HAD, a Data Bank of Heavy-Atom Binding Sites in Protein Crystals: a Resource for Use in Multiple Isomorphous Replacement and Anomalous Scattering

Suhail A. Islam,[a] David Carvin,[a,c] Michael J. E. Sternberg[a] and Tom L. Blundell[b,c]*

[a]*Biomolecular Modelling Laboratory, Imperial Cancer Research Fund, 44 Lincolns Inn Field, London WC2A 3PX, England,* [b]*Department of Biochemistry, University of Cambridge, 80 Tennis Court Road, Cambridge CB2 1GA, England, and* [c]*Department of Crystallography, Birkbeck College, Malet Street, London WC1E 7HX, England. E-mail: tom@cryst.bioc.cam.ac.uk*

Abstract

Information on the preparation and characterization of heavy-atom derivatives of protein crystals has been collected, either from the literature or directly from protein crystallographers, and assembled in the form of a heavy-atom data bank (HAD). The data bank contains coordinate data for the heavy-atom positions in a form that is compatible with the crystallographic data in the Brookhaven Protein Data Bank, together with a wealth of information on the crystallization conditions, the nature of the heavy-atom reagent and references to relevant publications. Some statistical information derived from the data bank, such as the most popular heavy-atom derivatives, is also included. The information can be directly accessed and should be useful to protein crystallographers seeking to improve their success in preparing heavy-atom derivatives for the methods of isomorphous replacement and anomalous dispersion. The World Wide Web address of HAD is http://www.icnet.uk/bmm/had.

1. Introduction

The method of multiple isomorphous replacement (MIR), first introduced by Perutz and co-workers in 1954 (Green *et al.*, 1954), and often enhanced by anomalous scattering (MIRAS) (see Blundell & Johnson, 1976, for a review) is still widely used in protein crystallography. Protein crystals comprise an open lattice of protein molecules with solvent occupying the channels and spaces which normally comprise between 30 and 80% of the crystal volume. The preparation of a heavy-atom derivative requires the binding of a heavy atom to a specific position, usually on the protein surface, for example by the displacement of a lighter solvent molecule or an ion, without distorting the protein or crystal lattice. Ideally rational selection of suitable heavy-atom reagents requires a comprehensive knowledge and understanding of the crystalline structure of the protein. Normally this information is unavailable as it is the objective of the crystal structure

analysis! Thus, the preparation of heavy atom derivatives has tended to remain an art.

Attempts to make chemically synthetic analogues of specific amino acids have included substituting selenium for sulfur residues or replacing an amino-terminal residue by an amino acid modified by a heavy-atom, but such chemical methods have not proved very useful. A very successful approach is to use site-directed mutagenesis to replace methionines by seleno-methionines (Hendrickson *et al.*, 1990) or more recently by teluro-methionines (Budisa *et al.*, 1997). However, recombinant approaches to replace amino acids have yet to provide a general method for introducing heavier atoms. Nevertheless, the sequence or function of a protein can give clues as to which heavy-atom reagents might be employed. The presence of a particular amino acid may suggest a covalent modification, for example the reaction of the sulfydryl groups of cysteine with mercury or tyrosines with iodine. The replacement of a metal ion cofactor, such as calcium or zinc, or the modification of a ligand by a heavy atom, can also give a useful derivative.

In many early studies the protein was covalently modified, purified and characterized before crystallization. However, pre-reaction of the protein often gives rise to conformational changes in the protein and crystallization occurs frequently in a different or non-isomorphous form. Most heavy-atom derivatives are produced by direct soaking of the crystals in a solution of the heavy-atom compound. However, with this approach heavy-atom substitution patterns tend to be complex, with sites frequently only partially occupied. Often the specificity is determined by entropic factors. Thus, sites between molecules in the crystal lattice, or between several different side chains brought together by the tertiary structure, may bind the metal ion even if the side chains individually do not have strong affinity for the metal.

In 1968 Blake (Blake, 1968) reviewed the data available for heavy-atom binding to proteins and suggested some generalizations. These were extended in a comprehensive review of protein heavy-atom derivatives (Blundell & Johnson, 1976; Blundell & Jenkins,

Table 1. *pH ranges of use for some heavy-atom reagents*

Min	Mean	Max	
3.0	6.5	9.1	Potassium tetrachloro platinate (II)
4.2	6.7	8.7	Potassium dicyano aurate (I)
4.2	6.8	9.5	Mercuric (II) chloride
2.8	6.6	8.5	Mercuric (II) acetate
4.7	6.6	9.3	4-Chloromercuribenesulfonic acid
2.0	6.4	9.3	Potassium tetra-iodo mercurate (II)
5.4	6.8	8.4	Thiomersal, ethyl mercury thiosalicylate (EMTS)
4.0	6.1	8.0	Potassium pentafluoro oxyuranate (VI)

1977) which analysed the dependence of reactivity on protein side-chain identity, nature of the reagent, pH, concentration, buffer *etc.* Over the past two decades there have been discussions of the binding of some particular metal ions, but there have been no comprehensive analyses. Furthermore, protein heavy-atom interactions have sometimes not been fully described in publications of protein crystallographic analyses and in any case the information has not been available in a format that could be used for systematic computer-based analysis.

