COMMUNICATION

A Structural Analysis of Phosphate and Sulphate Binding Sites in Proteins

Estimation of Propensities for Binding and Conservation of Phosphate Binding Sites

Richard R. Copley and Geoffrey J. Barton

Laboratory of Molecular Biophysics
University of Oxford, Rex Richards Building, South Parks Road, Oxford OX1 3QU, U.K.

The high resolution X-ray structures of 38 proteins that bind phosphate containing groups and 36 proteins binding sulphate ions were analysed to characterise the structural features of anion binding sites in proteins. 34 of the 66 phosphates found were in close proximity to the amino terminus of an α-helix.

27% of phosphate groups bind to only one amino acid, but there is a wide distribution, with 3% of phosphates binding to seven residues. Similarly, there is a large variability in the number of contacts each phosphate group makes to the protein. This ranges from none (3% of phosphates) to nine (3% of phosphates). The most common number of contacts is two (23% of phosphates). The most commonly found residue at helix-type binding sites is glycine, followed by Arg, Thr, Ser and Lys. At non-helix binding sites, the most commonly found residue is Arg followed by Tyr, His, Lys and Ser. There is no typical phosphate binding site.

There are marked differences between propensities for phosphate binding at helix and non-helix type binding sites. Non-helix binding sites show more discrimination between the types of residues involved in binding when compared to the helix set. The propensities for binding of the amino acids reveal the expected trend of positively charged and polar residues being good at binding (although that for lysine is unexpectedly low) with the bulky non-polar residues being poor at binding. Bulky residues are less likely to bind with the amide nitrogen. Sulphate binding sites show similar trends.

Analysis of multiple sequence alignments that include phosphate and sulphate binding proteins reveals the degree of conservation at the binding site residues compared to the average conservation of residues in the protein. Phosphate binding site residues are more conserved than other residues in the protein with phosphate binding sites more highly conserved than sulphate binding sites.

Keywords: 3D-structure; phosphate-binding; anion-binding; conservation; propensity

Around 50% of known proteins bind or process compounds possessing phosphoryl groups (Schulz & Schirmer, 1979). The types of molecules bound include factors such as NAD, control intermediates such as phosphotyrosine residues, and substrates (e.g. AMP binding to adenylate kinase). A clear understanding of the factors governing phosphate binding would be of benefit for protein engineering studies, drug design and structure prediction.

Johnson (1984) made a comprehensive review of 30 phosphate binding sites in 19 proteins and sub divided them into binding-only and catalytic sites. She observed that catalytic sites were less well defined than pure binding sites and that arginine was a common ligand for phosphate. More recently, studies have concentrated on restricted protein classes such as dinucleotide binding proteins (e.g. Baker *et al.*, 1992). The limited number of phosphate binding proteins available to Johnson precluded a general statistical analysis of phosphate sites. However, with the expansion in structural studies, the high resolution three-dimensional structures of a large number of proteins co-crystallized with a phosphate containing moiety have become available. Recently, Chakrabarti (1993) has exploited the enlarged database to examine the geometry of, and residues involved in, anion binding. Chakrabarti (1983) restricts his analysis to the inorganic ions SO_4 and PO_4 and the majority of the ligands considered (41 out of 52) were sulphate ions. Since SO_4 is less frequently

Correspondence: G. J. Barton, Laboratory of Molecular Biophysics, University of Oxford, South Parks Road, Oxford, OX1 3QU, England.

0022-2836/94/390321-09 $08.00/0

implicated in a specific functional role than phosphate, the sites discussed by Chakrabarti (1993) are of limited biochemical interest. Accordingly, we here examine the interaction of all phosphate containing ligands (not just PO_4) to identify the structural features required for phosphate binding, the propensity of each amino acid for binding and the conservation of phosphate binding residues during the course of protein evolution. Since sulphate ions may bind to similar sites, we also compare and contrast the same features for sites identified by the crystallographer as binding to sulphate.

Phosphate groups are negatively charged, so are expected to bind to regions of positive charge in the protein. It has been suggested that this role may be met by positively charged amino acids, by the peptide group, by the *α-helix macrodipole* (Hol *et al.*, 1978; Hol, 1985) or by a contribution from all three. The nucleotide pyrophosphates found in the nucleotide binding proteins are all bound near the amino termini of α-helices (Wierenga *et al.*, 1985, 1986; Baker *et al.*, 1992). The Rossmann fold (Rossmann *et al.*, 1974) comprises a parallel six-stranded β-sheet with helices on both sides of the sheet. The pyrophosphate moiety is located at the end of an α-helix and at the "switch point" of the β-sheet (Branden, 1980). Such structures provide support for the helix dipole hypothesis, however, the general importance of the helix dipole in stabilising anion binding has been questioned recently in studies of SBP (Sulphate Binding Protein) (He & Quiocho, 1993). In SBP charge stabilization appears to be performed predominantly by peptide units confined to the first turns of the α-helices, with no significant role played by helix macrodipoles (He & Quiocho, 1993). Since the importance of the α-helix macrodipole to anion binding remains open to debate, we here sub classify phosphate binding sites into helix and non-helix types in order to locate structural differences between the two types.

