protein interactions among themselves, but also with DNA and RNA, their enzymatic substrate and product. Proteins further interact with low molecular weight substances like sugars, fats, or salts. However, due to limited space these aspects will not be covered in this chapter. They should still be kept in mind since they are of considerable importance to cell metabolism.

23.1.1 Classification and Specificity: Protein Domains

Because of their great variety, it is almost impossible to classify protein interactions in a meaningful way. Arbitrarily, interactions can be divided into **strong (stable)** and **weak (transient) interactions**, but there is no clear-cut border between the two types. Many protein complexes are assembled quite stably, as their integrity is essential to their functions; ribosomes, for example, are largely stable as protein–RNA complexes. On the other hand, even form-giving structures like actin filaments are constantly being assembled and disassembled.

Although protein interactions need to be extremely specific (e.g., the binding of a peptide hormone like insulin to its receptor), many weak interactions appear to be relatively unspecific and thus without immediate meaning. They are tolerated without consequences, however, as long as they do not set the organism at a disadvantage. Unspecific interactions should not be confused with chance collisions caused by **Brownian motion**, as the latter do not create cohesion. Weak interactions might have played an important part in evolution, as they can be enhanced through mutation and selection and thus be made useful.

Biologically, it makes more sense to classify interactions by protein domains involved. **Domains** are the structural and functional units of protein interaction. They fold independently of other protein areas and tend to be globular with a length between 40 and 150 amino acids (Fig. 23.1). Many domains can be attributed with certain interaction qualities; for example, **SH3-domains** bind proline-rich sequences, **SH2-domains** bind peptide sequences that contain phosphotyrosine, and so on. The two mentioned domains have been named after their homology to the **oncoprotein Src**, which causes sarcomas. All Src-related proteins have such Src homology (SH) domains; the SH1 domain represents a **kinase domain.** However, there are still numerous examples from the about 750 protein domains in the human proteome, whose binding qualities are barely, if at all, known. Even when a domain's principle qualities have been discovered (e.g., binding proline-rich sequences), it is still impossible to exactly predict its interaction partners as numerous proline-rich proteins are encoded by most genomes. Predicting such interactions remains an important challenge to structural biologists and bioinformaticists.

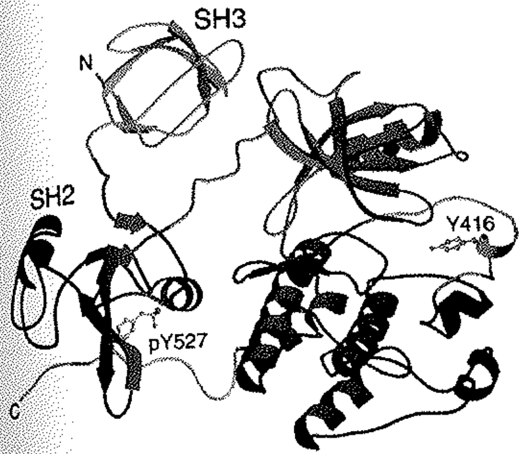

Fig. 23.1 Protein domains of the Src oncoprotein. The Src protein has three major domains: SH3, SH2 and SH1, the latter of which is the kinase domain. All three enter several well-defined interactions. The smaller domains do not simply interact with other proteins, but also with sequences within Src: SH3 binds a proline-rich sequence between SH2 and the kinase domain; the SH2 domain binds to a phosphorylated tyrosine at position 527, close to the C-terminal (pY527). This figure also appears with the color plates.

23.1.2 Protein Networks and Complexes

Eukaryotic cells are estimated to contain a hundreds of different discrete **protein complexes.** Many of these complexes contain dozens if not hundreds of different proteins (ribosomes, splicosomes, sarcomer elements in muscles, RNA polymerases; see Fig. 23.2). But even well-defined complexes interact with other, transiently associated proteins, e.g., translation factors and ribosomes. Proteins within a cell can thus be thought of as nodes within a giant protein network, which links most of a cell's proteins (see Fig. 23.3). Indeed, estimates say that each protein interacts with three other proteins on average. Although systematic protein interaction analyses have been performed for only a few organisms (like

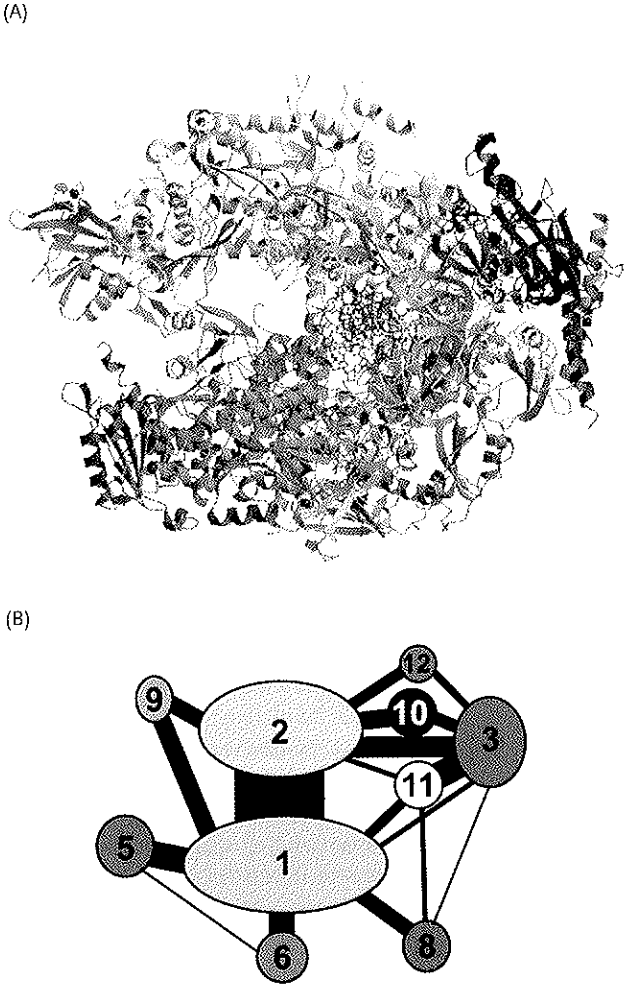

Fig. 23.2 RNA polymerase II: a multimeric protein complex. (A) Diagram of yeast RNA polymerase II. (B) Schematic diagram showing the interactions between the 10 subunits. The thickness of the connecting lines corresponds to the size of the contact area between the individual subunits. The colors correspond to those in (A) [1]. This figure also appears with the color plates.

several viruses, few bacteria, yeast, and a few other model organisms), it is estimated that the 30000 or more proteins of the human body interact with each other in more than 100000 ways. So far, only several thousand of these interactions have been experimentally proven and catalogued in data bases (Table 23.1). However, modern high-throughput methods will quickly increase our knowledge of such interactions.

