Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Heavy-atom Database System: a tool for the preparation of heavy-atom derivatives of protein crystals based on amino-acid sequence and crystallization conditions

Michihiro Sugahara,[a] Yukuhiko Asada,[a] Haruhiko Ayama,[a] Hisashi Ukawa,[b] Hideyuki Taka[b] and Naoki Kunishima[a]*

[a]Advanced Protein Crystallography Research Group, RIKEN Harima Institute at SPring-8, 1-1-1 Kouto, Mikazuki-cho, Sayo-gun, Hyogo 679-5148, Japan, and [b]Hitachi Software Engineering Co. Ltd, 4-12-7 Higashishinagawa, Shinagawa-ku, Tokyo 140-0002, Japan

Correspondence e-mail: kunisima@spring8.or.jp

Received 14 April 2005
Accepted 22 June 2005

Heavy-atom Database System (*HATODAS*) is a WWW-based tool designed to assist the heavy-atom derivatization of proteins. The conventional procedure for the preparation of derivatives is usually a time-consuming 'trial-and-error' process. The present program provides a solution for this problem using a database of known heavy-atom derivatives. A database search suggests potential heavy-atom reagents for any target protein based on its amino-acid sequence and crystallization conditions. A mining of the database identified 93 preferred motifs for heavy-atom binding. The motifs are observed frequently at the actual heavy-atom-binding sites encountered in the process of structure determination.

1. Introduction

The October 2004 release of the Protein Data Bank (http://www.rcsb.org/pdb/) contains 22 881 X-ray structures and 3402 NMR structures. Although the proportion of NMR structures has increased in the recent years, X-ray structures are still predominant. One of the reasons for the recent remarkable increase in the number of deposited structures is the promotion of several structural genomics initiatives. RIKEN Structural Genomics Initiative (RSGI; http://www.rsgi.riken.go.jp/) is one of the major structural genomics projects in Japan (Yokoyama *et al.*, 2000). The Advanced Protein Crystallography Research Group (formerly the Highthroughput Factory) of RIKEN Harima Institute promotes high-throughput X-ray crystallography using synchrotron radiation at SPring-8, Japan. In the RIKEN Harima Institute, several systems for rapid structural determination have recently been developed, including the automated crystallization robot TERA (Sugahara & Miyano, 2002) and the SPring-8 Precise Automatic Cryo-sample Exchanger (SPACE) at the RIKEN structural genomics beamline BL26B2 (Ueno *et al.*, 2004). In 2003, 46 protein crystal structures were solved by the Advanced Protein Crystallography Research Group. Of these, 22 structures (48%) were determined by experimental phasing from heavy-atom derivative crystals. This indicates that in one out of every two cases the technically easier molecular-replacement method, which requires a known probe structure sharing a high amino-acid identity of 30% or more with the target protein, could not be used for phasing. Even if phasing by the molecular-replacement method is successful in such cases, the model-rebuilding process is time-consuming and automated model-building programs such as *ARP*/*wARP* (Morris *et al.*, 2003) and *RESOLVE* (Terwilliger, 2000) may not be useful. Although the multiwavelength anomalous diffraction method using selenomethionyl proteins (Hendrickson *et al.*, 1990) has become a powerful tool for structure determination, it requires synchrotron radiation, which is not always fully accessible to most researchers. The expression of selenomethionyl protein using conventional expression systems is sometimes restrained owing to its toxicity to the host organisms. Furthermore, selenomethionyl protein does not necessarily crystallize under the same conditions as the native protein and sometimes leads to crystals that are non-isomorphous. In such cases, the conventional methods of preparation of heavy-atom derivatives such as soaking or cocrystallization with heavy-atom reagents are still worth trying for rapid and effective determination of crystal structures. However, the preparation of a heavy-atom derivative is usually a 'trial-and-error' process. Prediction of feasible heavy-atom reagents based on knowledge of protein/inorganic chemistry is an effective approach to reduce unnecessary trials, as described for example in the

