Chapter 23.3. Protein–ligand interactions

A. E. HODEL AND F. A. QUIOCHE

23.3.1. Introduction

There are currently over a thousand unique protein-bound ligands described in the Protein Data Bank, illustrating the enormous variety of small molecules with which proteins interact. These ligands serve as cofactors in protein-mediated reactions, substrates in these reactions, and elements that maintain or alter protein structure or macromolecular assembly. The specific binding of small molecules to proteins is a primary means by which living systems interact and exchange information with their environment. The atomic details of protein–ligand interactions are often quite similar to the intramolecular interactions observed within a protein molecule. Examples of all the various non-covalent interactions described in Parts 20 and 22, such as hydrogen bonds, van der Waals forces and other electrostatic phenomena, are observed between proteins and their small-molecule ligands.

Through the diverse interactions observed between proteins and their ligands, a few fundamental patterns of recognition emerge. In general, ligand binding requires that the protein partially or fully sequesters the ligand from the solvent. This demands that the energy of interaction between the protein and the ligand must be strong enough to overcome the interactions between both species and the solvent as well as the translational and rotational entropy which is lost upon fixing the orientation of the ligand relative to the protein. The protein achieves this level of interaction by presenting a binding site that is complementary to the ligand both in shape and electrostatic functionality. Beyond this generalization, each ligand has its own unique and complicated story. Rather than attempt to summarize the enormous subject of protein–ligand interactions in a comprehensive manner, we will instead illustrate several of the unique interactions observed between proteins and other molecules.

23.3.2. Protein–carbohydrate interactions

The interaction between proteins and carbohydrates provides a prototypic example of how proteins specifically recognize small organic ligands. Protein-mediated recognition of carbohydrates is crucial in a diverse array of processes, including the transport, biosynthesis and storage of carbohydrates as an energy source, signal transduction through carbohydrate messengers, and cell–cell recognition and adhesion (Rademacher et al., 1988). Many aspects of protein–carbohydrate recognition are observed in the interactions of proteins with other small-molecule ligands. For example, carbohydrate recognition is largely conferred through a combination of hydrogen bonding and van der Waals interactions with the protein (Quiocho, 1986). These interactions are generally presented in a binding site that is highly complementary to the target ligand in shape and functionality. Carbohydrate-binding proteins also occasionally employ ordered water molecules and bound metal ions to facilitate ligand recognition (Quiocho et al., 1989; Vyas, 1991; Weis & Drickamer, 1996).

There are also many examples of ‘induced fit’ recognition of carbohydrates, where the binding of the ligand induces a conformational change in the protein. This phenomenon of ligand-induced conformational change was observed in one of the first ligand–protein interactions visualized by X-ray crystallography, namely the binding of a carbohydrate substrate to lysozyme (Blake et al., 1967). Since these early studies, a wide variety of carbohydrate–protein structures have been determined, and a number of general themes have emerged from this continually active field (Vyas, 1991). These themes are further applicable in the general study of protein–ligand recognition.

Two general classes of carbohydrate-binding modes have been observed in their complexes with proteins (Quiocho, 1986; Vyas, 1991). The first group of proteins completely sequester the carbohydrate from the surrounding solvent. These proteins, including the periplasmic proteins and the catalytic site of glycogen phosphorylase, tend to have a high affinity for their ligand \(K_d \approx 10^{-6}–10^{-7}\). The high affinity of these proteins can be attributed to the entropic gain of desolvating the protein and carbohydrate surfaces as well as the exceptionally high degree of functional complementarity in the ligand-binding site. The periplasmic proteins nearly saturate the hydrogen-bonding potential of their carbohydrate ligands (Quiocho, 1986). The second group of proteins bind their ligands in shallow clefts in the solvent-exposed protein surface. This mode of binding, observed in the lectins, lysozyme and the storage site of glycogen phosphorylase, generally has a lower binding constant than the first class of binding proteins \(K_d \approx 10^{-3}–10^{-4}\). Some members of this second class of proteins, such as the lectins, are able to increase the affinity for their ligands dramatically by clustering a number of low-affinity sites through the oligomerization of the polypeptides (Weis & Drickamer, 1996). The specific arrangement of lectin oligomers allows them to discriminate between large cell-surface arrays of polysaccharides with high affinity and selectivity.

23.3.2.1. Carbohydrate recognition at the atomic level

An early review of protein–carbohydrate interactions revealed several atomic level interactions that continue to appear ubiquitously in the structures of protein–carbohydrate complexes (Quiocho, 1986). A generic cyclic sugar in mono- or oligosaccharides appears to be recognized as a disc that displays two flat non-polar surfaces surrounded by a ring of polar hydroxyls. Proteins recognize these features by hydrogen bonding to the ring of polar hydroxyls while stacking flat aromatic side chains against the non-polar disc faces.