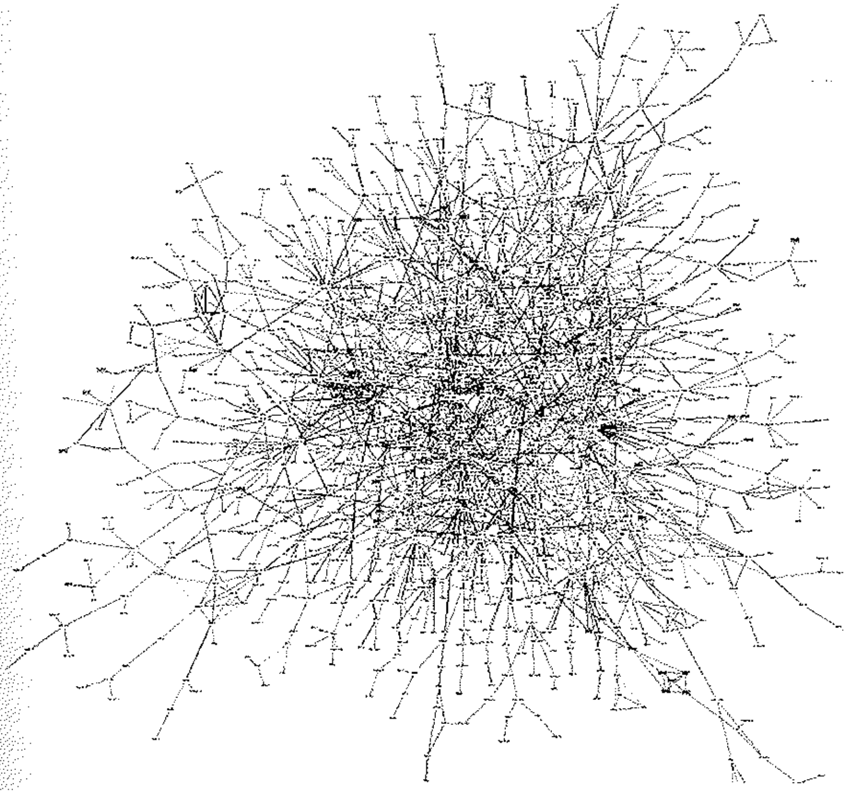

Fig. 23.3 Protein interaction network of a yeast cell. This map was reconstructed from published interaction data and contains 1548 proteins linked by 2358 interactions. The proteins are colored according to their biological function: proteins involved in membrane fusion are blue, chromatin proteins are grey, structure proteins are green, lipid metabolism proteins yellow and mitotic proteins are red [2]. This figure also appears with the color plates.

Table 23.1 Databases.

Database	Web Address
Database of Interacting Proteins (DIP)	http://dip.doe-mbi.ucla.edu/
BIND	http://www.bind.ca/
AMAZE	http://www.amaze.ulb.ac.be/
MINT	http://cbm.bio.uniroma2.it/m int/
PDB (3D-structures)	http://www.rcsb.org/
NDB (Nucleic acids and proteins)	http://ndbserver.rutgers.edu/
Protein Domains	http://smart.embl-heidelberg.de/

23.1.3 Structural Properties of Interacting Proteins

Several hundred protein complexes have already been examined by x-ray structure analysis and other methods; their structural data is available from the protein data bank (PDB, http://www.rcsb.org/). The following statements regarding the geometry and energetics of protein interaction have been derived from the analysis of several dozens to about 100 crystallized protein pairs.

The **contact area** between two proteins is almost always greater than 1100 Å^2, with each interaction partner contributing at least 550 Å^2 to the entire surface. On the average, every partner loses about 800 Å^2 of **solvent contact surface** area per interaction, which corresponds to about 20 amino acids per partner. In other words: Every amino acid residue involved in an interaction covers about 40 Å^2.

On an average, dimers contribute 12% of their surface to an interaction, trimers contribute 17.4%, and tetramers contribute 21%. There are considerable differences between individual complexes, however; the entire contact surface ranges from 6% for dimers of inorganic **pyrophosphatase** up to 29% for **Trp repressor** homodimers. This also concludes that protein surfaces almost always allow for interactions with several proteins at a time.

Eighty-three to eighty-four percent of contact surfaces are more or less flat. With a few exceptions, contact surfaces are almost round areas on the surface of stable or transient complexes. Contact surfaces in stable interactions tend to be larger, less plane, more strongly segmented (on a sequential level) and more densely packed than contact surfaces of unstable interactions.

Concerning **secondary structure**, one investigation showed that loop interactions on the average constitute about 40% of the contact area. Another study of 28 homodimers showed 53% of contact surfaces to be α-helical, 22% to be β-sheets, 12% to be $\alpha\beta$, and the remaining 11% to be coils.

Complementarity can be defined as fitting surface shape. Contact areas in homodimers, enzyme-inhibitor complexes, and stable heterodimers tend to be the most complementary. Antigen-antibody complexes and unstable heterodimers appear to possess the weakest complementarity.

Concerning **amino acid composition**, contact surfaces between proteins tend to be more hydrophobic than their outsides, but less hydrophobic than the protein interior. One study showed 47% of interacting amino acid residues to be hydrophobic, 31% to be polar and 22% to be charged. Stable complexes have contact surfaces with hydrophobic residues while unstable complexes tend to prefer polar residues. Mutagenesis experiments have shown that often more than half of a contact surface's amino acids can be changed to alanine without significantly altering the affinity constant (K_d). This concludes that the **functional epitope** is only a fraction of the structural epitope.

23.1.4 Which Forces Mediate Protein–Protein Interactions?

Interestingly, the average contact area of two interacting proteins is hardly any more polar or hydrophobic than the rest of the protein that is in contact with the solvent. **Transient complexes**, however, tend to have more hydrophilic contact areas, which makes sense, as both components need to exist separately in the cell's aqueous environment. Water is usually excluded from the contact site.

Some authors have proposed that **hydrophobic interactions** provide the energetic basis for the interaction while **hydrogen and salt bridges** ensure specificity.

Although van der **Waals forces** affect all neighboring atoms, these forces are by no means stronger between two proteins than between a protein and the solvent. Still, they contribute to protein interaction energetically, because they are more frequent on the densely packed contact sites than on the solvent interface.

Hydrogen bonds between proteins are often energetically favored over those with water. Stable protein complexes often feature fewer hydrogen bonds than transient complexes. The number of hydrogen bonds roughly equals 1 per 170 Å^2 of surface area. The average interaction area (ca. 1600 Å^2) thus contains about 900 Å^2 of unpolar surface, 700 Å^2 of polar surface, and about 10 (±5) hydrogen bonds. A random sample of relatively stable dimers featured 0.9 to 1.4 hydrogen bonds per 100 Å^2 on the average (with entire contact surfaces usually >1000 Å^2). However, the spread from zero (e.g., uteroglobin) to up to 46 (variant surface glycoprotein) was considerable. The amino acid side chains are involved in about 76 to 78% of hydrogen bonds.

Only 56% of homodimers even possess salt bridges; those that do can have up to five.

23.1.4.1 Thermodynamics

Protein interactions can be described as simple chemical reactions of the following form:

$$A + B \underset{k_d}{\overset{k_a}{\rightleftharpoons}} AB$$

A and B represent two proteins that form the complex AB. Even multi protein complexes are assumed to form through successive binding of subunits.

Protein–protein interactions can be very weak and short-lived as well as strong and permanent. The former are referred to as **transient**, the latter as **stable**, although all numbers of intermediate grades exist. For example, an enzyme can bind its substrate, phosphorylate it, and dissociate afterwards in less than a microsecond. On the other end of the scale, some protein complexes like the collagen triple helix stably persist in bones or other tissues for weeks or even years without dissociating.