Proteins are frequently crystallized from ammonium sulphate. The chemical properties of

Table 1
Choice of proteins

PDB code	Name	Ligand	Res (Å)	*R*
1ak3	Adenylate kinase	AMP & SO_4	1·9	0·284
1cdt	Cardiotoxin V4	PO_4	2·5	0·197
1cmb	*E. coli* Met repressor	PO_4	1·8	0·196
1cox	Cholesterol oxidase	FAD	1·8	0·153
1csc	Citrate synthase	CMC	1·7	0·188
1ctf	L7/L12 50 S ribosomal protein	SO_4	1·7	0·174
1drf	Dihydrofolate reductase	SO_4	2·00	0·189
1fnr	Ferredoxin reductase	NADP	2·2	0·226
1gal	Glucose oxidase	FAD	2·3	0·181
1gd1	Glyceraldehyde-3-phosphate dehydrogenase	NAD & SO_4	1·8	0·177
1gky	Guanylate kinase	GMP & SO_4	2·0	0·173
1gox	Glycolate oxidase	FMN	2·0	1·89
1gpb	Glycogen phosphorylase-*b*	PLP	1·9	0·190
1hho	Haemoglobin	PO_4	2·1	0·223
1ipd	3-Isopropylmalate dehydrogenase	SO_4	2·2	0·185
1lld	Lactate dehydrogenase	NADH	2·0	0·179
1mbd	Myoglobin (deoxy)	SO_4	1·4	0·188
1mrm	Mandelate racemase	SO_4	2·5	0·183
1nxb	Neurotoxin *b*	SO_4	1·38	0·24
1ofv	Oxidized flavodoxin	FMN	1·7	0·190
1pfk	Phosphofructokinase	ADP & FBP	2·4	0·165
1pgd	6-Phosphogluconate dehydrogenase	SO_4	2·5	0·185
1phh	*p*-Hydroxybenzoate hydroxylase	FAD	2·3	0·193
1pii	*N*-(5' phosporibosyl)anthranilate isomerase	PO_4	2·0	1·73
1pk4	Human plasminogen kringle 4	SO_4	1·90	0·142
1prc	Photosynthetic reaction centre	SO_4	2·3	0·193
1rn4	Ribonuclease T_1	PO_4	1·8	0·148
1rnb	Barnase	SO_4	1·9	0·214
1rnd	Ribonuclease A	DCG	1·5	0·190
1rnh	Selenomethionyl ribonuclease H	SO_4	2·0	0·198
1sar	Ribonuclease SA	SO_4	1·8	0·172
1snc	Staphylococcal nuclease	PTP	1·65	0·161
1thm	Thermitase	SO_4	1·37	0·166
1tmd	Trimethylamine dehydrogenase	FMN	2·4	Unrefined
1tpp	Beta-trypsin complex	SO_4	1·40	0·191
1wsy	Tryptophan synthase	PO_4	2·5	0·253
1ycc	Cytochrome *c*	SO_4	1·23	0·192
256b	Cytochrome *b*562	SO_4	1·4	0·164
2alp	Alpha-lytic protease	SO_4	1·7	0·131
2aza	Azurin (oxidised)	SO_4	1·8	0·157
2er7	Endothia aspartic proteinase	SO_4	1·6	0·142
2imn	Immunoglobulin domain	SO_4	1·97	0·149
2mhr	Myohemerythrin	SO_4	1·7	0·158

continued

Table 1 *continued*

PDB code	Name	Ligand	Res (Å)	*R*
2sar	Ribonuclease SA	GMP	1·8	0·175
2tgp	Trypsinogen complex	SO_4	1·90	0·200
2tsc	Thymidylate synthase	UMP	1·97	0·180
2wrp	Trp repressor	SO_4	1·65	0·18
3chy	Che Y	SO_4	1·66	0·151
3dfr	Dihydrofolate reductase	NADPH	1·7	0·152
3fgf	Basic fibroblast growth factor	SO_4	1·6	0·161
3gap	Catabolite gene activator protein	c-AMP	2·5	0·256
3grs	Glutathione reductase	FAD & PO_4	1·54	0·186
3icb	Calcium-binding protein	SO_4	2·3	0·178
3rn3	Ribonuclease A	SO_4	1·45	0·223
4blm	Beta-lactamase	SO_4	2·0	0·151
4enl	Enolase	SO_4	1·9	0·149
4sgb	Serine proteinase	SO_4	2·1	0·142
5enl	Enolase	2PG	2·2	0·148
5fbp	Fructose-1-6-biphosphate	F6P	2·1	0·177
5p21	c.-H-ras-p21	GNP	1·35	0·196
5p2p	Phospholipase A_2	DHG	2·4	0·189
5pti	Trypsin inhibitor	PO_4	1·0	0·197
5tim	Triose phosphate isomerase	SO_4	1·83	0·183
6ldh	Lactate dehydrogenase	SO_4	2·0	0·202
6tim	Triosephosphate isomerase	G3P	2·2	0·137
7aat	Aspartate aminotransferase	PLP	1·9	0·166
8atc	Aspartate carbamoyltransferase	PAL	2·5·	0·165
8cat	Catalase	NADPH	2·5	0·191
8rub	Rubisco	CAP	2·4	0·24
8rxn	Rubredoxin	SO_4	1·0	0·147
9icd	Isocitrate dehydrogenase	NADP	2·5	0·181

Phosphate binding proteins were extracted from the November 1992 release of the Protein Data Bank (PDB; Bernstein *et al.*, 1977). 203 proteins containing phosphorus in HETATM records were identified. This set was screened to select only those proteins that were solved by X-ray crystallography to a resolution of ≤2·5 Å, and with the exception of trimethylamine dehydrogenase (1tmd), all are refined. Ideally, we would use only the highest resolution refined structures. However, limiting the resolution to ≤2·0 Å as recommended by Morris *et al* (1992) would nearly halve the number of available phosphate binding structures from 146 to 79 PROCHECK (Laskowski *et al.*, 1993), a program that identifies atypical torsion angles and contacts, was used to examine a sample of the proteins analysed. In addition, regions of the proteins interacting with phosphates were visually inspected on a molecular graphics system. These screens suggested that the 2·5 Å cutoff provided data of adequate quality for the analysis

In order to limit the bias towards proteins that are highly represented in the PDB, all pairs of the sequences of the remaining 144 proteins were compared to each other using a standard dynamic programming algorithm (Smith & Waterman, 1981) and grouped into families on the basis of their sequence similarity. This revealed 48 families of proteins that did not show strong sequence similarity to each other The highest resolution structure in each family was selected for further analysis. Where the resolution of 2 family members was identical, the structure with the lowest *R*-factor value was chosen.