Cooperative hydrogen bonding, where the hydroxyl group of a carbohydrate participates as both a donor and acceptor of hydrogen bonds, is often observed in the direct interactions between proteins and carbohydrate ligands (Fig. 23.3.2.1). The sp²-hybridized oxygen atom of a carbohydrate hydroxyl may act as both an acceptor of two hydrogen bonds through the two lone pairs of electrons as well as a donor of a single hydrogen bond. Cooperative hydrogen bonding generally follows a simple pattern in which the carbohydrate hydroxyl accepts a hydrogen bond from a protein amide group while simultaneously donating a hydrogen bond to a protein carbonyl oxygen. Hydrogen bonding to protein hydroxyl groups is observed only infrequently.
Metal ions associated with proteins

Metal ions are important in protein function and structure. They provide a number of functions, including stabilization of enzyme catalysis and protein function, coordination to specific protein functional groups, and structure and orientation of the protein tertiary structure through ligand interactions. Tryptophan, the aromatic amino acid with the largest surface area and highest electronegativity, is the most common side chain employed in van der Waals ‘stacking’ with carbohydrates. The infrequent use of aliphatic groups in the binding of the non-polar carbohydrate faces suggests that the aromatic moieties are employed in a specific manner. The electron-rich electron clouds of the aromatic side chains may provide a strong electrostatic interaction with the aliphatic carbohydrate protons that could not be satisfied by protein aliphatic groups. The anionic character of aromatic side chains is observed in a number of protein–intramolecular (Chapter 22.3) and protein–ligand interactions (see below).

This pattern is thought to be a result, in part, of the entropic cost of fixing a freely rotating protein hydroxyl group while simultaneously fixing the ligand hydroxyl group. Amides and carbonyls are usually fixed in a planar geometry and thus do not require as much energy to compensate for their loss of entropy in ligand binding.

The vicinal hydroxyl groups of carbohydrates provide an ideal geometry for the formation of ‘bidentate’ hydrogen bonds, where the pair of hydroxyls interacts with two functional groups of a single amino-acid side chain or the main-chain amide groups of two consecutive residues (Fig. 23.3.2.1). These interactions occur when the adjacent carbohydrate hydroxyls are either both equatorial, or one is equatorial and the other axial. The interatomic distance for the carbohydrate hydroxyl oxygens is \(\sim 2.8\) Å in these cases, allowing for a bidentate interaction with the planar side chains of aspartate, asparagine, glutamate, glutamine and arginine. Bidentate hydrogen bonds have not been observed for consecutive axial hydroxyls where the oxygen–oxygen distance increases to \(\sim 3.7\) Å.

Carbohydrates often present a disc-like face of non-polar aliphatic hydrogen atoms which proteins recognize through the use of aromatic side chains. The protein aromatic groups are ‘stacked’ on the flat face of the carbohydrate, thus generating both specificity and binding energy through van der Waals interactions. Tryptophan, the aromatic amino acid with the largest surface area and highest electronegativity, is the most common side chain employed in van der Waals ‘stacking’ with carbohydrates. The infrequent use of aliphatic groups in the binding of the non-polar carbohydrate faces suggests that the aromatic moieties are employed in a specific manner. The electron-rich electron clouds of the aromatic side chains may provide a strong electrostatic interaction with the aliphatic carbohydrate protons that could not be satisfied by protein aliphatic groups. The anionic character of aromatic side chains is observed in a number of protein–intramolecular (Chapter 22.3) and protein–ligand interactions (see below).