doi:10.1107/S0907444905019670

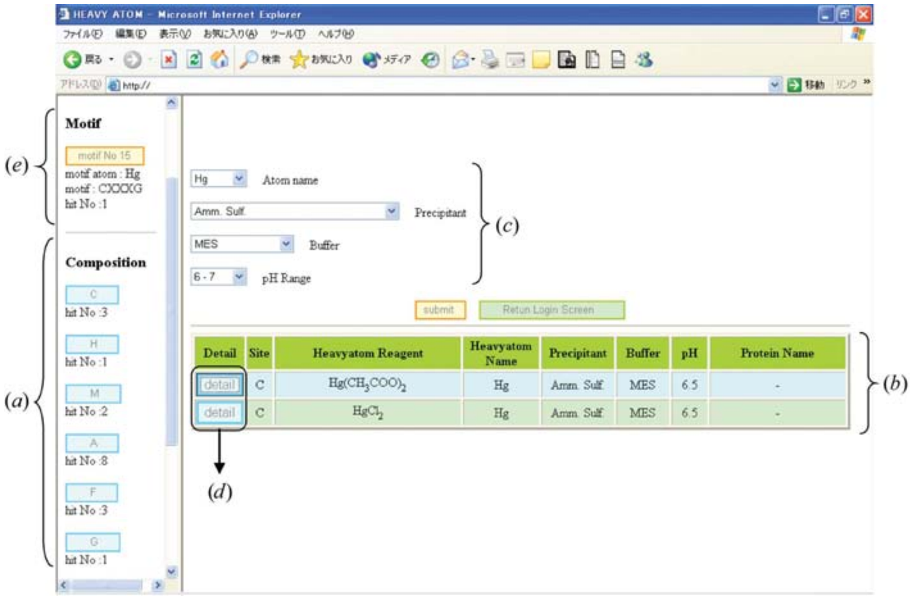
Table 1
The number of deposits of heavy-atom bound residues in the heavy-atom database.

	Hg	Au	Ag	Pt	Ir	Pd	Os	Pb	Tl	La	Sm	Eu	Gd	Dy	Er	Lu	U	Zn	Co	Br	Kr	Xe
Ala	2	—	—	2	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	1	—	—
Val	4	1	—	—	—	—	—	—	—	—	—	—	—	—	—	—	2	—	—	1	—	1
Leu	11	—	—	2	—	—	—	—	—	—	—	—	—	—	—	—	2	—	—	1	5	5
Ile	3	—	—	1	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	1	3
Met	18	1	—	36	—	1	1	—	—	—	—	—	—	—	—	—	—	—	—	1	—	2
Trp	3	—	—	1	1	—	—	—	—	—	—	—	—	—	—	—	—	—	—	2	—	—
Phe	4	1	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	1	1	3
Pro	2	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	4	—	—
Gly	2	—	—	—	—	—	—	—	—	—	1	—	—	—	—	—	1	—	—	—	—	—
Ser	5	1	2	—	—	—	—	—	1	—	—	—	—	—	—	—	—	—	—	17	1	—
Thr	12	2	—	2	—	—	—	—	1	—	—	—	—	—	—	—	2	—	—	7	—	—
Cys	142	15	6	4	—	1	—	1	—	—	—	—	—	—	—	—	—	—	—	—	—	—
Tyr	15	—	—	1	—	1	—	1	2	—	1	—	—	—	—	—	2	—	—	6	—	—
Asn	15	—	—		—1	2	—	1	2	—	4	—	—	—	—	—	1	—	—	12	—	—
Gln	13	—	—	1	1	—	—	1	—	—	1	—	—	—	—	—	—	—	—	7	1	1
Lys	14	2	1	5	1	2	1	2	—	—	—	—	—	—	—	—	2	—	—	6	—	—
His	52	9	4	23	2	3	1	3	1	—	1	—	—	—	—	—	—	1	1	5	—	—
Arg	9	4		2	1	3	2	—	—	—	1	—	1	—	—	—	—	—	—	15	1	—
Asp	12	2	1	1	—	—	1	7	2	1	16	3	3	—	1	6	6	—	3	2	—	—
Glu	13	2	—	2	—	1	1	13	3	—	16	—	9	1	—	2	10	1	1	3	—	—
Total	351	40	14	83	7	14	7	29	12	1	41	3	13	1	1	8	28	2	5	91	10	15

textbook by Blundell & Johnson (1976). However, prediction by a manual search of literature is laborious and time-consuming. A WWW-based tool *Heavy-Atom Data Bank* (*HAD*) has been reported which provides information on the crystallization conditions and the heavy-atom-binding sites (Islam *et al.*, 1998). However, the *HAD* suite is not suitable for direct and automatic identification of heavy-atom reagents. Here, we present the *Heavy-atom Database System* (*HATODAS*), which is the first program to provide key information on feasible reagents using a database of known heavy-atom derivatives.