Sulphate binding proteins were selected using a similar procedure. The 254 proteins containing sulphur in the HETATM records reduced to 53 families after resolution screening and clustering of which 36 sulphate binding proteins of ≤2·5 Å were selected for analysis

Abbreviations used for ligands. 2PG, 2-phospho-d-glyceric acid; ADP, adenosine di-phosphate; AMP, adenosine monophosphate; c-AMP, cyclic AMP; CAP, 2-carboxyarabinitol-1,5-biphosphate; CMC, carboxymethyl coenzyme A; DCG, cytidylyl 2',5'-guanosine; DHG, 2-dodecanoyl-amino-1-hexanol-phosphogylycol; F6P, fructose-6-phosphate; FAD, flavin adenine dinucleotide; FBP, fructose-1,6-biphosphate, G3P, glycerol-3-phosphate; GMP, guanosine monophosphate; GNP, guanosine-5'-(beta,gamma,imido)triphosphate, NAD,NDH nicotinamide adenine dinucleotide; NADP,NAP, nicotinamide adenine dinucleotide phosphate; PO_4, inorganic phosphate ion; PAL, *N*-(phosphonacetyl)-l-aspartate; PLP, pyridoxal phosphate; PTP, deoxythymidine 3'-5'-biphosphate; SO_4, sulphate ion; UMP, 2'-deoxyuridine 5'-monophosphate

phosphate and sulphate ions are sufficiently similar to allow the two groups to bind in locations with similar properties. Accordingly, all locations found to bind sulphate ions must also be viewed as potential phosphate binding sites. The exception to this view is when a site normally binds phosphate in its monobasic form as in phosphate binding protein (Luecke & Quiocho, 1990). Such binding proteins show a high degree of specificity, with sulphate unable to bind to the same site as phosphate. However, this specificity is unusual and many other binding sites have been shown to bind either anion. For example, the anion binding site in triosephosphate isomerase exploits the same residues for both sulphate and phosphate binding, but the details of the geometry are different for the two ions (Verlinde *et al.*, 1991).

Table 1 lists the phosphate and sulphate binding proteins selected for this study. The details of the analysis are explained in the legend. Of the 65 phosphate groups identified, 34 classed as helix-type, are within 7 Å of the amino terminus of an α-helix and with the exception of FAD in *p*-hydroxybenzoate hydroxylase (1phh; Schreuder *et al.*, 1988), all the ligands have at least one phosphate group within 5·2 Å of an amino terminus C^{α}.

Three phosphate groups appear within 5·6 Å of the end of a 3_{10}-helix. The hydrogen bonds stabilizing the 3_{10}-structure are weaker and the dipoles of the peptides are not aligned. This non-alignment prevents the occurrence of a large helix-dipole and thus phosphate binding by these structures would not be expected on the basis of the helix macrodipole theory. The role of the 3_{10}-helix in the observed examples appears to be to allow multiple backbone nitrogen atoms to bind the phosphate group. The ability of 3_{10}-helices to bind phosphate indicates that the stabilizing effect of the helix is limited to a favourable conformation of the polypeptide, such that the bound anion can interact with a number of individual peptide dipoles.

Phosphate groups were classified according to the number of distinct residues with which each group was in contact. This varied from none (two examples) to seven (one example) as illustrated in Figure 1a. Three of the four phosphate groups in contact with six or more residues are found at helix-type binding sites, but overall, there is no significant difference between the number of residues involved in binding a phosphate group at helix or non helix-type sites (data not shown). The most common number of residues in contact with a phosphate group is one, but there is not a strong preference for a particular number of residues. Neither is there any abrupt cutoff after which no more residues can be fitted around a phosphate group.

Figure 1b illustrates the number of phosphate groups with a given number of contacts on them. A contact is counted every time two atoms approach each other $\leqslant 3{\cdot}2$ Å. For example if a tyrosine hydroxyl group (OH) is within 3·2 Å of two phosphate oxygen atoms, two contacts are counted. This number will overestimate the number of possible hydrogen bonds available as one hydrogen atom will usually only form one hydrogen bond. There is a peak in the number of contacts at two per phosphate group but the spread of contacts is large, and the falling off does not occur very rapidly. This again leads to the conclusion that there is no critical number of contacts necessary to bind a phosphate group. The average number of contacts per phosphate group is 3·5 with a standard deviation of 2·3. This compares with the value of $5(\pm 3)$ given by Chakrabarti (1993) for inorganic phosphate and sulphate ions. The distribution of contacts for both subdivisions of helix and non-helix type sites is similar (data not shown). Two phosphate groups are not in contact with any atoms from the protein (1csc and 1rnd). In both examples, the groups are part of a larger molecule which is in contact. In 1rnd (a ribonuclease, Aguilar *et al.*, 1992) the phosphate is 3·5 Å away from an Arg residue and would thus be stabilised by a long range electrostatic interaction. Both phosphate groups are exposed on the surface of the protein and are likely to be extensively solvated. There is no striking correlation between the number of contacts on a phosphate group and the type of phosphate, i.e. if it is a free phosphate, a terminal phosphate or a phosphate connected to the ligand through two oxygen atoms (data not shown).

The number of each type of amino acid in contact with each phosphate group was recorded. In order to see the binding site from the level of the individual phosphate group, if a single amino acid was in contact with more than one phosphate group, it was counted twice. The total number of each amino acid binding phosphate is summarized in Table 2, and split by helix or non-helix type sites in Tables 3 and 4.

When there is no helix to stabilise the anion, arginine is the most commonly occurring residue that binds to phosphate groups. The guanidinium group is well suited to phosphate binding, since it is positively charged and can form multiple hydrogen bonds. Since the group is resonance stabilised, it is also a poor proton donor, and thus unlikely to hydrolyse phosphorylated intermediates.