The interaction between two proteins can be described quantitatively with the **mass action law**:

$$\frac{[A][B]}{[AB]} = \frac{1}{K_a} = K_d = \frac{k_d}{k_a} \tag{23.2}$$

with k_a being the reaction constant of second degree for the biomolecular association, k_d equalling the reaction constant of first degree for the unimolecular dissociation, $K_d = k_d/k_a$ being the reaction constant of the dissociation (K_a for association).

K_d depends on the concentrations of A, B, and AB at thermodynamic equilibrium; K_d has the dimension of a concentration (mol L^{-1} or M). K_a and K_d values for protein–protein interactions vary extremely and range over 12 orders of magnitudes from 10^{-4} to 10^{-16} M.

Interactions with K_d values in the mM range are considered weak, values in the nM range and below quite strong. The interaction between **trypsin** and **pancreas trypsin inhibitor**, for example, has a dissociation constant in the range of 10^{-14} M, the binding is thus very strong and stable. Biological interaction strength can also depend on other factors, however, for example **cooperativity**. Several weak interactions between the subunits of a complex can form a very stable complex.

23.1.4.2 Energetics

Values of K_d between 10^{-4} and 10^{-14} M correspond to free enthalpies ΔG_d of 6 to 19 kcal mol^{-1}, i.e., 19 kcal are required to dissociate one mol of the complex. Dehydration of the nonpolar groups on the contact surface is definitely decisive for stable association. Real K_d values for protein–protein interaction can be looked up in special databases, e.g., the database of interacting proteins, DIP (see Table 23.1).

Interactions between single amino acids can contribute up to 6 kcal mol^{-1} to a single protein interaction. The greatest energy gain, however, is provided by salt bridges and hydrogen bonds between charged amino acids. The strength of neutral hydrogen bonds lies in the range of 0–3 kcal mol^{-1}. This amount is significantly below a normal hydrogen bond's energy and means that the interaction between two amino acid residues within a complex is hardly stronger than interaction with the surrounding water of a soluble protein. In complexes of known three-dimensional structure, the peptide bonds form at least half of the hydrogen bonds between interacting proteins. Bonds between side chains and primary chain are especially common, although bonds between both primary chains are also observed at times.

Estimates say that the nonpolar contact areas of hydrophobic interactions provide an energy gain of about 25 to 70 calories per Å^2. Sometimes protein–protein interactions can be so strong (i.e., with a K_a value lower than 10^{-16} M^{-1}) that the components can only be separated by denaturing them.

23.1.5 Methods to Examine Protein–Protein Interactions

Several methods of examining protein–protein interactions are described in Chapters 8 and 28 (**recombinant antibodies** and phage display). One of the dominant methods is the purification of proteins that have been fused to a foreign protein such as **glutathione S-transferase (GST)**. The fusion proteins and associated proteins can then be isolated on a glutathione-linked matrix and identified with mass spectrometric methods (see Chapter 8). Ideally, the structures of the interacting proteins can be determined both individually and in the complex. The methods used are **nuclear magnetic resonance spectroscopy (NMR**, for smaller proteins) or – especially for larger complexes – **x-ray crystallography**. Another possible technique is selecting interacting proteins with phages (**phage display**, see Chapter 28).

In vivo **methods** involve expressing genes in such a manner that their interaction activates a so-called **reporter gene** (e.g., the **two-hybrid system**; see Fig. 23.4). Today, screenings are performed systematically by using robotic aid to simply test all possible protein pairs. Alternatively, light signals can also serve as reporter systems (e.g., **FRET**, **fluorescence resonance energy transfer**). This method observes two fluorescent proteins in regards to their spatial proximity. One of the proteins to be examined is fused with the **cyan fluorescent protein (CFP)**, a possible partner protein with **yellow fluorescent protein (YFP)**. When these proteins interact or otherwise come into close proximity of each other (at least 100 Å, 30 Å at best), the colocalization can be detected by irradiation with

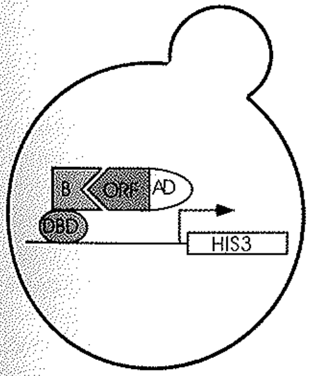

Fig. 23.4 Working principle of the two-hybrid system. The two-hybrid system is based on the expression of two fusion proteins within a cell. One of the proteins contains a DNA-binding domain (DBD), which can bind to the promoter of a reporter gene (here: His3), and a second protein B, the bait. The second fusion protein consists of a transcription-activation domain (AD) and a second protein, ORF (for open reading frame). If the proteins inserted in the B and ORF positions interact, a transcription factor is formed and the reporter gene is activated. In this case, that means that the cell can grow on histidine-free medium. A yeast colony growing on such medium thus indicates an interaction of the two inserted proteins.

blue light with a wavelength of 434 nm. This wavelength is absorbed by CFP, which immediately transfers the absorbed energy to YFP; YFP then emits its characteristic yellow light with a wavelength of 527 nm. A yellow signal in the fluorescence microscope thus indicates protein interaction (or close proximity).

Protein interactions can also be measured quantitatively. **Dissociation constants** are determined on a micromolar scale through equilibrium centrifugation or **microcalorimetry.** More accurate measurement on a nanomolar scale requires radioactive markers or antibody reactions. These methods are not frequently employed, however, and will therefore not be covered in this book.

23.1.6 Regulation of Protein–Protein Interactions

Protein interactions are subjected to strong regulation; if this regulation is disturbed or gets out of control, diseases like cancer may be the result.

The most important regulator of protein–protein interactions is **expression control**, because, naturally, proteins can only interact if they are expressed in the same place at the same time. The central control mechanisms are those for transcription and translation (see Chapter 4). For example, most **growth factors,** like some fibroblast growth factors (FGF), are expressed only in certain tissues, like limbs, brain or kidneys. Some FGFs are given off into the blood stream, from where they can reach and bind to receptors that are expressed only in certain tissues. FGFs are also strongly regulated in a temporal dimension; for example, FGF4 and FGF8 are only expressed in the embryo, while the other FGFs are primarily found in adult animals. The same is true for the respective receptors. **Protein localization** within a cell is also of great importance. Some transcription factors like **NF-κB** (composed of two subunits: relA and p50) are normally found as inactive protein complexes in the cytoplasm (Fig. 23.5). NF-κB is bound to its inhibitor IκB that dissociates after phosphorylation and is then degraded. The liberated NF-κB protein then enters the nucleus where it regulates the activity of target proteins. **Protein stability** is similarly important as expression, as the final concentration is determined by the equilibrium between synthesis and degradation. Numerous proteins are regulated at this aspect. **Cyclins,** for example, are specifically degraded during certain phases of the cell cycle and thus can no longer interact with their partners, the **cycline-dependent kinases (CDK).**

Covalent modifications are substantial regulators for protein–protein interactions. An important example beyond the above mentioned phosphorylation is acetylation of histone proteins, that allows for the association of **bromo domains** (see Chapter 4). These protein domains can bind only to acetylated histones.