We have now collected, either from the literature or directly from protein crystallographers, information on the preparation and characterization of heavy-atom derivatives of protein crystals. We have defined heavy atoms as those with atomic weight greater than rubidium. We have assembled the information in the form of a data bank (Carvin *et al.*, 1991), in which the coordinate data for the heavy-atom positions is compatible with the crystallographic data in the Brookhaven Protein Data Bank (Bernstein *et al.*, 1977). The heavy-atom data bank (HAD) contains a wealth of information and provides the basis for further, more detailed analyses of heavy-atom binding to proteins. The information can be directly accessed and should be useful to protein crystallographers seeking to improve their success in preparing heavy-atom derivatives for the methods of isomorphous replacement and anomalous dispersion. The World Wide Web (WWW) site is still not fully completed but can be accessed at http://www.icnet.uk/bmm/had.

2. Methods

2.1. *File systems*

Six file systems contain raw data. Each data file consists of a variable number of fields and each field is flagged by a four-character alpha code. This describes the nature of the information that may be deposited in each distinct field.

The conditions data file gives conditions for preparation of heavy-atom derivatives and information on the composition and concentration of the heavy-atom solution used in the experiments. This includes details of the chemical compound, precipitant, buffer, additives, pH, time of soak and source of protein. Additional techniques employed, such as variation of temperature, stabilization of the crystal by cross linking or mutagenesis of the primary structure, are described, as are the side chains of the protein involved at each heavy-atom binding site. An example is

```
HEAVY ATOM CONDITIONS DATA
REFS        20.00
BCOD        2AZA
PNAM        AZURIN (ALCALIGENES DENITRIFICANS)
HCOD        GHB
HCON        1.74
PPPT        AMMONIUM SULPHATE
PCON        -8.,75.,
BUFF        PHOSPHATE
BCON        0.10000
ACON        0.00000
SOAK        504.00000
PHPH        6.00000
TMAU        +3,2 MOLECULES
BIND        1A0.41/RES 42 CO:TRP 118 +3:
BIND        1B
BIND        2A0.11/RES 42 CO +3:TRP 118:
BIND        2B
BIND        3A0.06/RES 37 CO +3:TRP 118:
BIND        3B
```

The heavy-atom coordinates file contains the atomic coordinates and associated data as derived from the primary literature or as provided by personal communications. An example is

```
COMPND   3SGB
HEADER    10
REMARK
CRYST1  44.160 108.910 37.340 90.00 90.00 90.00 P21212  18
HET     PEC   DICHLORODIAMMINE PLATINUM (II)
FORMUL        1 PEC   PT1,N2,H6,CL2
FORMUL        2 PEC   PT1,[N1,H3]2,CL2 -0
HET     HBC   [3-[2(CARBOXYMETHOXY)BENZOYL] AMINO-2
FORMUL        1 HBC   C13,H17,O6,N1,HG1
FORMUL        2 HBC   [C6,H4]1,[(O1),(C1,H2),(C1,O2)]1,[(C1,O1),(N1,H1)
HET     AJA   POTASSIUM HEXACHLOROIRIDATE (III)
FORMUL        1 AJA   K3,IR1,CL6
FORMUL        2 AJA   IR1,CL6 -3
HETATM   1 PT1 PEC    1   0.8198 0.3490 0.7848
HETATM   2 PT2 PEC    2   0.8350 0.3345 0.6830
HETATM   3 PT3 PEC    3   0.8210 0.3648 0.7230
HETATM   4 PT4 PEC    4   0.8040 0.3515 0.8350
HETATM   5 HG1 HBC    5   0.6823 0.4886 0.0317
HETATM   6 HG2 HBC    6   0.3422 0.3624 0.2312
HETATM   7 HG3 HBC    7   0.4690 0.4080 0.4470
HETATM   8 IRA AJA    8   0.5000 0.5000 0.8085
HETATM   9 IR1 AJA    9   0.5000 0.5000 0.8065
HETATM  10 IR2 AJA   10   0.4658 0.3903 0.2342
TER     10    3SGB
```

The heavy-atom compound data file contains physical and chemical characteristics of each compound that has proved successful in past protein crystallographic analyses. This includes the IUPAC name, trivial name, molecular formula, oxidation state, solution chemistry and stereochemistry. To assist analysis an in-house three-character alphabet code was developed to

designate the heavy-atom compound (*i.e.* PEN = K_2PtCl_4).

HCOD	PEN
CNAM	POTASSIUM TETRACHLORO PLATINATE (II)
TNAM	POTASSIUM CHLOROPLATINITE
FORM	K2,PT1,CL4
OXYN	2 D8
STER	SPL
SOLC	PT1,CL4 -2
COOR	4

The reference data file contains literature citations including author(s), title, journal name, year of publication, volume number, first and last page number.

The multiderivative data file includes details of the composition and concentration of the two or more heavy-atom solutions used in making double and more complex derivatives.