Glycine is the most prevalent amino acid at helix-type phosphate binding sites even though Gly only has its main chain available for hydrogen bonding. Chakrabarti (1993) has suggested that the absence of a side-chain on Gly allows a large anion to approach the amide nitrogen. The abundance of Gly specifically at helix-type sites may be explained by the occurrence of glycine-rich loops, which are a frequently recurring motif in phosphate binding (Saraste *et al.*, 1990; Dreuscike & Schulz, 1986) for example in tryptophan synthase (1wsy; Hyde & Miles, 1990) and *ras* p21 oncogene protein (5p21; Pai *et al.*, 1990).

Of the 66 phosphate groups studied only six were found that were not at a helix binding site and also not in contact with a positively charged residue. Of these, two were in contact with no amino acids, two were close to a 3_{10}-helix and one was close to a calcium ion. The other was in a large cavity near the surface of the protein, and so would be well solvated. Of the helix-type set taken alone, 16 out of the 34 phosphate groups were not in contact with any positively charged residues. Accordingly there appears less need for a positively charged residue at the helix-type binding sites. This could be explained either by the α-helix dipole hypothesis or by the increased number of individual peptide dipoles available at such sites.

The three amino acids Phe, Pro and Leu are not represented at phosphate binding sites even though none of these amino acids are particularly rare (Phe 3·6%, Pro 4·36% and Leu 8·50%). Low abundances are also found for the amino acids Cys, Trp and Met (Cys 1·13%, Trp 1·27%, Met 2·45%). Thus, the under representation of amino acids such as Leu may be taken as an indication of poor phosphate binding ability. For amino acids that occur infrequently at binding sites, the mode of interaction is with the amide nitrogen. The data give no indication of a typical phosphate binding site. For example, it is not possible to say that a phosphate usually binds to an arginine, a serine and a glycine residue.

The large size of the phosphate group enables it to bridge gaps between chains and residues making contact to phosphates can be quite distantly removed from each other on the protein chain. The large size of some of the residues commonly involved in binding phosphate (e.g. Arg) means that residues that are distant on the sequence can be close in the native folded protein.

Each type of residue involved in phosphate binding was analysed, in order to detect preferences for binding with particular atoms (data not shown). Non-helix binding sites present a clear picture of phosphates being bound predominantly by the Arg guanidinium group and some hydroxyl groups from tyrosine. Helix-type binding sites show less specificity for the type of atom involved in the contact. Chakrabarti (1993) found a preference for histidine to bind using the distal nitrogen NE2. This is not

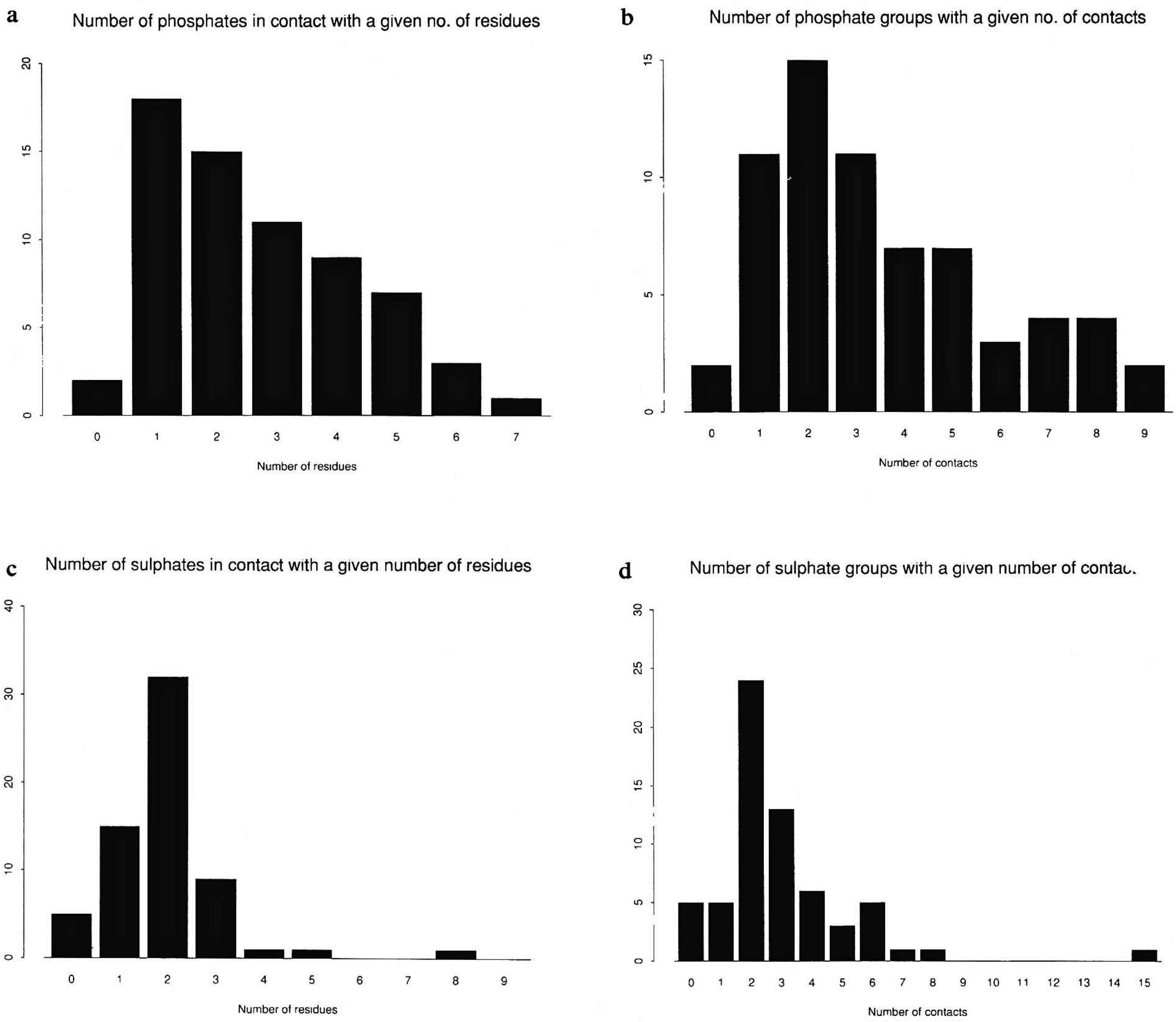

Figure 1. a, The number of phosphate groups in contact with a given number of amino acid residues. b, Number of phosphate groups with a given number of atomic contacts. c, Number of sulphate groups in contact with a given number of amino acid residues. d, Number of sulphate groups with a given number of atomic contacts.