Table 23.3.3.1
Metal ions associated with proteins

<table>
<thead>
<tr>
<th>Metal ion</th>
<th>Concentration in blood plasma (mM)</th>
<th>Common cofactors</th>
<th>Hard/soft classification</th>
<th>Common coordination number and geometry</th>
<th>Preferred ligand atom</th>
</tr>
</thead>
<tbody>
<tr>
<td>Na(^+)</td>
<td>1.38</td>
<td></td>
<td>Hard</td>
<td>6</td>
<td>O</td>
</tr>
<tr>
<td>K(^+)</td>
<td>4</td>
<td></td>
<td>Hard</td>
<td>8</td>
<td>O</td>
</tr>
<tr>
<td>Ca(^{2+})</td>
<td>3</td>
<td></td>
<td>Hard</td>
<td>8</td>
<td>O</td>
</tr>
<tr>
<td>Mg(^{2+})</td>
<td>1</td>
<td></td>
<td>Hard</td>
<td>6</td>
<td>O</td>
</tr>
<tr>
<td>Fe(^{2+})</td>
<td>0.02</td>
<td></td>
<td>Intermediate</td>
<td>6 (octahedral)</td>
<td>O</td>
</tr>
<tr>
<td>Zn(^{2+})</td>
<td>0.02</td>
<td></td>
<td>Intermediate</td>
<td>4 (tetrahedral), 6 (octahedral)</td>
<td>N,S</td>
</tr>
<tr>
<td>Cu(^{2+})</td>
<td>0.015</td>
<td></td>
<td>Soft</td>
<td>4 (tetrahedral)</td>
<td>S</td>
</tr>
<tr>
<td>Co(^{2+})</td>
<td>0.002</td>
<td></td>
<td>Hard</td>
<td>6 (octahedral)</td>
<td>O</td>
</tr>
<tr>
<td>Mn(^{2+})</td>
<td>0</td>
<td></td>
<td>Hard</td>
<td>6 (octahedral)</td>
<td>O</td>
</tr>
<tr>
<td>Ni(^{2+})</td>
<td>0</td>
<td></td>
<td>Intermediate</td>
<td>6 (octahedral)</td>
<td>N</td>
</tr>
<tr>
<td>Mo</td>
<td>0</td>
<td></td>
<td>Intermediate</td>
<td>6 (octahedral)</td>
<td>S</td>
</tr>
<tr>
<td>W</td>
<td>0</td>
<td>Pterin</td>
<td>Intermediate</td>
<td>6 (octahedral)</td>
<td>S</td>
</tr>
<tr>
<td>V</td>
<td>0</td>
<td>Pterin</td>
<td>Intermediate</td>
<td>5 (trigonal bipyramidal)</td>
<td>S</td>
</tr>
</tbody>
</table>

23.3.3. Metals

Metal ions provide a number of important functions in their diverse and ubiquitous interactions with proteins. The most common function for a protein-bound metal ion is the stabilization and orientation of the protein tertiary structure through coordination to specific protein functional groups. In addition to this structural role, metal ions are also often directly involved in enzyme catalysis and protein function. Examples of these functions include redox reactions, the activation of chemical bonds and the binding of specific ligands. Myoglobin, the first protein structure determined by X-ray crystallography, specifically binds molecular oxygen through an iron ion of a haem cofactor. Myoglobin provides a prototypic example of a protein and a metal ion providing a unique and specific functionality through their combination.

23.3.3.1. Metals important in protein function and structure

A number of metals are relatively abundant and available in living systems (Table 23.3.3.1) (Glusker, 1991). The most common ions include sodium, potassium, magnesium and calcium. Along with these ions, a large variety of trace metals are also found coordinated to proteins. The structures of protein complexes with some of these trace ions, including iron, zinc and copper, have been studied extensively for some time (Glusker, 1991). More recently, the structures of protein complexes with more unusual ions, such as nickel, vanadium and tungsten, have been determined (Volbeda et al., 1996).

Specificity in the interactions between proteins and metal ions is conferred through each ion’s preference for the coordinating atoms and the geometry of the binding site. All four of the more common metals, i.e. sodium, potassium, magnesium and calcium, are classified as ‘hard’ metals, referring to the polarizability of the

...
23.3. PROTEIN–LIGAND INTERACTIONS

23.3.1. Coordination number

The coordination number of a metal ion is defined as the number of ligands or donor atoms that interact with the ion. This number can range from 1 to 6, with 4 and 6 being the most common. It is determined by the size of the metal ion, the size and basicity of the ligands, and the stability of the resulting complexes. The coordination number also affects the stability of the complex, with higher coordination numbers generally leading to more stable complexes due to the increased overlap of orbitals and the resulting increase in charge delocalization. However, the coordination number is not always the only factor determining the stability of a complex, as other factors such as the nature of the ligand and the stability of the coordination sphere can also play a role.

23.3.2. Specific binding

Specific binding is a type of ligand–protein interaction where the ligand is tightly bound to the protein, resulting in a high affinity and specificity. This type of binding is typically observed in protein–ligand interactions, where the ligand is a small molecule that binds to a specific site on the protein. Specific binding is mediated by a combination of non-covalent interactions, including hydrogen bonding, van der Waals forces, and hydrophobic interactions. The specificity of binding is often determined by the fit of the ligand to the binding site, with the ligand conforming to the shape and chemical features at the site.

23.3.3. PROTEIN–NUCLEIC ACID INTERACTIONS

DNA is a double-stranded helix composed of two polynucleotide strands held together by hydrogen bonding between complementary bases. The bases are adenine (A), thymine (T), guanine (G), and cytosine (C). The two strands are antiparallel, with the A-T and G-C base pairs forming the hydrogen bonds. The stability of the DNA double helix is determined by the number of base pairs and the nature of the ligand binding to the DNA.