The metalloprotein data files record information on conditions, including details of type, quantity, geometry and function of the metal cofactor(s) present, together with the procedure for metal cofactor substitution, including the composition and concentration of the reagent. It also records the interatomic distances and angles between the substituted heavy-atom and protein ligands.

REFS	210.1
BCOD	3PCY
PNAM	PLASTOCYANIN (MERCURY SUBSTITUTED)
NMET	COPPER
OMET	NONE
FUNC	MECHANISTIC (ELECTRON TRANSFER)
GEOM	DISTORTED TETRAHEDRAL
PLIG	HIS 37:HIS 87:CYS 84:MET 92:
SUBS	YES
PROS	CRYSTALLINE
MEBA	NONE
MEBC	MERCURIC ACETATE
HCOD	HDP
HCON	50.0
PPPT	AMMONIUM SULPHATE
PCON	3.3
BUFF	SODIUM PHOSPHATE
BCON	0.1
SOAK	23.0
SPEC	DIRECT DIFFUSSION:RETURNED TO FRESH STABILISER
SPEC	SOLUTION TO REMOVE EXCESS MERCURY REAGENT:
SPEC	SLIGHT ENLARGEMENT COORDS POLYHEDRON TO
SPEC	ACCOMMODATE MERCURY ATOM.
PHPH	6.0
TMAU	MONOMER
PD01	HG - HIS 37 (ND) - 2.34:
PD02	DOUBLE DERIVATIVE DATAHG - HIS 87 (ND) - 2.36:
PA01	HIS 37 (ND) - HG - HIS 87 (ND) - 100:
PA02	HIS 37 (ND) - HG - CYS 84 (SG) - 133:
RA01	CHURCH, W.B., GUSS, J.M., POTTER, J.J., AND FREEMAN,
RA01	H.C.
RJ01	J. BIOL. CHEM. (1986) 261, 234 - 237.

A second metalloprotein file describes the geometry of coordination of the metal cofactor and its protein ligands in the native protein.

There are also two file systems that contain processed data. These are sites containing geometrical details of heavy-atom sites and site coordinates containing atomic coordinates for the entire binding site *i.e.* protein residues making contact with the heavy atom.

2.2. *Method of analysis*

We have used a number of in-house computer programs to create, check and analyse the heavy-atom data bank, in addition to a relational database ORACLE (ORACLE corporation) and computer graphics. The principal programs carried out the following.

(*a*) Creation, maintenance and check of the data bank.

(*b*) Generation of the heavy-atom environment *i.e.* atomic coordinates for the protein and solvent interacting with the heavy atom (using predefined criteria for interatomic interactions). This was performed using symmetry operators so that the heavy-atom coordinates are appropriate to the asymmetric unit of the crystallographic cell used and all interactions are identified for the protein coordinates deposited in the Brookhaven Protein Data Bank.

(*c*) Preparation of data suitable for generation of relational database tables. A number of the file systems have been tabulated and placed in the relational database. The tabulated data can be made suitable for incorporation into most database systems.

3. Results and discussion

3.1. *Contents of the heavy-atom data bank*

The heavy-atom data bank (HAD) is a computer-based archival file system which contains experimental and derived information from successful multiple isomorphous replacement analyses in the determination of protein crystal structures. HAD is available *via* the WWW and in the form of a flat file system. The data bank makes information available which is otherwise only accessible in a fragmented form in the scientific literature or even unpublished in laboratory files. The data bank contains information about heavy-atom derivatives for 374 protein crystals, of which 176 are deposited in the Brookhaven Protein Data Bank. A further 600 proteins are being processed at present. The data bank contains information on the physical and chemical characteristics of each chemical compound that has proved successful in past protein crystallographic analyses: this includes the IUPAC name, trivial name, molecular formula, oxidation state, solution chemistry and stereochemistry. Experimental details of the preparation of the heavy-atom derivatives include the source of the protein, concentration of the heavy-atom solution, pH values, soak times and details of the buffer used in the experiments. The atomic coordinates are given in the same format as the PDB coordinates for the 5500 heavy-atom binding sites of the heavy atoms. A

statistical analysis is included for each of the 376 heavy-atom reagents; this includes range of pH values and a summary of the amino acids involved at the binding sites. For metalloproteins we give details on the details of the type, number, geometry of coordination and function of the native metal(s) present. This is followed by a description of the procedure for native-metal substitution and details of the coordination of the substituted heavy atom. We also include an extensive bibliography and references to other relevant WWW sites.

The information within HAD relates not only to proteins whose atomic coordinates have been deposited in the Brookhaven Protein Data Bank, but also to other proteins whose structures have yet to be deposited. The general scheme for the collation and categorization of HAD is shown in detail in Fig. 1.