Ligands are another important form of regulation. GTP, as a prominent example, binds to the α-subunit of **trimeric G-proteins** and causes the dissociation of the βγ-subunit (see Chapter 2). The unbound subunits bind other proteins in turn and regulate their activity. Exchanging GTP for GDP triggers the reassociation of the subunits.

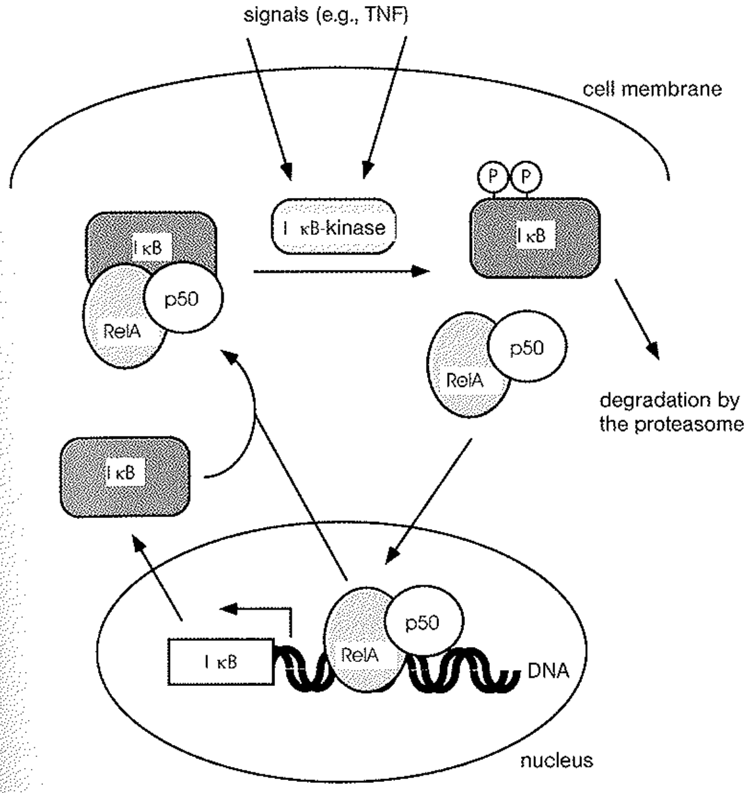

Fig. 23.5 The NF-κB signalling pathway as an example for protein–protein and protein–DNA interactions. Several signals influence the activity of the IκB-kinase complex (IKK), e.g., coming from the TNF receptor (tumor necrosis factor). When induced, IKK phosphorylates IκB. This phosphorylation triggers the recognition of IκB by ubiquitination enzymes, which modify it, so that it is recognized and degraded by the proteasome. The removal of IκB exposes a previously covered nuclear localization signal on the NF-κB complex, which can now migrate into the nucleus and bind to specific DNA sequences there. It functions as a transcription factor that activates several target genes, among them the gene for IκB. The now newly expressed IκB in turn binds the NF-κB complex at the promoter and deactivates the gene once more. Many other target genes and interaction partners of NF-κB exist beyond IκB. This example illustrates how complex regulatory networks can be and how many layers of regulation they are composed of (here: transcription, localization, modification by phosphorylation, and ubiquitination). Also see Fig. 23.15.

23.1.7 Theoretical Prediction of Protein–Protein Interactions

Even when the three-dimensional structures of two interacting proteins are understood, predicting their contact sites is by no means trivial. Interaction sites are often located at sites with hydrophobic properties. In one study of protein interactions, 25 out of 29 interactions could be correctly predicted based on this feature. Other parameters suitable for prediction are solvation potential, the residue interface propensity of a given amino acid, planarity, protrusion, and the available surface. In a sample of 28 homodimers, the contact site was often planar, well exposed and thus accessible, and the area with the highest residue interface propensity. Still, one of the prediction algorithms employed (PATCH), which utilizes several of the parameters mentioned above, predicted the contact sites correctly in only 66% of the examined structures.

Predicting Interacting Proteins by their Genome Sequence Several attempts have been made to predict protein–protein interactions *de novo*. One of these approaches is the **Rosetta Stone method** (Fig. 23.6): It uses the observation that parts of some proteins that form a single entity in one organism are distributed over two proteins in another. From this, it was concluded that the respective protein halves must interact, since they interact similarly in the fusion protein. These fusion proteins are hence called Rosetta Stone proteins after the famous Rosetta Stone found in Egypt in 1799, engraved with the same text in Greek, Demotic (an early Egyptian common writing), and Hieroglyphs. The stone marked a breakthrough in the translation of hieroglyphs. One example of a Rosetta Stone protein is **human succinyl CoA transferase**, that is found split in two halves in *E. coli*, namely the α- and β-subunits of acetyl CoA transferase (Fig. 23.6).

Phylogenetic Profiles Some gene combinations are retained stably over the course of evolution, which means that these genes apparently cannot exist in isolation. It is concluded that these genes code for components of metabolic pathways or protein complexes that require all respective subunits or otherwise lose their function. Although such phylogenetic profiles do not necessarily allow for conclusions regarding physical interactions, a functional cohesion of the respective gene groups has been found in many cases. Good examples are **ribosomal proteins** or the **yeast proteins Hog1 and Fus3**, two kinases in the yeast MAP-kinase signalling pathway.

23.1.8
Biotechnological and Medical Applications of Protein–Protein Interactions

Substantial fields of biotechnological research are concerned with the production of **interacting proteins** such as antibodies or peptide hormones. One therapeutically used antibody for instance, is **Herceptin**, which binds to the cancer protein HER2 that is overexpressed in 25 to 30% of all cases of breast cancer. **Erythropoietin** is an example of a peptide hormone that stimulates the formation and maturation of red blood cells (erythrocytes) in the bone marrow. It has been produced biotechnologically for several years now and has become infamous for its involvement in doping cases in professional sports. With exact knowledge of protein–protein interactions, substances that specifically block those interactions can be identified. For example, it is desirable to block the binding of **HIV** to its **target receptors** CD4, CCR5, and CXCR4. Meanwhile, there is also a number of substances available that develop their effect *within* cells by blocking specific protein interaction. The **immunosuppressant FK506**, for instance, binds the **FK506-binding protein (FKBP)**. The resulting complex in turn blocks the activity of the phosphatase **calcineurin** through direct interaction, which then triggers the actual immunosuppressant effect.