2. The program

Currently, our heavy-atom database comprises information on 776 heavy-atom-binding sites from the PDB, from the textbook on protein crystallography by Blundell & Johnson (1976) and from the experimentally phased protein crystal structures determined at our Advanced Protein Crystallography Research Group (Table 1). The contents of the database are heavy-atom-binding residues, heavy-atom names, heavy-atom reagents, concentrations of heavy-atom reagents, preparation methods (cocrystallization or soaking), soaking times, compositions of harvest solutions, crystallization conditions (precipitant, additive reagent, buffer name, buffer concentration and pH), PDB codes, protein names and reference literature. In the process of the database construction, we found that some heavy atoms preferentially bind to certain amino-acid residues (Table 1). As pointed out by Blundell & Johnson (1976), Hg tends to bind Cys or His, Pt binds to Met or His and Pb



Motif

motif No 15
motif atom : Hg
motif : CXXXG
hit No :1

Composition

C
hit No :3

H
hit No :1

M
hit No :2

A
hit No :8

F
hit No :3

G
hit No :1

Hg Atom name

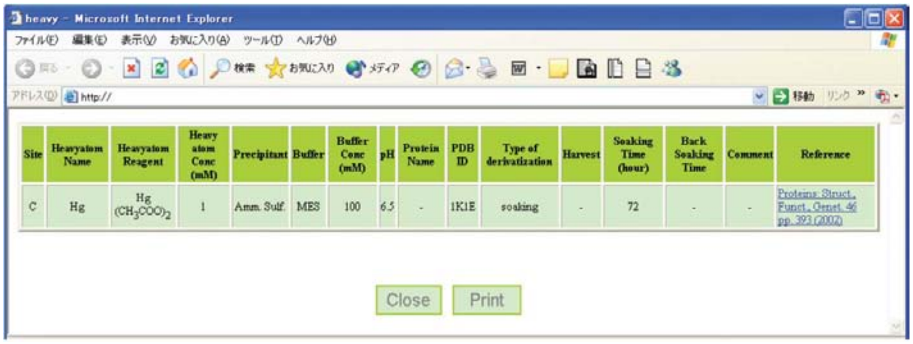
Amm. Sulf. Precipitant

MES Buffer

6 - 7 pH Range

Detail	Site	Heavyatom Reagent	Heavyatom Name	Precipitant	Buffer	pH	Protein Name
detail	C	$Hg(CH_3COO)_2$	Hg	Amm. Sulf.	MES	6.5	-
detail	C	$HgCl_2$	Hg	Amm. Sulf.	MES	6.5	-

Figure 1
Search tool and output windows. The window is separated into five parts (*a*)–(*e*). (*a*) The region displaying the amino-acid composition of the sequence. (*b*) The region displaying the search results. (*c*) Selection tool for heavy atom and crystallization conditions. (*d*) 'Detail' window display icon. (*e*) The region displaying the results of the motif search. The number of motifs found within the target sequence is displayed.


heavy – Microsoft Internet Explorer
ファイル(F) 編集(E) 表示(V) お気に入り(A) ツール(T) ヘルプ(H)
戻る
検索
お気に入り
メディア
アドレス(D) http://
移動
リンク
Close
Print

Site	Heavyatom Name	Heavyatom Reagent	Heavy atom Conc (mM)	Precipitant	Buffer	Buffer Conc (mM)	pH	Protein Name	PDB ID	Type of derivatization	Harvest	Soaking Time (hour)	Back Soaking Time	Comment	Reference
C	Hg	$Hg(CH_3COO)_2$	1	Amm. Sulf.	MES	100	6.5	-	1K1E	soaking	-	72	-	-	Proteins: Struct., Funct., Genet. 46 pp. 393 (2002)

Figure 2
'Detail' window. Detailed information for suggested heavy-atom reagents and putative binding residues are displayed.