The upper limit of 3·2 Å between ligand oxygen and protein O or N atoms chosen by Bass *et al.* (1992) and Noble *et al.* (1991) was adopted to identify possible hydrogen bonds (contacts). This compares with 3·1 Å for the sum of the Van der Waals radii of nitrogen and oxygen and 3·0 Å for oxygen with oxygen while in small molecule structures mean O· · ·O distances are of the order of 2·75 to 2·85 Å and mean O· · ·N distances 2·85 to 2·95 Å. No geometry check was performed since many of the interactions of interest were principally electrostatic and non-directional (i.e. with positively charged side-chains of His, Lys and Arg). Interactions with water molecules were not considered in this study. It requires good electron density maps to distinguish water molecules from noise and the presence of disordered side-chains on the surface may make it hard to distinguish solvent from alternative locations for side-chain atoms (Baker & Hubbard, 1984). In addition, not all PDB files contain solvent molecules (e.g. 1fnr, 8rub).

The co-ordinates of the atoms surrounding the target phosphorus atom were used to scan through the ATOM records of the PDB file and thus identify any contacts $\leqslant 3{\cdot}2$ Å to oxygen or nitrogen. The total number of amino acids interacting with the group, and the total number of contacts were recorded, as were all the atoms involved.

Secondary structure was defined by the method of Kabsch & Sander (1983), using the program DSSP. In phosphate binding sites that are thought to be stabilised by the helix dipole, the phosphate will be 3 to 5 Å from the end of the α-helix (Hol *et al.*, 1978). However, as the interaction is electrostatic, there is no clear cutoff distance. 7 Å was chosen as the cutoff for this study.

Table 2
All phosphate binding sites

Amino acid	Total no.	No. binding	Propensity
Arg	664	43	4·78
Gly	1063	35	2·43
Tyr	490	16	2·41
Ser	733	17	1·71
His	332	7	1·56
Thr	721	15	1·53
Lys	751	10	0·98
Cys	183	2	0·81
Asp	771	7	0·67
Asn	549	5	0·67
Gln	450	4	0·66
Trp	166	1	0·44
Ala	1210	7	0·43
Val	944	4	0·31
Met	320	1	0·23
Glu	810	2	0·18
Ile	752	1	0·10
Phe	474	0	0·00
Pro	569	0	0·00
Leu	1110	0	0·00

Phosphate binding propensities calculated from 66 phosphate groups identified in 38 proteins. Total no. the frequency of each amino acid type in the set of 38 proteins. No. binding the number of each amino acid type identified as binding to phosphate. Propensity binding propensity calculated according to the method of Chou & Fasman (1974), where propensity P is given by

$$P = \frac{\frac{N_b}{T_b}}{\frac{N_p}{T_p}},$$

where N_b = no. of particular amino acid at binding sites, N_p = no. of particular amino acid in proteins studied, T_b = total no. of amino acids at binding sites and T_p = total no. of amino acids in proteins studied. The percentage composition for each amino acid over the protein data set as a whole, was calculated using only those chains that actually bound phosphates (e g only the B chain of 1wsy was used) The number of amino acids at binding sites was calculated by examining the environment of each individual phosphate group. Thus, if one arginine residue was bound between 2 phosphate groups, it was counted twice as it was at the binding site for both phosphates.

Tables showing the detailed breakdown of atomic interactions in phosphate binding sites are available from the authors by anonymous ftp from geoff.biop.ox.ac.uk.

Table 3
Phosphates from helix-dipole binding sites

Amino acid	Total no.	No binding	Propensity
Gly	740	26	3·14
Arg	466	14	2·70
Thr	488	14	2·57
Ser	476	11	2·06
Lys	485	6	1·11
Gln	298	3	0·90
Cys	102	1	0·88
His	214	2	0·83
Tyr	327	3	0·82
Trp	116	1	0·77
Asn	358	3	0·75
Asp	522	4	0·68
Val	656	4	0·55
Ala	842	5	0·53
Met	226	1	0·40
Ile	528	1	0·17
Phe	308	0	0·00
Pro	371	0	0·00
Glu	581	0	0·00
Leu	743	0	0·00

Phosphate binding propensities calculated for helix-type binding sites for 34 phosphate groups from 21 proteins

supported by our results. We find seven examples of histidine residues involved in binding phosphate containing ligands. Of these three bind with the amide nitrogen, three with ND1 and two with NE2.

The proportion of contacts made to the amide nitrogen of a residue decreases with increasing bulk of the side-chain, irrespective of whether the side-chain itself makes a large number of contacts to phosphate groups. Thus residues with small side-chains such as serine and threonine have a similar number of phosphate contacts to both main chain and side-chain atoms (for Ser, 12 to the amide nitrogen and 14 to the side-chain oxygen), whereas contacts to Tyr are almost exclusively with its hydroxyl group (a total of 15 contacts to the hydroxyl group and one to the amide nitrogen) and only a small proportion of contacts to arginine are to its amide nitrogen.

The propensities of each amino acid for binding phosphate were calculated as described in the legend to Table 2. The ordering of the propensities for the non-helix type binding sites follows the broad trend of good hydrogen bond donors with positively charged side-chains down to non-polar residues with bulky hydrophobic side-chains. Not only do these residues have one suitable donor atom (the amide N), but the bulky side-chains also restrict the approach of a phosphate group.