23.3.4. Protein–nucleic acid interactions

DNA provides one of the more compelling protein ‘ligands’ for biophysical study, as the sequence-specific binding of proteins to the DNA double helix mediates the interaction between the environment surrounding the living cell and the information ‘programmed’ into the cell within its genome. A classic example of such a process is the response of the bacteria Escherichia coli to the nutrients in the surrounding media through the regulation of gene expression. A simple case of this interaction is found in the biosynthesis of the amino acid tryptophan. The transcription of the genes necessary for the synthesis of tryptophan is suppressed when tryptophan is present in the environment. This process is mediated by the tryptophan-dependent sequence-specific binding of the trp repressor protein to the trp operon within the genes encoding the metabolic enzymes (Joachimiak et al., 1983). In the absence of tryptophan, the affinity of the aporepressor for the trp operon is dramatically reduced. Thus, when tryptophan is not available in the environment, transcription of the biosynthetic genes proceeds. In mammalian cells, the analogous process is observed in the activation of gene expression through hormones, cytokines and other stimuli.

Although DNA has often been considered to be a long, nearly featureless cylindrical double helix, proteins have evolved with exquisite specificity for their cognate DNA sequences. This apparent contradiction can be reconciled with the acknowledgment of two recently appreciated properties of DNA (Harrington & Winicov, 1994). First, the local structure of DNA is actually highly variable and dependent on the specific sequence of the base pairs in the helical ladder. Second, the DNA double helix is a relatively soft structure that is easily deformed into concerted bends, kinks and other distortions. DNA-binding proteins thus recognize their cognate sequences both by utilizing the unique local structure of the double helix and by inducing distortions into the helix which facilitate recognition.

The most intuitive features of the double helix that are important in sequence-specific recognition are the unique surfaces presented by the bases in the helix grooves. DNA is primarily found in a B-form helix that presents a wide, accessible major groove and a deep, narrow minor groove. An analysis of the arrangement of hydrogen-bonding functional groups presented by DNA bases (Fig. 23.3.4.1) suggests that the sequence-specific recognition of the DNA helix is best facilitated through the major groove, where each of the four possible base-pair combinations present unique hydrogen-bonding patterns (Steitz, 1990). The majority of sequence-specific DNA-binding proteins of known structure appear to utilize this direct readout of the major groove by inserting a portion of an α-helix, a two-stranded β-hairpin, or even a peptide coil which presents complementary hydrogen-bonding arrangements with the DNA bases (Pabo & Sauer, 1992; Steitz, 1990). The narrow surface of the minor groove presents some characteristic hydrogen-bonding patterns: however, the absolute identity of each base pair is

23.3.6. DNA-protein interactions

DNA-protein interactions are critical for many biological processes, including transcription, replication, and repair. The interaction between DNA and proteins is mediated by a combination of non-covalent forces, including hydrogen bonding, electrostatic interactions, hydrophobic interactions, and van der Waals forces. The specificity of these interactions is determined by the interaction between the protein and the DNA sequence, with the protein recognizing specific DNA sequences through the formation of protein–DNA complexes. The interaction between DNA and proteins can be further modulated by the presence of other factors, such as nucleosomes or transcription factors, which can affect the stability and specificity of the interaction.
The major groove at an angle of approximately 30° (15). Protein–DNA complexes are quite variable. Most helices bind in the major groove (Fig. 23.3.4.2). The exact orientations of helices in various proteins suited for binding in the major groove of B-DNA helix make it hard to distinguish AT base pairs from TA base pairs and GC base pairs from CG base pairs. Although there are proteins that recognize DNA through the minor groove, such as the TATA-box binding protein, the recognition of their target is completed through dramatic distortion of the DNA helix through intercalation (see below).

α-Helices are the most frequently observed structural motif for recognition in the major groove of DNA (Pabo & Sauer, 1992). The overall shape and dimensions of the α-helix are geometrically suited for binding in the major groove of a B-DNA helix (Fig. 23.3.4.2). The exact orientations of helices in various protein–DNA complexes are quite variable. Most helices bind in the major groove at an angle of approximately 30° (15°) from the plane normal to the DNA helical axis (Fig. 23.3.4.3). However, the numerous variants to this rule would include the trp repressor/operator complex, where only the N-terminal end of the ‘recognition’ helix is inserted into the major groove (Otwinowski et al., 1988a,b). Interactions observed between these inserted elements and the DNA bases include the common direct hydrogen bond between the protein side chain and base, the less common hydrogen bond between the protein backbone and base, indirect but specific hydrogen bonding through water molecules, and hydrophobic interactions.

There appears to be no simple correlation between the primary sequence of the peptide segments which make specific base contacts and the DNA sequence that those segments recognize (Pabo & Sauer, 1992; Steitz, 1990). Examples of every polar protein side chain participating in specific hydrogen bonds with DNA bases have been observed, but each amino acid does not show any preference for any one particular base. What is observed is that conserved residues within families of DNA-binding proteins tend to make conserved base-specific interactions in DNA–protein complexes. Strikingly, this subset of interactions which are conserved within protein families include cooperative hydrogen bonding reminiscent of the pairs of hydrogen bonds often observed in carbohydrate–protein complexes. These interactions, which include the pairing of arginine with guanine and glutamine or asparagine with adenine, were predicted early on by Seeman et al. (1976).