Table 2
Heavy-atom-binding motif.

Heavy atom	Motif	Hit No.	Heavy atom	Motif	Hit No.
Hg	C*X*G	16	Pt	M*X*K	5
	G*X*C	12		M*X*G	4
	H*X*H	2		M*XX*G	2
	H*X*E	4		M*XX*T	3
	C*XX*C†	11		R*XX*M	6
	C*XX*G	11		K*XX*M	5
	C*XX*A	15		T*XX*M	2
	G*XX*C	12		G*XX*M	3
	A*XX*C	30		M*XXX*M	2
	C*XXX*C	12		M*XXX*K	7
	C*XXX*D	18		M*XXX*A	3
	C*XXX*G	16		M*XXX*Y	2
	C*XXX*M	8		M*XXX*T	2
	C*XXX*V	26		M*XXX*H	2
	C*XXX*F	18		M*XXX*G	5
	C*XXX*T	10		K*XXX*M	2
	C*XXX*A	14		Q*XXX*M	3
	T*XXX*C	9		L*XXX*M	5
	I*XXX*C	17		G*XXX*M	3
	A*XXX*C	14		GM*XX*T	2
	M*XXX*C	2		G*XX*M*X*G‡	1
	K*XXX*C	16			
	P*XXX*C	7	Sm	EE	3
	G*XXX*C	9		DD	3
	G*XX*CG	2		EXE	2
	H*XXX*M	4		DXD	3
	H*XX*Q*X*Q	2		I*X*D	3
	GMTC*XX*C§	1		E*XX*E	6
	MT/HC*XX*C¶	4		E*XX*G	2
	G*XXX*CG*XX*T	2		D*XX*E	2
				D*XX*D	2
Ag	GMTC*XX*C††	1		E*XXX*E	4
	CQ*XX*C	2		P*XXX*Y	2
Au	LC	3	Gd	EE	5
	C*X*E	4		E*X*G	3
	H*X*V	3		E*XX*E	5
	IH*X*V	2		E*XXX*E	5
	C*X*K	4		RE*XX*E	4
	C*XX*G	2		RE*XX*EE	4
	C*X*ED	2		D/EY*XX*V	2
	C*XXX*E	2			
	C*XXX*D	2	Pb	EE	6
	C*XXX*G	2		E*X*E	5
	V*XXX*C	2		D*X*D	3
	L*XXX*C	2		E*XX*E	4
	D*XX*H*X*V	2		E*XX*D	2
	K*X*C*XXX*G	2		E*XXX*E	6
				EQ*XX*E	3
				L*XXX*E	4

† Hopfner *et al.* (2002). ‡ The motif corresponds to the combination of G*XX*M and M*X*G. § Steel & Opella (1997). ¶ Wernimont *et al.* (2000). †† Gitschier *et al.* (1998).

Table 3
Results of experiments for the preparation of heavy-atom derivatives.

Successful motifs that bound predicted heavy atoms are in bold.