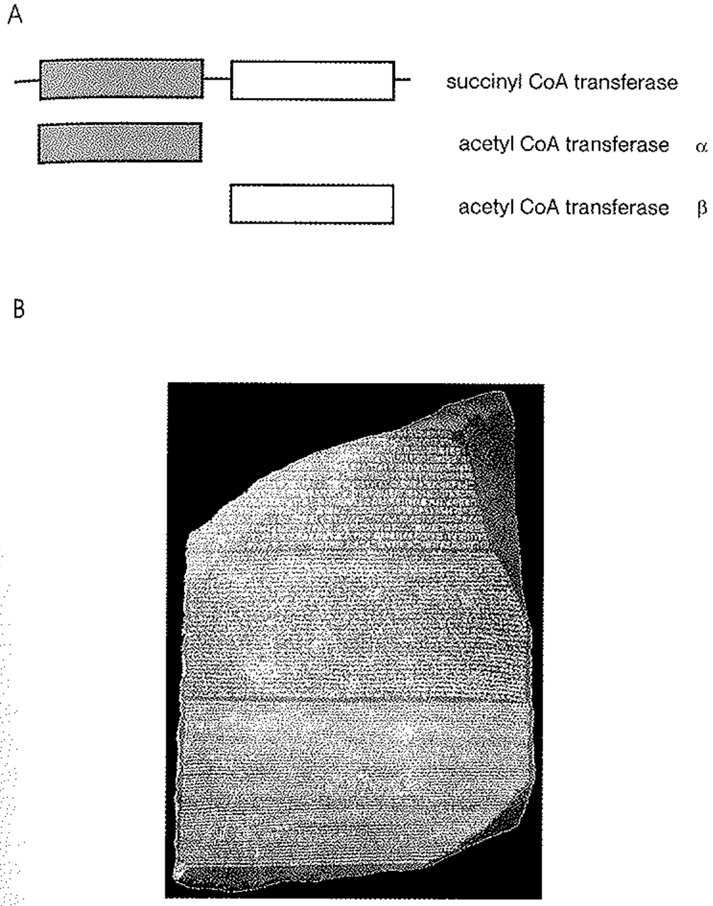

Fig. 23.6 The Rosetta Stone method. (A) Some proteins, like human succinyl CoA transferase, are distributed over two proteins in other organisms. In *E. coli*, the protein's function is fulfilled by the α- and β-subunits of the enzyme acetate CoA transferase. (B) The historical Rosetta Stone from Egypt. Further details in the text.

Occasionally protein–protein interactions prove disadvantageous, like with insulin, which tends to form dimers or hexamers that show less activity than monomers. This tendency toward oligomerization can be suppressed via genetic modification, and thus **insulin** of higher activity can be produced.

23.2 Protein–DNA Interactions

Protein–DNA interactions play a major role in all fields of genetics from regulation and transcription of individual genes to repair of damaged sequences, even to the stabilization of DNA in chromatin and the replication of entire genomes. Presently, two to three percent of prokaryotic and six to seven percent of eukaryotic genes are estimated to code for **DNA-binding proteins**. Additionally, these proteins do not merely bind DNA, but also interact with other proteins and sometimes, as is shown in the example of RNA polymerase, only display their full activity when organized in multimeric complexes.

23.2.1 Sequence-specific DNA-binding

Protein recognition of specific sequences on the DNA double helix is of critical importance to fundamental genetic processes like gene regulation and transcription (see Chapter 4). The specific binding occurs on an atomic level via an interaction of certain side chains of the protein with nucleotides of the DNA. In spite of numerous attempts to formulate them, no simple or general rules to explain or predict sequence specificity exist to date. Instead, the same basic interactions known from **protein–protein** or **protein–ligand** interactions are also found in **protein–DNA** complexes. The sequence-specific binding results from an individual combination of the different possible interactions. The first aspect of this specificity are the different hydrogen bond patterns of the base pairs **adenine–thymine** (A–T) and **guanosine–cytosine** (G–C), that are presented in the minor or major groove of the DNA double helix (Fig. 23.7).

Statistic analyses of known protein–DNA complex x-ray structures could show that the positively charged side chains of **arginine and lysine** preferably form hydrogen bonds with guanine; **asparagine and glutamine** mainly form hydrogen bonds with adenine. Not quite as specific, the shorter side chains of serine and threonine have been shown to bind to the sugar-phosphate chain and are probably concerned with overall stability rather than specificity. Aside from hydrogen bonds, hydrophobic interactions are also important: Although van der Waals contacts tend to be less specific than hydrogen bonds, certain preferences could still be observed. **Arginine** shows a preference for guanine while **threonine** prefers methyl–methyl van der Waals interaction with thymine. **Phenylalanine, proline** and **histidine** form hydrophobic stacks with the planar bases, especially in those structures in which the DNA is sufficiently deformed (e.g., in TATA box-binding proteins, see below). A third possibility for interaction is indirect contact mediated by water molecules. The importance of these interactions has newly become apparent in recent years with the growing number of high-resolution x-ray structures (with resolutions better than 2 Å). Unfortunately, the resulting binding patterns are still too complex to formulate general rules.

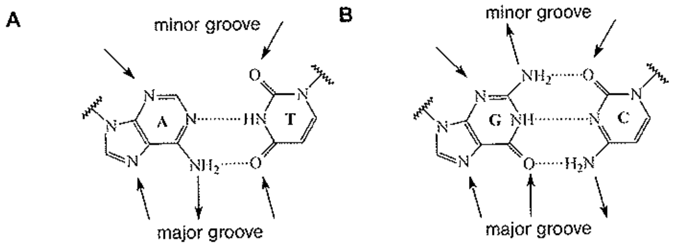

Fig. 23.7 Watson-Crick hydrogen bonds of the base pairs A–T and G–C. The arrows indicate potential hydrogen donors or acceptors.

23.2.2 Thermodynamic Considerations Regarding Protein–DNA Complexes

Although only a small number of protein–DNA complexes has been characterized thermodynamically (much less than are structurally known), some important and basic considerations can be made and a few conclusions be drawn. **Sequence-specific proteins** generally have a high DNA affinity with association constants of $K_a > 10^{-7}$ M. In contrast to this, the respective values for **unspecific binding** are lower by up to three orders of magnitude, to ensure sufficient discrimination. The upper limit for K_a is estimated to be in the range of 10^{-12} M, because at higher values complex formation would not be reversible under physiological conditions and would react too sensitively to minor concentration fluctuations within the cell. Furthermore, it could be shown that the association constants (and thus the free binding enthalpies ΔG^0) of several protein–DNA complexes tended to be quite similar, while the individual contributions of ΔH^0 and $T\Delta S^0$ varied greatly. Stabilizing enthalpy contributions result from the formation of hydrophilic and hydrophobic interactions, whereas the loss of hydrogen bonds to solvent molecules has a destabilizing effect. Still, the liberation of water molecules constitutes the most important contribution to $T\Delta S^0$; the actual complex formation lowers entropy. Further destabilizing enthalpy contributions result when complex formation forces one of the partners into a disadvantageous conformation. In many protein–DNA complexes the DNA double helix is bent from its canonical **B-conformation** in this manner.

23.2.3 Methods to Study Protein–DNA Interactions

The methods used to examine protein–DNA interactions are similar to those that have already been described in the first part of this chapter (Fig. 23.8). The special methods used in addition to these are summarized in Fig. 23.8 and Table 23.2.

The only method capable of exploring protein–DNA complexes in atomic detail is x-ray crystallography.