Although the elements of protein structure in direct contact with the DNA bases play a prominent role in sequence specificity, these elements are not sufficient to impart the specificity of the DNA-binding protein. This statement is supported by the variety of orientations in which the ‘recognition’ helices bind to the major groove. The structural context of the recognition elements and the overall docking of the protein to the DNA helix play an important role in specificity as the direct base interactions.

The contacts between the protein and the ribose–phosphate backbone of the DNA appear to be one of the more important aspects of the ‘indirect readout’ of the DNA sequence (Pabo & Sauer, 1992). On average, more than half of the interactions between protein and DNA in complex structures involve the backbone of the DNA helix. Thus, the sheer number of interactions suggests that these contacts serve an important function in recognition. Although several of the protein–DNA-backbone contacts observed involve salt bridges between the phosphates and basic protein side chains, these interactions are not as highly represented as one might expect. This could be a result of the high degree of flexibility inherent in the long side chains of arginine and lysine. Instead, examples of every basic and neutral residue and occasionally even acidic residue with some hydrogen-bonding potential interacting with the phosphate backbone have been observed. These contacts may contribute to specificity through two mechanisms. First, they can establish the exact
orientation of the base-specific contacts relative to the ‘rungs’ in the phosphate backbone. Second, they may read the base sequence indirectly through sequence-specific backbone distortions or flexibility. There are numerous examples of DNA–protein complexes with highly distorted DNA helices. There is also evidence that certain DNA sequences inherently confer bends within the B-form helix. Thus, it is conceivable that protein interactions with the DNA backbone may confer specificity by selecting for a specific distorted conformation of the helix.

The most dramatic distortion of the DNA helix has been observed in DNA–protein complexes where the protein induces a kink or bend through the intercalation of the DNA helix at the minor groove (Werner et al., 1996a,b). Intercalation involves the insertion of a hydrophobic protein side chain into the helix, disrupting the stacking of two adjacent base pairs, and, in some cases, the side chain itself then stacks with one of the base pairs. Examples of this mode of binding include the complexes of the TATA-box binding protein (TBP), the PurR repressor and the human oncogene ETS1 with their cognate DNA partners (Werner et al., 1996a,b). The ETS1–DNA complex provides the only current example of complete intercalation of the DNA extending from the minor groove to the major groove. A tryptophan side chain extends into the helix from the minor groove and stacks with one of the displaced base pairs. The remaining base pair contacts the ring system of the tryptophan edge in forming a pseudo-hydrogen bond between the indole hydrogens and the π-rings of the DNA bases. In ETS1, the deformation of the DNA helix resulting from protein intercalation results in the kinking of the helical axis from 45° to about 60°.

Examples of protein intercalation of the DNA helix from the major groove are found in proteins, such as the methyltransferases, that perform chemistry on the bases of the DNA. To perform their enzymatic function, these proteins must extract the target base from the DNA helix and ‘flip’ the base out into the enzyme active site (Cheng, 1995). The resulting void in the DNA is then filled by protein side chains that partially satisfy the hydrogen-bonding and van der Waals interactions that were broken when the target base was flipped. Although there are only a few known structures of DNA–protein complexes with extra-helical bases, base flipping is thought to be a relatively common feature of DNA-modifying enzymes.

23.3.4.3. RNA

Although RNA and DNA are chemically similar, RNA presents a much greater variety of shapes and surfaces compared to the relatively simple B-form helix of DNA. Generally single-stranded, RNA often forms secondary structures driven by the base pairing of complementary stretches of sequence within the same strand. The formation of base-paired regions can result in stem loops, bulges and helices which can further assemble into more complicated tertiary structures, such as that observed for transfer RNAs. Protein-mediated recognition of RNA often depends as much on the three-dimensional structure presented by these secondary structures as on the specific identity of the base sequence.

Very little information is currently available on the structural details of protein–RNA interactions (Nagai, 1996). Only a handful of protein–RNA complex structures have been determined. These fall into three basic categories, depending on the secondary structure of the RNA: four tRNA–protein complexes, two stem-loop–protein complexes and a capped single-stranded RNA–protein complex.

23.3.4.4. Transfer RNA

In the four known structures of tRNA bound to their aminoacyl tRNA synthetases (Cusack et al., 1996a,b; Goldgur et al., 1997; Rould et al., 1991), the effects of RNA's preference for A-form helices on recognition are immediately apparent. The proteins make numerous contacts in the shallow and exposed minor grooves of the RNA helices. This contrasts with the extensive use of the major groove in the recognition of B-form DNA helices. Beyond this generalization, the details of tRNA recognition differ in each specific case. Comparison of the protein-bound tRNA to the structure of free tRNA reveals that the proteins tend to distort the RNA conformation and partially unwind the helices near the anti-codon loop. In one case, namely the structure of glutamyl-tRNA synthetase (Rould et al., 1991), the final base pair near the acceptor stem of the tRNA is broken, and the CCA acceptor makes a dramatic hairpin turn into the enzyme active site.