3.2. Conditions of preparation of heavy-atom derivatives described in the data bank

The data bank records 2993 conditions of soak. This very large number of conditions reflects the wide range of buffers, salting-out agents, stabilities and solubilities of metal ions and pH of crystallization.
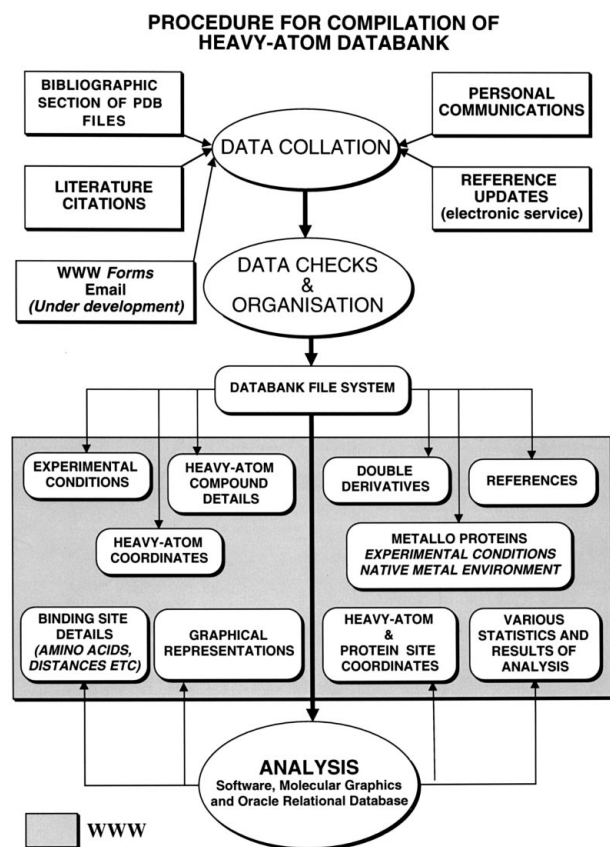

Fig. 1. Procedures for compilation of the heavy-atom data bank.

The pH has proved particularly important. For example, below pH 3.5 cations bind less well to aspartic and glutamic acids due to the protonation of the carboxylate groups. The nucleophilicity of histidine increases when it loses its proton around pH 6.0 to 7.0. Similarly the nucleophilicity of cysteine increases dramatically when the thiolate ion is formed at pH $\simeq$ 8.0. The thiolate ion is a stronger nucleophile than the thioether group of methionine, but when it becomes protonated it is considerably less effective. The attacking groups have the order

$$RS^- > R_2S > RSH.$$

Thus, the number and occupancy of sites can be manipulated by varing the pH, often after crosslinking the crystals to stabilize them. Extremes in pH can give rise to considerable difficulties in establishing suitable derivatives, as hydrogen and hydroxyl ions compete with the metal ion/complex for the protein and with the protein for the metal ion/complex. At extremely high pH values, metals in solution tend to form insoluble hydroxides. Thus, variation of the reactivity of amino-acid side chains by manipulation of the pH can enable the same heavy-atom ion/complex to bind at different sites, so producing more than one derivative useful for phase determination. From the data bank we can find the range of pH where each heavy atom has proved successful; a sample of a much more extensive table is included in Table 1.

Components present in the derivatization solution can also have a profound effect on protein heavy-atom interactions. The precipitant and buffer are the principal source of alternative ligands for the heavy-atom reagents, whilst protons compete with the heavy-atom ion/complex for the reactive amino-acid side chains. For example, ammonium sulfate is the most successful precipitant in protein crystallization experiments, but its continued presence in the mother liquor can cause problems by interfering with protein heavy-atom inter-actions. At high hydrogen ion concentrations the NH_3 group is protonated (i.e., NH_4^+), but as the pH rises the proton is lost, typically around pH 6.0–7.0, enabling the group to compete with the protein for the heavy-atom reagent. For example, the anionic complex $PtCl_4^{2-}$ in excess ammonia at pH > 7.0 will react:

$$PtCl_4^{2-} \rightarrow cisPtCl_2(NH_3)_2 \rightarrow Pt(NH_3)_4^{2+}.$$
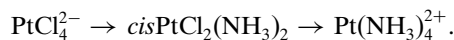
The resultant cationic complex is less susceptible to reaction because of the trans-effect of NH_3 (Petsko et al., 1978). Pd, Au, Ag and Hg complexes react in a similar way. Decreasing the pH of the solution reduces the amount of free ammonia available through protonation (Sigler & Blow, 1965). Such a technique may give rise to other problems (i.e., cracked crystal, decrease nucleo-philicity of the protein ligands). The data bank allows one to investigate the conditions that are best suited to such heavy-atom reagents.

Other important conditions can be investigated using the heavy-atom data bank. These include concentration of reagent, length of soak and temperature.

3.3. Analysis of heavy atoms and their binding sites in the data bank

The data bank records 42 different elements that have been used as heavy atoms by protein crystallographers. The most popular heavy-atom reagents are given in Table 2(a). These include uranyl, platinum, mercury, lead and gold. For any heavy-atom site the location in the protein can be displayed easily using the data bank, either as a position in the whole protein represented in terms of its elements of secondary structure or in terms of its detailed atomic coordinates (see Fig. 2).