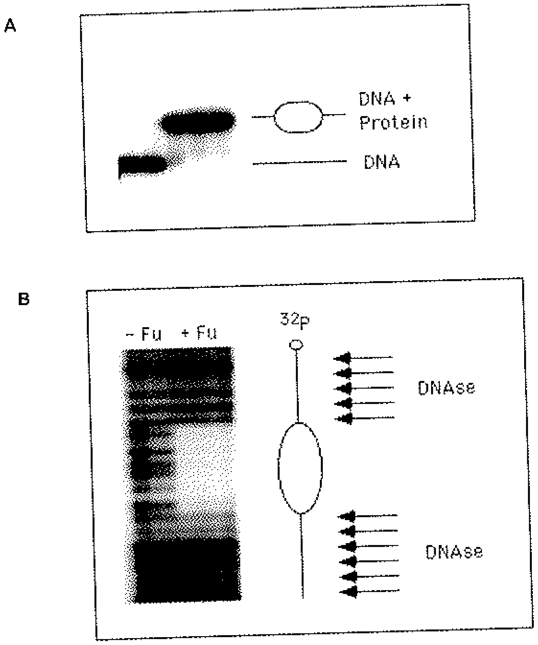

Fig. 23.8 DNA band shift assay and DNAse I protection assay of the ferric uptake regulator (Fur). (A) DNA band shifts are based on the fact that free DNA and free proteins migrate through a gel faster than a DNA–protein complex. (B) DNAse I footprinting involves digesting DNA incompletely (with or without bound protein), resulting in fragments of different length. DNA-binding proteins protect DNA from digestion. If the (radioactively marked) fragments are loaded onto a gel, a gap between the bands corresponding to the binding site can be seen (with kind permission from M. L. Vasil).

Table 23.2 Important methods to examine protein–DNA interactions.

Method	Description
DNAse I footprinting	A (e.g., radioactively) marked DNA fragment is incubated with a protein and digested with DNAse I. The following gel electrophoresis provides information on the DNA area protected by the protein (Fig. 23.8 B).
Hydroxyl radical footprinting	In this case, unprotected DNA areas are damaged by hydroxyl radicals that are produced *in situ* via a reaction of H_2O_2 and iron. Since OH radicals are much smaller than, e.g., DNAse I, the interaction can be characterized in greater resolution.
Band shift assays	With this method, radioactively or chemically marked DNA fragments are incubated with the protein and analyzed with nondenaturing polyacrylamide gel electrophoresis (PAGE). The difference in the migration speed of the individual components and the complex can help determine the complex properties (Fig. 23.8 A).
Fluorescence spectroscopy	This method measures changes in the intrinsic fluorescence of DNA and determines stoichiometry and equilibrium constants.
Isothermal titration calorimetry (ITC)	This method involves pipetting small amounts of DNA to the protein solution at constant temperature; ΔH^0 can then simply be measured.
Electron microscopy	Electron microscopy can be used to roughly determine form and shape. It is mostly used for large multiprotein complexes.

Structural Classification of Protein–DNA Complexes **DNA-binding proteins** can be divided into eight groups based on their structure and function; each of these groups uses similar motifs to recognize and bind DNA (see Fig. 23.9 to Fig. 23.16). Note that this classification is merely based on the several hundred complex crystal structures solved so far. The introduced groups can be further subdivided into 54 families according to their structure. To document the vari-

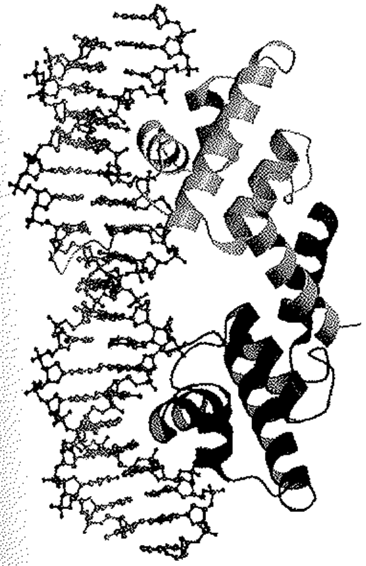

Fig. 23.9 Helix-turn-helix proteins. This motif is characterized by two almost orthogonal α-helices linked by a turn. The second helix usually lies in the major groove. One example is the λ repressor. This figure also appears with the color plates.

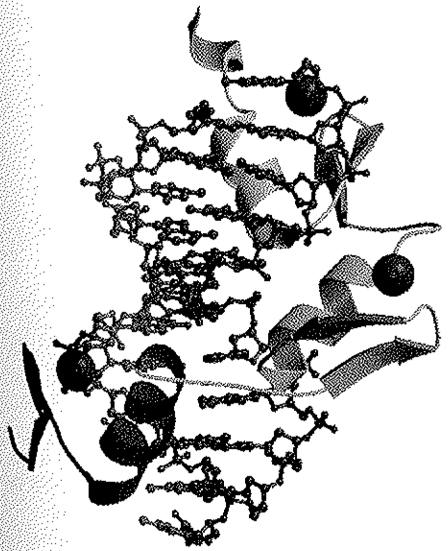

Fig. 23.10 Zinc-coordinating proteins. This greatest group so far contains many eukaryotic transcription factors that can already be identified by their sequence patterns. The characteristic feature is one or two zinc atoms coordinated by conserved cysteins or histidines. One example is the so-called zinc finger of the mouse transcription factor Zif268. DNA-binding in this group also occurs via a helix in the major groove. This figure also appears with the color plates.

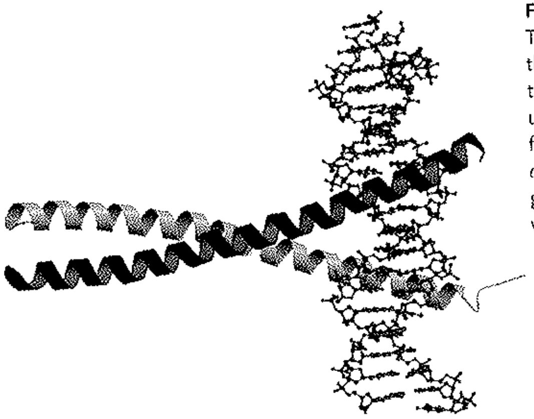

Fig. 23.11 Zipper-type proteins. This group name is derived from the groups' mode of dimerization that resembles a zipper. This figure shows the yeast transcription factor GCN4; one end of the long α-helices rests in the DNA's major groove. This figure also appears with the color plates.

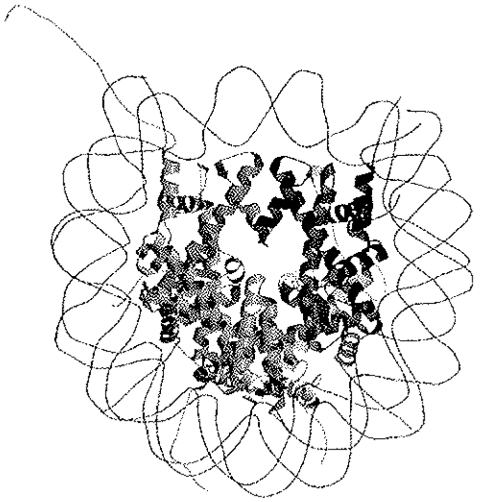

Fig. 23.12 α-Helix group. This group summarizes all other families that use an α-helix as DNA-binding element. The most prominent example are the histones, which form the nucleosome core particle. This figure also appears with the color plates.