23.3.4.5. Stem loops

One fascinating observation in viewing the structures of RNA-binding proteins, even in the absence of RNA, is that aside from the tRNA-binding synthetases, they all appear to have evolved from or towards a very similar general fold (Burd & Dreyfuss, 1994). This fold, exemplified by the RNP domain found in numerous RNA-binding proteins, consists of a β-sheet surrounded on one side by α-helices and solvent-exposed on the opposing face. This general folding architecture is found in RNP domains, ribosome proteins, K-homologous domains (KH), double-stranded RNA-binding domains and cold shock proteins. Although each of these subsets of RNA-binding domains has a different topology and most probably bind to RNA with different surfaces, they all appear to have this alpha–beta–solvent architecture.
Two proteins with this architecture have been co-crystallized with their specific RNA stem-loop ligands (Nagai et al., 1995; van den Worm et al., 1998). In both cases, the loop of the RNA binds to the open face of the $\beta$-sheet where solvent-exposed aromatic amino-acid side chains stack with the extrahelical bases of the RNA. Unpaired bases from the RNA also form numerous specific hydrogen bonds with protein side chains and polar backbone groups, imparting sequence specificity in the interaction. These structures suggest that the flat, open face of a $\beta$-sheet provides a good surface for RNA binding, where the extrahelical bases can make extensive and specific contacts with the protein.

23.3.4.6. Single-stranded sequence-nonspecific RNA–protein interactions

There is a single example of a single-stranded RNA–protein complex which is sequence-nonspecific. The structure of the vaccinia RNA methyltransferase VP39 bound to a 5$'$m$^7$G-capped RNA hexamer reveals a mechanism of nonspecific recognition reminiscent of the Klenow fragment–DNA tetramer complex (Hodel et al., 1998). The RNA forms two short single-stranded helices of three bases each. The first of these helices binds in the active site of VP39 solely through hydrogen bonds between the protein and the ribose–phosphate backbone. The bases of the RNA strand stack together as trimers, but do not form any interactions with the protein (Fig. 23.3.4.4). Like the Klenow–DNA complex, this observation suggests an intuitive mechanism for sequence-nonspecific nucleic acid binding, where the single-stranded RNA forms short transient helices driven by intramolecular stacking interactions. The protein then recognizes and stabilizes the helical backbone conformation formed by this transient stacking without interacting with the bases themselves.

23.3.4.7. The recognition of alkylated bases

The complex of VP39 with capped RNA also illustrates a final example of the diversity of protein–ligand interactions in the specific recognition of the 7-methylguanosine cap. When guanosine is methylated at the N7 position, a positive charge is introduced to the $\pi$-ring system of the base. Eukaryotic cells utilize the methylation of a guanosine base at the N7 position as a tag or cap for the 5$'$ end of messenger RNA. The m$^7$G$^5$ppp mRNA cap is specifically recognized in the splicing of the first intron in nascent transcripts, in the transport of mRNA through the nuclear envelope and in the translation of the message by the ribosome (Varani, 1997). Two structures of specific m$^7$G binding proteins are now known: VP39 and the ribosomal cap-binding protein IF-4E (Hodel et al., 1997; Marcotrigiano et al., 1997). Each structure offers clues as to how the proteins can discriminate between the charged methylated m$^7$G base and the unmodified guanosine base. The m$^7$G base is stacked between aromatic protein side chains and hydrogen bonded to acidic protein residues (Fig. 23.3.4.5). One long-held hypothesis is that IF-4E, with dual tryptophan residues, binds specifically to the positively charged form of the base through a charge-transfer complex (Ueda, Iyo, Doi, Inoue & Ishida, 1991). The formation of a charge-transfer complex is evident in small-molecule studies and spectroscopic studies with IF-4E (Ueda, Iyo, Doi, Inoue, Ishida et al., 1991). However, VP39 performs the same discrimination with the much less electronegative phenylalanine and tyrosine side chains (Hodel et al., 1997). So far, no charge-transfer complex has been observed in VP39.