Uranium reagents are amongst the most popular A metals; the five top, all uranyl compounds, are given in Table 2(b). UO_2^{2+} is a linear, covalent group based on uranium (VI), the most stable oxidation state of
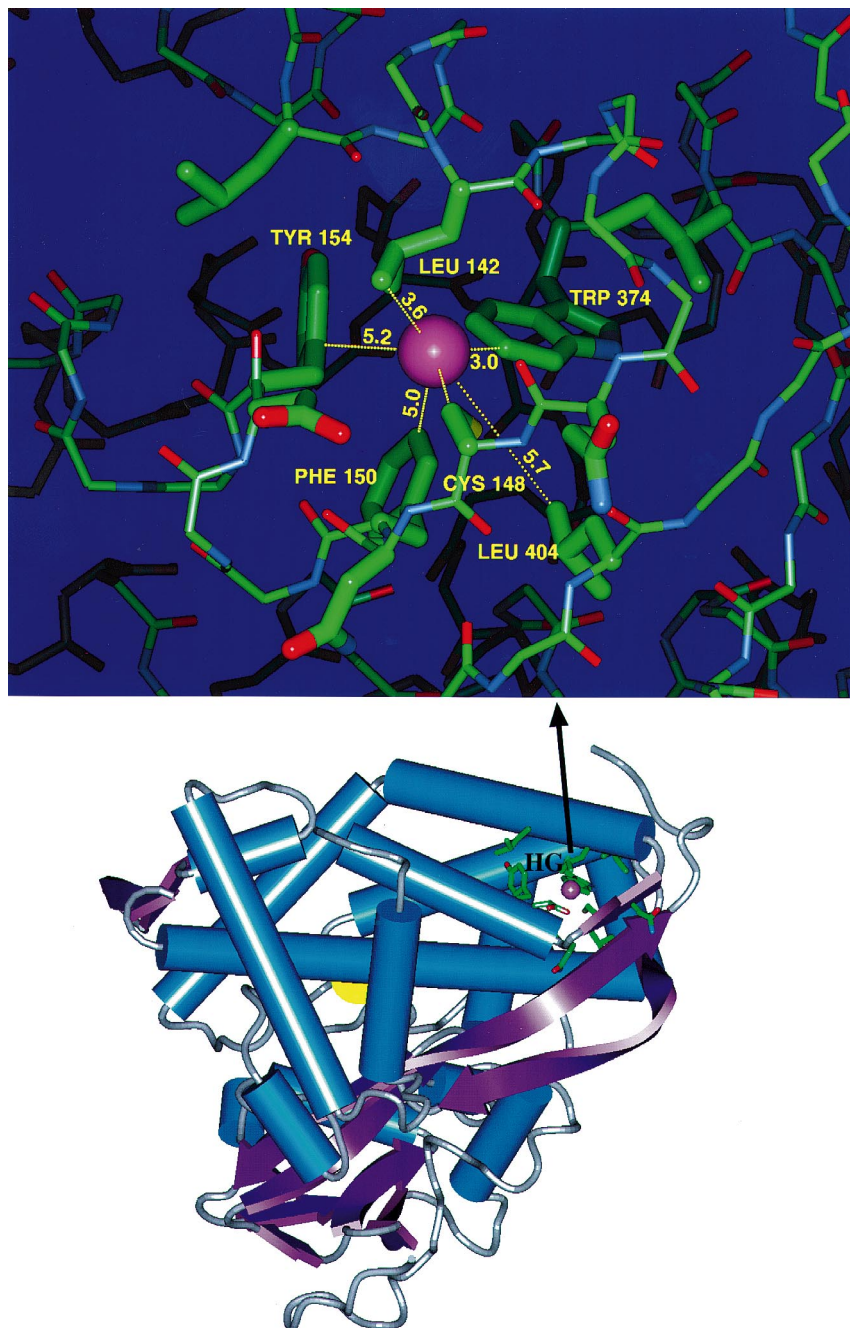

Fig. 2. The binding site for methyl mercury chloride in cytochrome P450. Such binding sites can easily be displayed as part of the whole protein represented by ribbons (β-strands) and cylinders (α-helices) or as a local binding site.

Table 2. *The most popular heavy-atom reagents*

The first number is the number of times the reagent has been used and the second is its precentage use.

(*a*) The most commonly used heavy-atom reagents

277	9	PT	Potassium tetrachloro platinate (II)
106	4	AU	Potassium dicyano aurate (I)
95	3	U	Uranyl acetate
95	3	HG	Mercuric (II) chloride
93	3	HG	Mercuric (II) acetate
81	3	HG	*Para* chloromercuri benzene sulfonate (PCMBS)
79	3	HG	Potassium tetraido mercurate (II)
79	3	HG	Thiomersal, ethyl mercury thiosalicylate (EMTS)
71	2	U	Potassium pentafluoro oxyuranate (VI)
68	2	HG	HG Phosphato [tris(ethylmercury)]
63	2	PB	Trimethyl lead (IV) acetate
59	2	PT	Potassium tetranitrito platinum (II)
58	2	U	Uranyl nitrate
56	2	PT	Dichlorodiammine platinum (II)
53	2	PT	Potassium tetracyano platinate (II)
50	2	PT	Potassium hexachloro platinate (IV)
45	2	HG	*Para* chloromercury benzoate (PCMB)
44	1	HG	Methyl mercury chloride
38	1	HG HBC	Mersalyl
37	1	PB DPC	Lead (II) acetate
37	1	AU GHF	Potassium gold chloride
36	1	HG HAA	Methylmercuriacetate
31	1	PT PEF	Dichloroethylene diamine platinum (II)