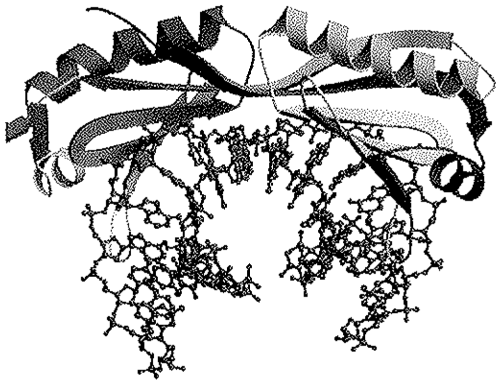

Fig. 23.13 β-Sheet group. So far, this group contains only one family, the TATA box-binding proteins (TBP), that are an essential component of the multiprotein transcription initiation complex. A prominent feature of the structures is a bending of the DNA by almost 90°. This figure also appears with the color plates.

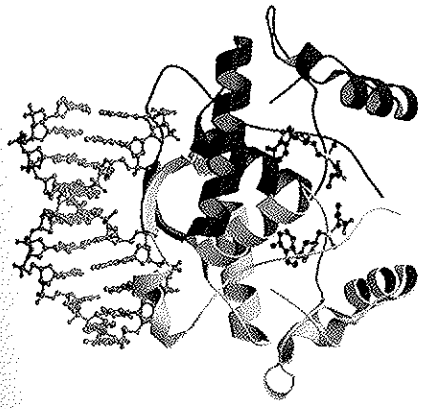

Fig. 23.14 β-Hairpin/ribbon group. Representatives of this group use short β-sheets or coil motifs in the DNA's minor or major groove. One example is the Met repressor. This figure also appears with the color plates.

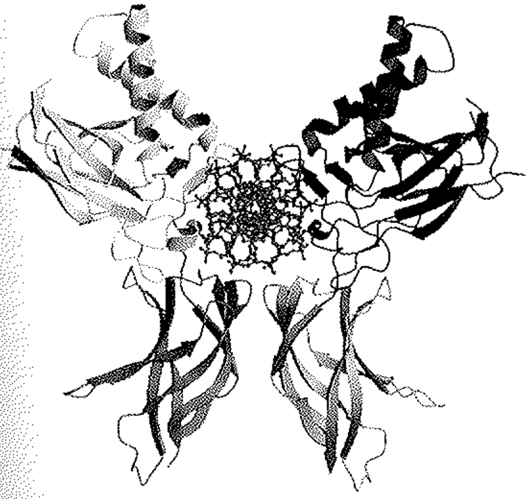

Fig. 23.15 Other protein–DNA interactions. This group contains two families so far, both of which show somewhat complicated protein–DNA interactions that utilize several structural elements. The example shown here is taken from the so-called *Rel homology region* family of the eurkaryotic transcription factor NF-κB. Also compare Fig. 23.5. This figure also appears with the color plates.

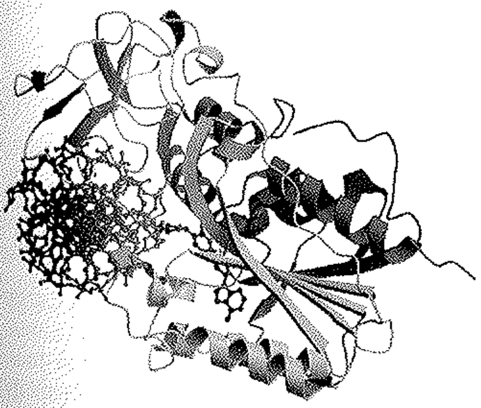

Fig. 23.16 Enzymes. This group defines families based on their members' enzymatic function instead of their structure. One example shown is the methyl transferase. This figure also appears with the color plates.

DNA and RNA-bindingDNA and RNA-binding",4,1> proteins in the human- ome as well as in the genomes of fully sequenced model organisms. fication is based on sequence. RNA indicates RNA-binding domains, all nains are DNA-binding.

	Human	*Drosophila*	*Caenorhabditis*	*Saccharomyces*	*Arabidopsis*
ore domain	75 (81)	5	71 (73)	8	48
p-helix DNA- domain	60 (61)	44	24	4	39
ox domain	160 (178)	100 (103)	2 (84)	6	66
DNA-binding	32 (43)	18 (24)	17 (24)	15 (20)	243 (401)
pper domain	114	55	36	16	134
A-binding	7	2	1	1	0
ding protein,	2 (4)	4 (8)	2 (4)	1 (2)	2 (4)
nc finger s	77 (100)	34 (37)	50 (72)	19 (21)	87 (102)
ger, C_2H_2-type	564 (4500)	234 (771)	68 (155)	34 (56)	21 (24)
ger, C_3HC_4-type finger)	135 (137)	57	88 (89)	18	298 (304)
NA-binding s	46 (47)	26 (27)	19	6	7
ox helicase	63 (66)	48 (50)	55 (57)	50 (52)	84 (87)
ain (RNA)	28 (67)	14 (32)	17 (46)	4 (14)	27 (61)
NA recognition RNA)	224 (324)	127 (199)	94 (145)	43 (73)	232 (369)

s in parentheses count the number of domains, e.g., 28 (67), in the KH do- e means, that the human genome contains 28 proteins with a total of 67 do- Several proteins thus contain more than one KH domain. Other DNA-binding contain the so-called ARID and *forkhead* domain. The RFX domain is an un- a helix-turn-helix motif and thus related to the homeobox proteins. The helix- x DNA-binding domain resembles the leucin zipper group as well as the nain (modified from [3]).

ety, one representative of each group is briefly introduced above. Table 23.3 gives an overview over the classes' respective frequency in the human genome.

23.2.4 Medical Relevance of Protein–DNA Interactions

Numerous diseases are caused by **incorrect protein–DNA interactions.** The great importance of these interactions stems from the fact that DNA-binding transcription factors are central switches of a cell's regulatory network. The **transcription factor SRY,** for instance, is sufficient to trigger the male sex. Mutations in the protein can lead to a sex change of affected embryos. Quite a number of **hormone receptors** like the glucocorticoid or oestrogen receptor are zinc finger proteins, which play an important part in the hormone-controlled metabolism. Finally, cancer can also be caused by mutated transcription factors. The proteins Jun and Fos are well examined **leucine zipper proteins** that do not simply have to bind the correct promoter sequences, but can only fulfill this physiological function as a Jun/Fos protein complex.

23.2.5 Biotechnological Applications of Protein–DNA Interactions

The detailed knowledge of **protein–protein** and **protein–DNA** interactions allows us to specifically manipulate them for different purposes. In some cases, DNA-binding proteins can be manipulated in such a manner, that they recognize specific DNA sequences. One aim is, for example, the creation of DNA-binding proteins that specifically recognize and bind to defined target genes to activate or deactivate them. If specific promoters were to be placed before such genes, they could be manipulated in a tissue-specific manner. Another aim is the specific activation of DNA binding through added substances. One example for the latter is the **Tet repressor** that binds to the Tet operator DNA sequence as well as to the antibiotic **tetracycline.** This system can be modified in such a manner that the addition of tetracycline or related substances can induce or inhibit the DNA binding. It is possible to insert the Tet repressor gene and a target gene under the Tet operator into mammalian cells and then switch the target gene on or off simply via the addition of tetracycline (Tet system). Many laboratories work on the manipulation of DNA-binding domains that recognize specific sequences (even unnatural sequences when required). In the case of restriction enzymes, proteins could be created which cleave DNA on predefined sites. With an increase of detailed knowledge, numerous applications are imaginable that not only allow for the manipulation of bacteria, animals, and plants, but also make alterations of the human genome possible. While therapeutic applications are desired, it will be a great challenge to avoid unintentional side effects and misuse.