The recognition of charged methylated bases is important not only in mRNA processing, but also in the repair and recognition of DNA damaged by alkylating carcinogens. The mechanism by which the charged m$^7$G base is recognized is probably similar to how other positively charged bases, such as 3-methyladenosine, O2-methylcytosine and O2-methylthymidine, are recognized. In fact, the E. coli DNA repair enzyme, AlkA, will catalyse the glycolysis of all of these bases (Lindahl, 1982). The structure of AlkA is known, but only in the absence of a substrate (Labahn et al., 1996). In this structure, a number of solvent-exposed tryptophan residues are found at the putative active site. This observation suggests that AlkA may recognize positively charged bases through an aromatic ‘sandwich’, much like that found in IF-4E and VP39.
23.3. Phosphate and sulfate

Novel features of molecular recognition and electrostatic interactions of these two tetrahedral oxyanions have emerged from our crystallographic and functional studies of the phosphate-binding protein (PBP) and sulfate-binding protein (SBP), which serve as extremely specific initial receptors for ATP-binding cassette (ABC)-type active transport or permease in bacterial cells. The complexes of these proteins have $K_d$ values in the low $\mu M$ range. Although phosphate and sulfate are structurally similar, at physiological pH PBP and SBP exhibit no overlap in specificity (Medveczky & Rosenberg, 1971; Pardee, 1966; Jacobson & Quiocho, 1988). This stringent specificity prevents one tetrahedral oxyanion nutrient from becoming an inhibitor of transport for the other. The specificity of the PBP-dependent phosphate transport system is also shared by other phosphate transport systems in eukaryotic cells and across brush borders and into mitochondria.

As described below, discrimination between anions is based solely on the protonation state of the ligand. Sulfate, a conjugate base of a strong acid, is completely ionized at pH values above 3, whereas phosphate, a conjugate base of a weak acid, remains protonated up to pH 13.

The structure of the PBP–phosphate complex was initially determined at 1.7 Å resolution (Luecke & Quiocho, 1990). The resolution has been pushed to an ultra high resolution of 0.98 Å, the first reported for a protein with a molecular weight as high as 34 kDa with a bound ligand (Wang et al., 1997). The bound phosphate is completely desolvated and sequestered in the protein cleft between two domains. It makes 12 hydrogen bonds with the proteins (11 with donor groups and one with an acceptor group), as well as one salt link to an Arg that is in turn salt-linked to an Asp residue (Fig. 23.3.5.1). The distances of the 12 hydrogen bonds between phosphate and PBP obtained from the ultra high resolution structure range from 2.432 to 2.906 Å (Wang et al., 1997). The Asp56 carboxylate, the lone acceptor group, plays two key roles in conferring the exquisite specificity of PBP. It recognizes, by way of the hydrogen bond, a proton on the phosphate and presumably disallows, by charge repulsion, the binding of a fully ionized sulfate dianion (Luecke & Quiocho, 1990).

The SBP binding-site cleft is also tailor-made for sulfate (Pflugrath & Quiocho, 1985). In keeping with the stringent specificity of SBP for fully ionized tetrahedral oxyanions (Pardee, 1966; Jacobson & Quiocho, 1988), the bound sulfate, which is also completely dehydrated and buried, is held in place by seven hydrogen bonds made entirely with donor groups from uncharged polar residues of the protein (Fig. 23.3.5.2) (Pflugrath & Quiocho, 1985). The absence of a hydrogen-bond acceptor group accounts for the inability of SBP to bind phosphate. Interestingly, the absence of a salt link and the formation of five fewer hydrogen bonds with the bound sulfate (Fig. 23.3.5.2b) than with the bound phosphate (Fig. 23.3.5.1b) do not make the affinity of the SBP–sulfate complex any weaker than that of the PBP–phosphate complex. In fact, the sulfate binds 10–20 times more tightly to SBP (Pardee, 1966; Jacobson & Quiocho, 1988). Also, the hydration energies of both anions are likely to be similar.

The ability of PBP and SBP to differentiate each oxyanion ligand through the presence or absence of proton(s) is an extremely high level of sophistication in molecular recognition. The importance of complete hydrogen bonding in recognition of buried ligands is powerfully demonstrated in PBP and SBP. As the sulfate is fully ionized (i.e. possesses no hydrogen at physiological pH), repulsion occurs at Asp56 of PBP specifically for this dianion. On the other hand, SBP is unable to bind phosphate because it contains no hydrogen-bond acceptor in the binding site. Significantly, despite the potential for a large number of matched hydrogen-bonding pairs, a single mismatched hydrogen bond (e.g. a fully ionized sulfate providing no proton for interaction with Asp56 of PBP and no acceptor group in SBP for a phosphate proton) represents a binding energy barrier of 6–7 kcal mol$^{-1}$ (1 kcal mol$^{-1}$ = 4.184 kJ mol$^{-1}$).