The ordering of propensities for helix type binding sites follows the same broad trends as the non-helix propensities, with the notable exception of glycine, which is the best phosphate binding residue in helix-type sites. This is presumably a consequence of its frequency of occurrence at the amino terminus of

Table 4
Phosphates not from helix-dipole binding sites

Amino acid	Total no	No. binding	Propensity
Arg	356	29	7·34
Tyr	267	13	4·40
His	184	5	2·45
Gly	576	9	1·41
Ser	393	6	1·37
Lys	421	4	0·86
Cys	129	1	0·70
Asp	409	3	0·66
Asn	284	2	0·63
Glu	419	2	0·43
Gln	236	1	0·38
Ala	636	2	0·28
Thr	381	1	0·24
Trp	77	0	0·00
Met	162	0	0·00
Phe	308	0	0·00
Ile	403	0	0·00
Pro	312	0	0·00
Val	513	0	0·00
Leu	620	0	0·00

Phosphate binding propensities calculated for non helix-type binding sites for 32 phosphate groups from 17 proteins.

phosphate binding helices and the commonly occurring glycine-rich loop, at the end of such helices. Although the ordering of the amino acids is similar for both helix and non-helix sites, the propensities for both types of binding sites are different. The preferences for the top two amino acids Arg (7·34) and Tyr (4·4) are higher in the non-helix-type binding sites compared to 3·14 and 2·7 for Gly and Arg, respectively, at helix-type sites. In non-helix-type sites the propensities fall off much more rapidly than they do in helix type binding sites. This suggests that the major binding factor in the helix set is the helix dipole, or the individual peptides but not the specific residues involved in binding.

Both serine and tyrosine score well, suggesting that the presence of a positive charge on the side-chain of an amino acid is not a prerequisite for phosphate binding. For serine and threonine, small side-chains allow close approach of phosphates to the amide nitrogen, and so increase possibilities for hydrogen bonding. Thr has a much lower propensity for binding at non-helix type binding sites than helix-type binding sites. Tyr is a better hydrogen bond donor than either Ser or Thr, owing to the ability of the ring system to stabilize negative charge.

A surprising result of this analysis is the low propensity observed for lysine. Since this is one of the three positively charged residues, it might be expected to play a significant role in phosphate binding. One suggested explanation is that the side-chain atoms of lysine are sometimes omitted from PDB files due to poor electron density. However, in the PDB files analysed here, there are only ten lysine residues with incomplete side-chains, of which only two are potential phosphate binding residues. A possible physical reason for the low occurrence of lysine is that lysine may act as a proton donor and could hydrolyse phosphorylated intermediates unlike the resonance stabilised guanidinium group of arginine (Riordan *et al.*, 1977).

The large differences in propensities for each amino acid, particularly for non-helix type sites, raise the possibility of predicting possible phosphate binding sites in proteins of known sequence, but unknown structure. However, propensity based methods of predicting protein structure (e.g. Chou & Fasman, 1974) work by finding regions of consecutive amino acids with an increased propensity. Since residues in contact with phosphates are often distant on the sequence (although not in space) such methods cannot readily be applied. Furthermore, the most commonly occurring number of residues found in contact with a phosphate group is one.

When the sulphate groups were scanned for interaction with α-helices, using the same method as for phosphate groups, only ten examples were found out of a total of 64 sulphate groups. With so few examples, splitting the binding sites into these two types would give unreliable statistics. Accordingly, sulphate sites were considered as a single set.

The number of residues involved in binding sulphate differs markedly when compared to phosphate binding proteins (Figure 1c). Most commonly, sulphate groups are in contact with two amino acid residues and there is a very rapid fall off in the number of residues the sulphate groups are found to be in contact with. Only three sulphate groups are in contact with more than three residues. This reflects the observation that sulphate groups are normally bound to the surface of the protein while phosphate groups are often buried. This may be a reflection on the lack of a functional role for sulphate ions and the fact that formally they bear a smaller charge than phosphate groups. The three sulphate groups that bind to more than three residues are all present in sites which, under physiological conditions, would be occupied by a different anion. Thus in guanylate kinase (1gky; Stehle & Schulz, 1990) and adenylate kinase (1ak3; Diederichs & Schulz, 1990) the sulphate site would be occupied by a phosphate group, and in mandelate racemase (1mrm; Neidhart *et al.*, 1991) the site is thought to be occupied by the carboxyl group of mandelate in the Michaelis complex.

Table 5
All sulphate binding sites

Amino acid	Total no.	No. binding	Propensity
Arg	317	25	5·00
Ser	529	25	3·02
His	157	7	2·84
Lys	515	19	2·34
Tyr	241	7	1·83
Glu	434	7	1·02
Thr	471	7	0·94
Gly	669	9	0·85
Asn	312	4	0·82
Asp	428	5	0·73
Trp	119	1	0·54
Ile	400	2	0·33
Gln	253	1	0·25
Phe	265	1	0·24
Ala	761	1	0·08
Cys	120	0	0·00
Met	139	0	0·00
Pro	326	0	0·00
Val	566	0	0·00
Leu	651	0	0·00

Sulphate binding propensities calculated for non helix-type binding sites for 63 sulphate groups from 36 proteins.

A summary of the number of contacts made by the protein on each sulphate group is shown in Figure 1d. This shows a large number of sulphate groups making just two or three contacts to the protein. Again this is in contrast to phosphates where there is greater variability. Figure 1d shows one sulphate making 15 contacts. This is the sulphate group in mandelate racemase (1mrm; Neidhart *et al.*, 1991).