Further Reading

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., Walter, P. **2002**, *Molecular Biology of the Cell*, 4th edn, Garland Science, New York.

Böhm, H.-J., Schneider, G. 2003, Protein-ligand interactions: from molecular recognition to drug design, in *Methods and Principles in Medicinal Chemistry*, vol. 19, Mannhold, R., Kubinyi, H., Folkers, G (eds.), Wiley-VCH, Weinheim.

Buchner, J., Kiefhaber, T. **2005**, *Protein Folding Handbook*, Wiley-VCH, Weinheim.

Cramer, P., Bushnell, D.A., Cornberg, R.D. **2001**, Structural basis of transcription: RNA polymerase II at 2.8 Å resolution, *Science* 292, 1863–1876.

Cesareni, G., Gimona, M., Sudol, M., Yaffe, M. **2004**, *Modular Protein Domains*, Wiley-VCH, Weinheim.

Collins, C.H., Yokobayashi, Y., Umeno, D., Arnold, F.H. **2003**, Engineering proteins that bind, move, make and break DNA, *Curr. Opin. Biotechnol.* 14, 371–378.

Darnell, J.E. **2002**, Transcription factors as targets for cancer therapy, *Nat. Rev. Cancer* 2, 740–749.

Frishmann, D., Mewes, H.aW. **1997**, PEDANTic genome analysis, *Trends Genet.* 13, 415–416.

Gavin, A.C., Superti-Furga, G. **2003**, Protein complexes and proteome organization from yeast to man, *Curr. Opin. Chem. Biol.* 7, 21–27.

Ghosh, S., Karin, M. **2002**, Missing pieces in the NF-κB puzzle, *Cell* 109 (Suppl.), S81–S91.

Golemis, E. (ed.) **2002**, *Protein–Protein Interactions: a Molecular Cloning Manual*, Cold Spring Harbor Laboratory Press, New York.

Heller, K.J. **2003**, *Genetically Engineered Food: Methods and Detection*, Wiley-VCH, Weinheim.

Janin, J. **2000**, Kinetics and thermodynamics of protein–protein interactions from a structural perspective, in *Protein–Protein Recognition*, Kleanthous, C. (ed.), Oxford University Press, Oxford.

Jen-Jacobson, L., Engler, L.E., Jacobson, L.A. **2000**, Structural and thermodynamic strategies for site-specific DNA-binding proteins, *Structure* 8, 1015–1023.

Jones, S., Thornton, J.M. **2000**, Analysis and classification of protein–protein interactions from a structural perspective, in *Protein–Protein Recognition*, Kleanthous, C. (ed.), Oxford University Press, Oxford.

Lodish, H., Berk, A., Zipursky, L., Matsudaira, P., Baltimore, D., Darnell, J. **2003**, *Molecular Cell Biology*, 5th edn, Freeman, New York.

Luscombe, N.M., Austin, S.E., Berman, H.M., Thornton, J.M. **2000**, An overview of the structures of protein–DNA complexes, *Genome Biol.* 1(1), 1–10.

Luscombe, N.M., Laskowski, R.A., Thornton, J.M. **2001**, Amino acid-base interactions: a three-dimensional analysis of protein–DNA interactions at an atomic level, *Nucl. Acids Res.* 29, 2860–2874.

Moss, T. (ed.) **2001**, *DNA-Protein Interactions: Principles and Protocols (Methods in Molecular Biology)*, Humana Press, Totowa.

Pawson, T., Nash, P. **2003**, Assembly of cell regulatory systems through protein interaction domains, *Science* 300, 445–452.

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

Salwinski, L., Eisenberg, D. **2003**, Computational methods of analysis of protein–protein interactions, *Curr. Opin. Struct. Biol.* 13, 377–382.

Schwikowski, B., Uetz, P., Fields, S. **2000**, A network of interacting proteins in yeast, *Nat. Biotechnol.* 18, 1257–1261.

Sensen, C.W. **2005**, *Handbook of Genome Research: Genomics, Proteomics, Metabolomics, Bioinformatics, Ethical and Legal issues*, vol. 2, Wiley, New York.

Thorner, J. (ed.) **2000**, Applications of chimeric genes and hybrid proteins, part C: protein–protein interactions and genomics, *Methods in Enzymology* 328, Academic Press, New York.

Travers, A. **2000**, *DNA-Protein Interactions: a Practical Approach*, Oxford University Press, Oxford.

Urnov, F.D., Rebar, E.J. **2002**, Designed transcription factors as tools for therapeutics and functional genomics, *Biochem. Pharmacol.* 64, 919–923.

Venter, C.J., Adams, M.D., Myers, E.W. **2001**, The sequence of the human genome, *Science* 291, 1343.

Walsh, G. **2002**, *Proteins: Biochemistry and Biotechnology*, Wiley, Chichester.

Westermeier, R., Naven, T. **2002**, *Proteomics in Practice: A Laboratory Manual of Proteome Analysis*, Wiley-VCH, Weinheim.

Whitford, D. **2005**, *Proteins: Structure and Function*, Wiley, New York.

24 Bioinformatics

Learning Objectives ***Bioinformatic methods hold the key to the analysis and understanding of large amounts of data gathered from genomics, functional genomics, proteomics, and molecular diagnostics to name but a few. This chapter will introduce you to different methods and problems of bioinformatics.***

24.1 Introduction

Bioinformatics as a discipline arose from the necessity to process and analyze sequencing data. The availability of large amounts of data provided by molecular biological techniques consequently led to the development of computer processes to store and compare this data. The process repeated itself about two decades later with the development of DNA chips, which promised insights into the **transcriptome** (after the genome had already been investigated). Ironically, bioinformatics was not meant to be more than an auxiliary discipline at first. However, it quickly rose to a **full discipline** in its own right (especially in the field of sequence analysis) and has significantly contributed to biological knowledge since. For instance, the investigation of evolutionary processes, which are not accessible to conventional experiments, has only become possible through the help of mathematical methods and statistical analysis of sequence data. In fact, today's arrangement of the tree of life is based on molecular similarity instead of morphological criteria (Fig. 1.1).

Bioinformatics can be subdivided based on fields of application or methods employed. Selected applications of **bioinformatic methods** (in temporal order) would be sequence alignment, database search, motif recognition, phylogenetic analysis, structure prediction of RNA and proteins, gene prediction, promoter analysis, transcriptome analysis, proteome analysis, and modelling of complex biological systems. These methods contain algorithms to determine similarities in series of characters (including addition, deletion and alteration of letters), methods from graph theory, statistical procedures (e.g., maximum likelihood estimation), methods of machine learning (e.g., artificial neural networks, hidden

An Introduction to Molecular Biotechnology. Edited by M. Wink