Protein ID/pH	Heavy-atom reagent	Success/ failure	Predicted motif
HTPF00126/4.6	$HgCl_2$	Success	Hg: **C*XXX*G**, **M*XXX*C**
	$KAu(CN)_2$	Success	Au: none
	$K_2Pt(CN)_4$	Failure	Pt: none
	K_2PtCl_6	Failure	Pt: none
	K_2PtCl_4	Failure	Pt: none
HTPF11492/6.5	K_2PtCl_4	Success	Pt: **M*XXX*K**, R*XX*M, **G*XXX*M**, **L*XXX*M**, **M*XXX*G**, **K*XX*M**, **M*X*G**, **K*XX*M**
	K_2PtCl_6	Success	Pt: **M*XXX*K**, **R*XX*M**, G*XXX*M, L*XXX*M, M*XXX*G, **K*XX*M**, M*X*G, **K*XX*M**
	$SmCl_3$	Success	Sm: **EE**, DD, E*XXX*E, I*X*D, E*XX*E, P*XXX*Y, E*XX*G
	$Hg(CH_3COO)_2$	Success	Hg: none
	EMTS	Failure	Hg: none
HTPF12045/5.7	K_2PtCl_4	Success	Pt: L*XXX*M, **K*XXX*M**, **K*XX*M**
	$HgCl_2$	Success	Hg: none
HTPF11055/4.8	$HgCl_2$	Success	Hg: **C*X*G**, **C*XXX*A**, H*XXX*M
	EMTS	Success	Hg: **C*X*G**, **C*XXX*A**, H*XXX*M
HTPF11731/7.4	K_2PtCl_4	Failure	Pt: G*XXX*M, M*XXX*K, L*XXX*M, M*XXX*G, GM*XX*T
HTPF11731/4.7	K_2PtCl_4	Success	Pt: G*XXX*M, M*XXX*K, L*XXX*M, M*XXX*G, GM*XX*T
HTPF00367/4.8	K_2PtCl_4	Success	Pt: none
	$Pb(CH_3COO)_2$	Failure	Pb: E*XX*E
HTPF10950/5.7	$Hg(CH_3COO)_2$	Success	Hg: **T*XXX*C**
	$KAu(CN)_2$	Success	Au: none
	EMTS	Failure	Hg: T*XXX*C
	K_2PtCl_4	Failure	Pt: R*XX*M, M*XXX*G
HTPF10017/4.4	$K_2Pt(CN)_4$	Success	Pt: none
	K_2HgI_4	Success	Hg: none
	K_2PtCl_4	Failure	Pt: none
	EMTS	Failure	Hg: none
HTPF00069/8.2	$SmCl_3$	Failure	Sm: EE, E*XX*E, E*XXX*E
HTPF00403/8.7	$HgCl_2$	Failure	Hg: C*XX*G

and lanthanides bind to Asp or Glu. Based on the heavy-atom database, the search tool finds reasonable heavy-atom reagents for the derivatization of any target protein.

The heavy-atom search is performed in three steps. In the first step, the program performs a single-amino-acid search against the heavy-atom database. The input to this program is the amino-acid sequence of a target protein. The program sorts the amino acids of the target sequence by residue type (Fig. 1*a*). The potential heavy-atom reagents and the probable binding residues are determined by searching against the database (Fig. 1*b*). Clicking the residue icon highlights the potential heavy-atom reagent (Fig. 1*a*). In the second step, the program filters the predicted reagents using the data on the crystallization conditions of the target protein; keys used for the filtering are precipitant, buffer and pH (Fig. 1*c*). The interactive consideration of these keys allows effective filtering. Thus, the number of candidates will decrease greatly in the second step. More detailed information is displayed by clicking the icon labelled 'detail' in Fig. 1(*d*) (Fig. 2). The third step is an advanced search based on heavy-atom-binding motifs (Fig. 1*e*). We examined the relationship between the heavy atoms and the heavy-atom-binding sites using the known structures of heavy-atom derivatives. Interestingly, we could identify 93 motifs that are frequent binding sites for certain heavy atoms (Table 2). In recent years, some heavy-atom-binding motifs have also been reported by others (Steel & Opella, 1997; Hopfner *et al.*, 2002; Wernimont *et al.*, 2000; Gitschier *et al.*, 1998). These known motifs are also included in *HATODAS*. The output of the advanced search lists the motif of the target protein by a search against the database.

In order to evaluate *HATODAS*, we performed a search on the ten proteins from *Thermus thermophilus* HB8 and *Pyrococcus horikoshii* OT3 (Table 3). Because there were no exact matches in the crystallization conditions, crystallization pH was mainly used as the key for heavy-atom filtering. Of the total of 27 heavy-atom reagents used in the derivatization experiments according to the *HATODAS* results, 16 were successfully predicted. In the motif search, the program found the motifs in the amino-acid sequence of nine proteins. For instance, the protein HTPF00126 was found to have two Hg motifs C*XXX*G and M*XXX*C. In agreement with the prediction, an Hg atom was found to be incorporated in the motif successfully. Furthermore, both Pt and Sm motifs in the protein HTPF11492 bound the predicted heavy atoms successfully.