![Figure 23.3.4.5](image)

The specific recognition of the messenger RNA 7-methylguanosine cap. (a) The residues contacting the m$^7$G base in the cap-binding protein, IF-4E (Marcotrigiano et al., 1997). (b) The residues interacting with the cap in the vaccinia RNA methyltransferase VP39 (Hodel et al., 1997). Both proteins bind to the charged, methylated base by stacking aromatic amino acids on both sides of the base.
23. STRUCTURAL ANALYSIS AND CLASSIFICATION

23.3.5.1. Dominant role of local dipoles in stabilization of isolated charges

A novel finding of further paramount importance and wide implication is how the isolated charges of the protein-bound phosphate and sulfate are stabilized. No counter-charged residues or cations are associated with the sulfate completely buried in SBP. Although a salt link involving Arg135 is formed with the phosphate bound to PBP, it is shared with an Asp residue (Fig. 23.3.5.1b). Moreover, site-directed mutagenesis studies indicate that phosphate binding is quite insensitive to modulation of the salt link (Yao et al., 1996). These findings are a powerful demonstration of how a protein is able to stabilize the charges by means other than salt links. Experimental and computational studies indicate that local dipoles, including the hydrogen-bonding groups and the backbone NH groups from the first turn of helices, immediately surrounding the sulfate and phosphate are responsible for charge stabilization (Pflugrath & Quiocho, 1985; Quiocho et al., 1987; Åqvist et al., 1991; He & Quiocho, 1993; Yao et al., 1996; Ledvina et al., 1996). Helix macrodipoles play little or no role in charge stabilization of the anions. The same principle of charge stabilization by local dipoles also applies for the following buried uncompensated ionic groups: Arg151 of the arabinose-binding protein (Quiocho et al., 1987), the zwitterionic leucine ligand bound to the leucine/isoleucine/valine-binding protein (Quiocho et al., 1987), the potassium in the pore of the potassium channel (Doyle et al., 1998) and Arg56 of synaptobrevin-II in a SNARE complex (Sutton et al., 1998).

Figure 23.3.5.1
12 hydrogen-bonding interactions between the phosphate-binding protein (PBP) and phosphate. (a) Displacement ellipsoids of the atoms involved in the interactions from the 0.98 Å atomic structure (Wang et al., 1997). (b) Schematic diagram of the interactions, including additional hydrogen bonds.

Figure 23.3.5.2
Seven hydrogen-bonding interactions between the sulfate-binding protein (SBP) and sulfate. (a) Interactions based on the 1.7 Å structure (J. Sack & F. A. Quiocho, unpublished data). (b) Schematic diagram of the interaction.
23.3. PROTEIN–LIGAND INTERACTIONS

23.3.5.2. Short hydrogen bonds

The ultra high resolution refined structure of the PBP–phosphate complex is the first to show structurally the formation of an extremely short hydrogen bond (2.432 Å) between the Asp56 carboxylate of PBP and phosphate. Although this short hydrogen bond is within the proposed range of low-barrier hydrogen bonds with estimated energies of 12–24 kcal mol⁻¹ (Hibbert & Emsley, 1990), its contribution to phosphate binding affinity has been assessed to be no better than that of a normal hydrogen bond (Wang et al., 1997). Thus, a unique role for short hydrogen bonds in biological systems, such as in enzyme catalysis (Gerlt & Gassman, 1993; Cleland & Kreevoy, 1994), remains controversial.

23.3.5.3. Non-complementary negative electrostatic surface potential of protein sites specific for anions

The presence of an uncompensated negatively charged Asp56 is unusual for an anion-binding site, as observed in PBP. In fact, a related discovery of profound ramification is that the binding-cleft region of PBP has an intense negative electrostatic surface potential (Fig. 23.3.5.3a) (Ledvina et al., 1996). Non-complementarity between the surface potential of a binding region and an anion ligand is not unique to PBP. We have reported similar findings for SBP, a DNA-binding protein, and, even more dramatically, for the redox protein flavodoxin (Fig. 23.3.5.3b) (Ledvina et al., 1996). Evidently, for proteins such as these, which rely on hydrogen-bonding interactions with only uncharged polar residues for anion binding and electrostatic balance, a non-complementary surface potential is not a barrier to binding. This conclusion is supported by very recent fast kinetic studies of binding of phosphate to PBP and the effect of ionic strength on binding (Ledvina et al., 1998).

FAQ is an HHMI Investigator. The work carried out in his laboratory is supported in part by grants from NIH and the Welch Foundation.

References


Cusack, S., Yarechuk, A. & Tukalo, M. (1996a). *The crystal structure of the ternary complex of T. thermophilus seryl-tRNA synthetase with tRNA(Ser) and a seryl-adenylate analogue reveals a conformational switch in the active site*. EMBO J. 15, 2834–2842.

Cusack, S., Yarechuk, A. & Tukalo, M. (1996b). *The crystal structures of T. thermophilus lysyl-tRNA synthetase complexed with E. coli tRNA(Lys) and a T. thermophilus tRNA(Lys) transcript: anticodon recognition and conformational changes upon binding of a lysyl-adenylate analogue. EMBO J*. 15, 6321–6334.


23.3. PROTEIN–LIGAND INTERACTIONS