(*b*) The five most popular platinum, mercury and uranium derivatives

(i) Platinum

277	9	PT PEN	Potassium tetrachloro platinate (II)
59	2	PT PEO	Potassium tetranitrito platinum (II)
56	2	PT PEC	Dichlorodiammine platinum (II)
53	2	PT PEP	Potassium tetracyano platinate (II)
50	2	PT PFD	Potassium hexachloro platinate (IV)

(ii) Mercury

95	3	HG HDO	Mercuric (II) chloride
93	3	HG HDP	Mercury (II) acetate
81	3	HG HBA	*Para* chloromercuribenzene sulfonate (PCMBS)
79	3	HG HDF	Potassium tetra iodo mercurate (II)
79	3	HG HAZ	Thiomersal, ethyl mercury thiosalicylate (EMTS)

(iii) Uranium

95	3	U UXC	Uranyl acetate
71	2	U UXG	Potassium pentafluoro oxyuranate (VI)
58	2	U UXB	Uranyl nitrate
4	0	U UXJ	Sodium triacetate oxyuranate (VI)
3	0	U UXK	Uranium (VI) oxyoxalate

uranium. The data base shows that uranyl compounds may show 2+4, 2+5, or 2+6 coordination, with ligands lying in or near a plane normal to the O—U—O axis. In the heavy-atom compounds these equatorial ligands may be neutral (*i.e.*, H_2O) or anionic (*i.e.*, NO_3^-, CH_3COO^-, F^-, Cl^- or NO_2^-); in the protein they are most likely substituted by carboxylates at the C terminus or side chains of glutamate or aspartate, as shown in Fig. 3(*a*). Quite often entropic factors introduce unexpected

ligands, such as the lysine in Fig. 3(*a*). The data bank indicates that, at low pH, uranyl groups are often located near the hydroxyl groups of threonine and serine.

The data bank shows that, amongst the A metals, lanthanide ions have greater selectivity than the uranyl ion, which often forms clusters on the protein surface. It also shows that thallium and lead can provide useful derivatives, especially in their lower oxidation states, Tl (I) and Pb (II), when they resemble class A metals.

The most useful members of the B-metal group, platinum, gold and mercury, give rise to an extensive range of heavy-atom compounds, which form covalent, electrostatic and van der Waals complexes with proteins. Some compounds can bind to the protein molecule in different ways, for example, $PtCl_4^{2-}$ can bind either covalently to the thioether group of methionine, or electrostatically with positively charged residues.

The most popular mercury compounds are given in Table 2(*b*). Their use is mainly due to the ease of formation of covalent bonds with cysteine residues; an example is given in Fig. 2. Four of the most popular Hg^{2+} complexes are two coordinate. The mercuric chloride and acetate tend to be the most reactive. The covalent character in Hg—L bonds, especially in the two-coordinate complexes, can cause solubility problems in aqueous solutions. However, an excess of an alkali metal salt (*i.e.*, $HgX_2 + 2KI \rightarrow K_2HgX_4$) will often convert the compound to a more soluble anionic complex of the type HgX_4^{2-}, where $X = Cl^-$, Br^-, I^-, SCN^-, NCS^-, CN^-, SO_4^{2-}, oxalate^{2-}, NO_3^-, NO_2^-. This is probably the reason why HgI_4^{2-} occurs in the most popular list. However, the success of linear covalent compounds is reflected by the presence in the most popular compound list of *para*-chloromercuribenzene sulfonate (PCMBS) and ethyl-mercury thiosalicylate (EMTS). The aromatic ring and the ethyl group both prefer some hydrophobic site in the protein, but PCMBS requires an ionic interaction also. In this way the reactivity and location of different cysteines can be explored. Indeed the data bank shows that variation in the charge on the aromatic groups of organo-mercurials can give rise to different substitution patterns.