3. Discussion

The outstanding advantage of Hg reagents in the derivatization of protein crystals is shown by its occurrence in 351 out of 776 cases in *HATODAS* (Table 1). Br and Pt are the next most often appearing entries in the database and constitute 80–90 cases. Sm, Au, Pb and U form the third group showing 30–40 cases; Xe, Ag, Pd, Gd, Tl and Kr belong to the fourth group showing 10–15 cases; the other heavy atoms occur in less than ten cases. An effective way of heavy-atom screening is to try the reagents suggested by *HATODAS* in this order of occurrence. In most failed cases it is observed that crystals crack on heavy-atom soaking. Alkaline crystallization conditions are considered to be one of the factors for the failure (Table 3). As pointed out by Blundell & Johnson (1976), alkaline conditions could disturb the derivatization owing to the chemical properties of most heavy-atom reagents. The results of using *HATODAS* suggest that the heavy-atom reagents $HgCl_2$, K_2PtCl_4 and $KAuCl_4$ provide better results in the pH ranges 4.6–8.5, 4.6–7.5 and 4.7–6.8, respectively. The structure–motif relationship is of particular interest. Some heavy-atom-binding motifs seem to reflect the secondary structure. In the case of the Hg-binding motif H*XXX*M, the two heavy-atom-binding residues His and Met on an α-helix tend to be located on the same side of the helix, thereby placing an interval of three arbitrary amino-acid residues between them. Interestingly, some motifs such as A*XX*C and C*XXX*V for Hg and M*XXX*K for Pt are most frequent, suggesting that the motif search may be useful to improve the efficiency of the heavy-atom screening. With further increases in the number of heavy-atom sites deposited, *HATODAS* is expected to provide greater insights into the motif search.

MS mainly contributed to this work and wrote the paper, YA, HU and HT wrote the program, HA contributed to the construction of the database including the motif search and NK supervised this work. The authors would like to thank M. R. N. Murthy for critical reading of the manuscript. This work was supported by the National Project on Protein Structural and Functional Analysis funded by the MEXT of Japan.

References

Blundell, T. L. & Johnson, L. N. (1976). *Protein Crystallography*, pp. 183–239. London: Academic Press.

Gitschier, J., Moffat, B., Reilly, D., Wood, W. I. & Fairbrother, W. J. (1998). *Nature Struct. Biol.* **5**, 47–54.

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

Hopfner, K. P., Craig, L., Moncalian, G., Zinkel, R. A., Usui, T., Owen, B. A., Karcher, A., Henderson, B., Bodmer, J. L., McMurray, C. T., Carney, J. P., Petrini, J. H. & Tainer, J. A. (2002). *Nature (London)*, **418**, 562–566.

Islam, S. A., Carvin, D., Sternberg, M. J. E. & Blundell, T. L. (1998). *Acta Cryst.* D**54**, 1199–1206.

Morris, R. J., Perrakis, A. & Lamzin, V. S. (2003). *Methods Enzymol.* **374**, 229–244.

Steel, R. A. & Opella, S. J. (1997). *Biochemistry*, **36**, 6885–6895.

Sugahara, M. & Miyano, M. (2002). *Tanpakushitsu Kakusan Koso*, **47**, 1026–1032.

Terwilliger, T. C. (2000). *Acta Cryst.* D**56**, 965–972.

Ueno, G., Hirose, R., Ida, K., Kumasaka, T. & Yamamoto, M. (2004). *J. Appl. Cryst.* **37**, 867–873.

Wernimont, A. K., Huffman, D. L., Lamb, A. L., O'Halloran, T. V. & Rosenzweig, A. C. (2000). *Nature Struct. Biol.* **7**, 766–771.

Yokoyama, S., Hirota, H., Kigawa, T., Yabuki, T., Shirouzu, M., Terada, T., Ito, Y., Matsuo, Y., Kuroda, Y., Nishimura, Y., Kyogoku, Y., Miki, K., Masui, R. & Kuramitsu, S. (2000). *Nature Struct. Biol.* **7**, 943–945.