Propensities for binding were calculated as for phosphate binding (see legend to Table 1) and the results summarised in Table 5. Propensities for sulphate binding follow the same general trends as for phosphate binding, with polar and positively charged residues good at sulphate binding and residues with bulky hydrophobic side-chains being poor at sulphate binding. If phosphate binding is of functional importance to the protein, then one would expect the

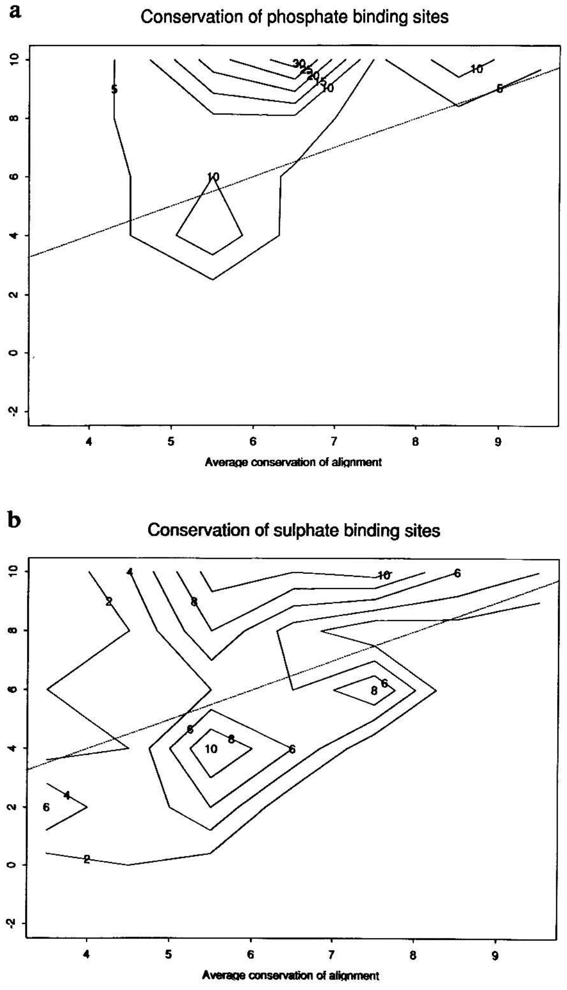

Figure 2. a, Contour plot of conservation at phosphate binding sites against average sequence conservation. Values shown above the dotted line indicate binding site residues that are more conserved than the average for the alignment. b, As for a, but sulphate binding sites.

In order to determine whether phosphate and sulphate binding residues are more highly conserved in aligned protein families, multiple sequence alignments were generated for each protein in this study. The PIR (Protein Identification Resource) sequence database (release 36) was scanned with each protein using a sensitive method (Barton, 1993) to identify all unequivocal family members. Multiple alignments were then generated using the Barton & Sternberg method (Barton & Sternberg, 1987).

From the multiple sequence alignments, the degree of conservation, *C*, at each position of the alignment was calculated using the method of Zvelebil *et al.* (1987), with gaps ignored. The average conservation value was calculated by summing the individual *C* values for each position and dividing by the total alignment length. Alignment positions corresponding to residues known to be involved in anion binding in one of the proteins in the alignment were extracted. Thus the relationship between the conservation at putative anion binding sites, and the alignments as a whole could be studied.

residues in the phosphate binding site to be conserved over the course of evolution. In order to evaluate the relationship between residue conservation and phosphate or sulphate binding we constructed multiple sequence alignments for each of the proteins shown in Table 1 (see legend to Figure 2 for details). The residue conservation at positions known to be anion binding was plotted against the average conservation for the alignment. The contour plots shown in Figure 2 show marked differences for phosphate (Figure 2a) and sulphate conservation (Figure 2b).

The principal conclusion is that residues involved in phosphate binding are more highly conserved relative to the alignment as a whole, than those involved in sulphate binding. This may be explained since while phosphate containing ligands are often functionally important, the sulphate anion is not. Accordingly, there will be no evolutionary pressure to conserve a sulphate binding site *per se*. The site may be conserved, but in such examples, the residues involved in binding may be serving some other structural or functional purpose. The presence of some highly conserved sulphate binding sites may be explained by the fact that *in vivo*, the site would be occupied by a phosphate group. This is true in guanylate kinase (1gky; Stehle & Schulz, 1990) and adenylate kinase (1ak3; Diederichs & Schulz, 1990) where a sulphate group is bound to the amino terminus of an α-helix. A phosphate in this position would be consistent with the kinase activity of these enzymes.

The degree of conservation at phosphate binding residues, is in general greater than the average conservation of the whole alignment irrespective of the degree of conservation of the alignment. This suggests that it is possible to predict phosphate binding sites by identifying highly conserved potential phosphate ligands in a multiple alignment. However, since our analysis suggests phosphate binding is insensitive to the number and type of residues involved, a residue which binds phosphate in one protein could mutate to a non-phosphate binding residue in a structurally homologous protein, without significantly impairing the strength of binding.

We thank Professor L. N. Johnson for encouragement and support. R.R.C. is supported by a Medical Research Council studentship and is a member of St. Catherine's College, Oxford. G.J.B. thanks the Royal Society for support. We thank Dr R. B. Rusell and Craig Livingstone for assistance with their programs.

References

Aguilar, C. F., Thomas, P. J., Mills, A., Moss, D. S. & Palmer, R. A. (1992). Newly observed binding mode in pancreatic ribonuclease. *J. Mol. Biol.* **224**, 265–267.

Baker, E. N. & Hubbard R. E. (1984). Hydrogen bonding in globular proteins. *Prog. Biophys. Mol. Biol.* **44**, 97–179.

Baker, P. J., Britton, K. L., Rice, D. W., Rob, A. & Stillman, T. J. (1992). Structural consequences of sequence patterns in the fingerprint region of the nucleotide binding fold. *J. Mol. Biol.* **228**, 662–671.

Barton, G. J. (1993). An efficient algorithm to locate all locally optimal alignments between two sequences allowing for gaps. *CABIOS*, **9**, 729–734.

Barton, G. J. & Sternberg, M. J. E. (1987). A strategy for the rapid multiple alignment of protein sequences: confidence levels from tertiary structure comparisons. *J. Mol. Biol.* **198**, 327 337.

Bass, M. B., Hopkins, D. F., Jaquysh, A. N. & Ornstein, R. L. (1992). A method for determining the positions of polar hydrogens added to a protein structure that maximizes protein hydrogen bonding. *Proteins*, **12**, 266 277.

Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Jr, E. F. M., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977). The Protein Data Bank, a computer based archival file for macromolecular structures. *J. Mol. Biol.* **112**, 535 542.