The class-B metals, platinum and gold, have proved very useful in making heavy-atom derivatives as shown by Table 2. They form stable covalent complexes with soft ligands such as chloride, bromide, iodide, ammonia, imidazole and sulfur groups. The stereochemistry of their complexes depends on the number of *d* electrons present. For instance the d^{10} ion of Au(I) gives a linear coordination of two [*i.e.*, $Au(CN)_2^-$], whereas d^8 ions of Pt(II) and Au(III) are predominantly square planar, giving cationic [*i.e.*, $Pt(NH_3)_4^{2+}$], anionic [*i.e.*, $Au(CN)_4^-$ and $PtCl_4^{2-}$] or neutral [*i.e.*, $Pt(NH_3)_2Cl_2$] complexes. These may accept an additional ligand to give a square pyramidal or two ligands to give octahedral coordination. The additional ligands are normally more weakly bound. Platinum (IV) has a d^6 configuration and forms
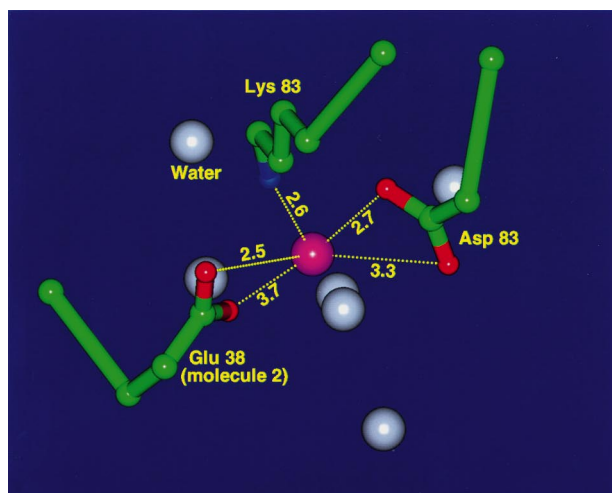
(a)
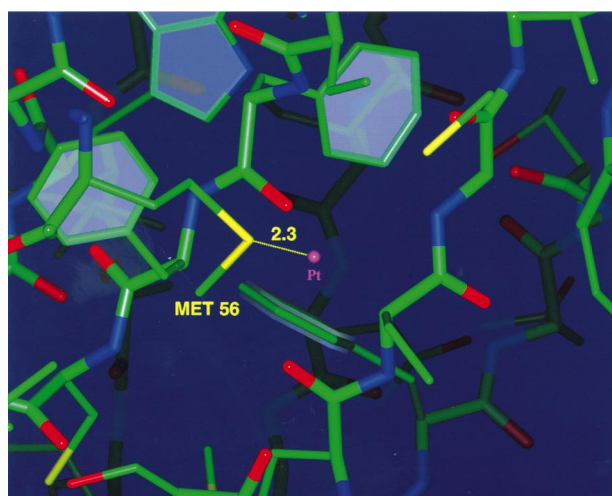
(b)

(c)

Fig. 3. Typical binding sites for (a) uranyl derivative, (b) $PtCl_4^-$ and (c) $PtCN_4^-$.

stable octahedral complexes such as $PtCl_6^{2-}$ with six equivalent covalent bound ligands.

$PtCl_4^{2-}$ remains by far the most successful heavy-atom reagent (Table 1). It generally reacts covalently with methionines as illustrated in Fig. 3(b); but the data bank shows that other polar and hydrophobic groups, often phenylalanines, can stabilize the complex. The data bank confirms the observation by Petsko et al. (1978) that the kinetic and thermodynamic stability of these complexes depends on the protein ligands, buffer, pH and salting-in/out agent (see above).

Positively charged groups of proteins, such as the α-amino terminus, ε-amino of lysine, guanidinium of arginine and imadazolium of histidine may form ion pairs with heavy-atom anionic complexes. For example, HgI_4^{2-} and HgI_3^- can bind through electrostatic interactions. Anionic metal cyanide complexes tend to be more resistant to substitution and consequently interact electrostatically on most occasions. For example, $Pt(CN)_4^{2-}$ binds at several sites involving lysines and arginines in proteins; an example is given in Fig. 3c. $Pt(CN)_4^{2-}$ and $Au(CN)_2^-$ can also act as inhibitors by binding at coenzyme phosphate sites.

As many heavy-atom reagents are hydrophilic, most interactions occur at the protein surface. However, substitution, addition or removal of non heavy-atom component(s) of the derivatization reagent can alter the hydrophilic hydrophobic balance and lead to penetration of the core. For example, anionic complexes such as $HgCl_4^{2-}$ and $PbCl_6^{2-}$ are hydrophilic and would not normally enter the protein core, although organometallics such as RHgCl and R_3PbCl (R = aliphatic or aromatic) are much more hydrophobic and can do so. We have already seen that hydrophobic organomercury compounds have proved very successful heavy-atom reagents. Inert gases, first used in the analysis of myoglobin by Schoenborn et al. (1965), are now proving to be a very useful alternative (Schiltz, 1997).

The structure determination of large multicomponent systems such as the 50S ribosomal subunit (Yonath et al., 1986) or the nucleosome core particle (O'Halloran et al., 1987) requires the addition of reagents with a greater number of electrons, preferably in a compact polynuclear structure. Polynuclear reagents should preferably be covalently bound to one or a few specific sites, either first in solution or later in the crystals. Spacers of differing length can be inserted into the reagent to increase accessibility. Tetrakis (acetoxy-mercury) methane (TAMM) and di-m-iodo-bis-ethylenediamine-di-platinum (II) nitrate (PIP) have better solubility in aqueous solutions than other polynuclear heavy-atom compounds. Cluster and multimetal reagents that have been successfully employed in protein structural determinations have been reviewed by Thygesen et al. (1996)