Branden, C. (1980). Relation between structure and function of α/β-proteins. *Quart. Rev. Biophys.* **13**, 317 338.

Chakrabarti, P. (1993). Anion binding sites in protein structures. *J. Mol. Biol.* **234**, 463 482.

Chou, P. Y. & Fasman, G. D. (1974). Conformational parameters for amino acids in helical, β-sheet, and random coil regions calculated from proteins. *Biochemistry*, **13**, 211 222.

Diederichs, K. & Schulz, G. E. (1990). Three dimensional structure of the complex between the mitochondrial matrix adenylate kinase and its substrate AMP. *Biochemistry*, **29**, 8138-8144.

Dreusike, D. & Schulz, G. E. (1986). The glycine-rich loop of adenylate kinase forms a giant anion hole. *FEBS Letters*, **208**, 301 304.

He, J. J. & Quiocho, F. A. (1993). Dominant role of local dipoles in stabilizing uncompensated charges on a sulphate sequestered in a periplasmic active transport protein *Protein Sci.* **2**, 1643-1647.

Hol, W. G. J. (1985). The role of the α-helix dipole in protein function and structure. *Prog. Biophys. Mol. Biol.* **45**, 149 195.

Hol, W. G. J., van Duijnen, P. T. & Berendsen, H. J. C. (1978). The α helix dipole and the properties of proteins. *Nature (London)*, **273**, 443-446.

Hyde, C. C. & Miles, E. W. (1990). The tryptophan synthase multienzyme complex. Exploring structure-function relationships with X-ray crystallography and mutagenesis. *Bio/technology*, **8**, 27-31.

Johnson, L. N. (1984). Enzyme-substrate interactions. In *Inclusion Compounds* (Atwood, J. L., Davies, J. E. D. & MacNicol, D. D., eds), vol. 3, pp. 529-541, Academic Press.

Kabsch, W. & Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577-2637.

Laskowski, R. A., MacArthur, M. W., Moss, D. S. & Thornton, J. M. (1993). PROCHECK: a program to check the stereochemical quality of protein structures. *J. Appl. Crystallogr.* **26**, 283-291.

Luecke, H. & Quiocho, F. A. (1990). High specificity of a phosphate transport protein determined by hydrogen bonds. *Nature (London)*, **347**, 402 406.

Morris, A. L., MacArthur, M. W., Hutchinson, E. G. & Thornton, J. M. (1992). Stereochemical quality of protein coordinates. *Proteins*, **12**, 345 364.

Neidhart, D. J., Howell, P. L., Petsko, G. A., Powers, V. M., Li, R., Kenyon, G. L. & Gerlt, J. A. (1991). Mechanism of the reaction catalyzed by mandelate racemase. 2. Crystal structure of mandelate racemase at 2·5 angstrom resolution: identification of the active site and possible catalytic residues. *Biochemistry*, **30**, 9264 9273.

Noble, M. E. M., Wierenga, R. K., Lambeir, A., Opperdoes, F. R., Thunissen, A. W. H., Kalk, K. H., Groendijk, H. & Hol, W. G. J. (1991). The adaptability of the active site of trypanosomal triosephosphate isomerase as observed in the crystal structures of three different complexes. *Proteins*, **10**, 50–59.

Pai, E F., Krengel, U., Petsko, G. A., Goody, R. S., Kabsch, W. & Wittinghofer, A. (1990). Refined crystal structure of the triphosphate conformation of h-ras p21 at 1·35 angstroms resolution: implications for the mechanism of GTP hydrolysis. *EMBO J.* **9**, 2351-2359.

Riordan, J. F., McElvany, K. D. & Borders, C. L., Jr (1977). Arginyl residues: anion recognition sites in enzymes. *Science*, **195**, 884-886.

Rossmann, M. G., Moras, D. & Olsen, K. W. (1974). Chemical and biological evolution of a nucleotide-binding protein. *Nature (London)*, **250**, 194 199.

Saraste, M., Sibbald, P. R. & Wittinghofer, A. (1990). The p-loop—a common motif in ATP- and GTP-binding proteins. *Trends Biochem. Sci.* **15**, 430–434.

Schreuder, H. A., van der Laan, J. M., Hol, W. G. J. & Drenth, J. (1988). Crystal structure of *p*-hydroxybenzoate hydroxylase complexed with its reaction product, 3,4-dihydroxybenzoate. *J. Mol. Biol.* **199**, 637 648.

Schulz, G. E. & Schirmer, R. H. (1979). *Principles of Protein Structure*. Springer-Verlag, Berlin.

Smith, T. F. & Waterman, M. S. (1981). Identification of common molecular subsequences. *J. Mol. Biol.* **147**, 195-197.

Stehle, T. & Schulz, G. E. (1990). Three-dimensional structure of the complex of guanylate kinase and its substrate GMP. *J. Mol. Biol.* **211**, 249-254.

Verlinde, C. L. M. J., Noble, M. E. M., Kalk, K. H., Groendijk, H., Wierenga, R. K. & Hol, W. J. (1991). Anion binding at the active site of trypanosomal triosephosphate isomerase. *Eur. J. Biochem.* **198**, 53-57.

Wierenga, R. K., De Maeyer, M. C. H. & Hol, G. J. H. (1985). Interaction of pyrophosphate moieties with α helixes in dinucleotide binding proteins. *Biochemistry*, **24**, 1346 1357.

Wierenga, R. K., Terpstra, P. & Hol, W. G. J. (1986). Prediction of the occurrence of the ADP-binding βαβ-fold in proteins, using an amino acid sequence fingerprint. *J. Mol. Biol.* **187**, 101-107.

Zvelebil, M. J., Barton, G. J., Taylor, W. R. & Sternberg, M. J. E. (1987). Prediction of protein secondary structure and active sites using the alignment of homologous sequences. *J. Mol. Biol.* **195**, 957-961.

Edited by F. Cohen

(Received 3 February 1994; accepted 15 July 1994)