Metal ion cofactors can sometimes be displaced by dialysis or diffusion by a heavy-atom solution, but usually the cofactor is removed first by a chelating agent

(*i.e.* EDTA) or by acidification. This is best carried out on the crystals. Alternatively the metal can be substituted by biosynthesis of the metalloprotein under enriched conditions of the substituting metal, an approach which has been successful in displacing zinc with cobalt and other lighter metals

The data bank confirms that metal ions are best substituted by a metal of similar character and radius. Thus, calcium is an A-metal and prefers ligands containing O atoms that may originate from carboxylic, carboxyamide, hydroxyl, main-chain carbonyl groups and water molecules. Divalent alkaline earth metal ions (*i.e.*, Sr^{2+}, Ba^{2+}) or trivalent lanthanide ions can bind at calcium sites but can give very different coordination geometry and stability. Nd^{3+} and Sm^{3+} can displace some Ca^{2+} ions with negligible change in structure. On the other hand zinc has a relatively small ionic radius and is more polarizing. Structural Zn atoms are often tetrahedrally coordinated by cysteine residues, while those at active sites frequently bind histidine, often in association with a water molecule and/or carboxylate ligands. The data bank shows that cadmium or mercury can replace zinc but often with a conformational change leading to lack of isomorphism.

3.4. *Use of the data bank*

The data bank is probably best exploited by first investigating the most commonly used heavy-atom reagents with a view to obtaining mercury, platinum and uranyl derivatives that tend to bind at different sites. The most common reagents (Table 2) can first be selected and tested for suitability in terms of amino-acid sequence, pH, buffer and salt. If there are many sulfydryls several mercurials might be exploited or if several methionines, other platinum agents might be investigated. A high pH would argue against use of some A metals due to insolubility of hydroxides; the presence of ammonium sulfate would argue for as low a pH as possible. The presence of citrate would imply changing the buffer for acetate if A-metals such as uranyl or lanthanides were to be used.

For each heavy-atom agent the conditions of its previous use can be checked against the conditions of crystallization in the current study. Conversely the data bank can be interrogated for reagents that have been used in similar conditions. In each case derivatives that maximize the variety of ligands can be exploited.

The time of soak should be first set according to previous experience indicated in the data bank. However, the progress of derivatization needs to be monitored by checking for change of colour, transparency or cracking. If cracking and disruption of the

crystals occurs quickly, a less reactive reagent can be tried, and conversely if substitution is insufficient a more reactive reagent can be tried. If there are several cysteines, different derivatives can be obtained with mercurials of different size and hydrophobicity. In each circumstance the data bank should provide useful information to assist decisions about the choice of reagents.

Please keep information about the heavy-atom binding sites and the heavy-atom structure-factor amplitudes. These data and other relevant information should be submitted to the Protein Data Bank.

References

Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer Jr, E. F., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977). *J. Mol. Biol.* **112**, 534–552.

Blake, C. C. F. (1968). *The Preparation of Isomorphous Derivatives*. In *Advances In Protein Chemistry*, Vol. 23, pp. 59–120.

Blundell, T. L. & Jenkins, J. A. (1977). *Chem. Soc. Rev. (London)*, **6**, 139–171.

Blundell, T. L. & Johnson, L. N. (1976). *Protein Crystallography*. New York: Academic Press.

Budisa, N., Karnbrock, W., Steinbacher, S., Humm, A., Prade, L., Neuefeind, T., Moroder, L. & Huber, R. (1997). *J. Mol. Biol.* **271**, 1–8.

Carvin, D. G. A., Islam, S. A., Sternberg, M. J. E. & Blundell, T. L. (1991). *Isomorphous Replacement and Anomalous Scattering*. Warrington: Daresbury Laboratory

Green, D. W., Ingram, V. M. & Perutz, M. F. (1954). *Proc. R. Soc. London Ser. A*, **225**, 287–307.

Hendrickson, W. A., Horton, J. R. & Lemaster, D. M. (1990). *EMBO J.* **9**, 1665–1672.

O'Halloran, T. V., Lippard, S. J., Richmond, T. J. & Klug, A. (1987). *J. Mol. Biol.* **194**, 705–712.

Petsko, G. A., Phillips, D. C., Williams, R. J. P. & Wilson, I. A. (1978). *J. Mol. Biol.* **120**, 345–359.

Schiltz, M. (1997). *Xenon at LURE*. http://www.lure.u-psud.fr/lure/sections/XENON/xenon_eng.html.

Schoenborn, B. P., Watson, H. C. & Kendrew, J. C. (1965). *Nature (London)*, **207**, 28–30.

Sigler, P. B. & Blow, D. M. (1965). *J. Mol. Biol.* **14**, 640–644.

Thygesen, J., Weinstein, S., Franceschi, F. & Yonath, A. (1996). *Structure*, **4**, 513–518.

Yonath, A., Saper, M. A., Makowski, I., Mussig, J., Piefke, J., Bartunik, H. D., Bartels, K. S. & Wittmann, H. G. (1986). *J. Mol. Biol.* **187**, 633–636.