JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

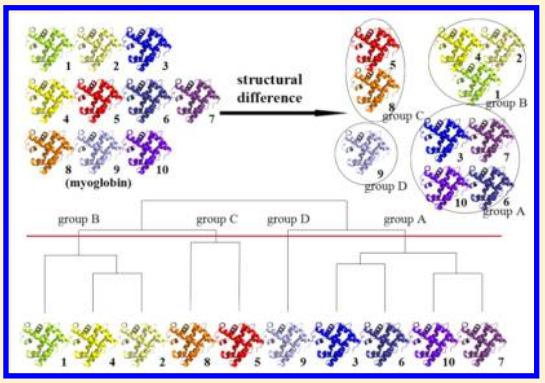
A Dominant Factor for Structural Classification of Protein Crystals

Fei Qi, Satoshi Fudo, Saburo Neya, and Tyuji Hoshino*

Graduate School of Pharmaceutical Sciences, Chiba University, Inohana 1-8-1, Chuo-ku, Chiba 260-8675, Japan

Supporting Information

ABSTRACT: With the increasing number of solved protein crystal structures, much information on protein shape and atom geometry has become available. It is of great interest to know the structural diversity for a single kind of protein. Our preliminary study suggested that multiple crystal structures of a single kind of protein can be classified into several groups from the viewpoint of structural similarity. In order to broadly examine this finding, cluster analysis was applied to the crystal structures of hemoglobin (Hb), myoglobin (Mb), human serum albumin (HSA), hen egg-white lysozyme (HEWL), and human immunodeficiency virus type 1 protease (HIV-1 PR), downloaded from the Protein Data Bank (PDB). As a result of classification by cluster analysis, 146 crystal structures of Hb were separated into five groups. The crystal structures of Mb (n = 284), HEWL (n = 336), HSA (n = 63), and HIV-1 PR (n = 488) were separated into six, five, three, and six groups, respectively. It was found that a major factor causing these structural separations is the space group of crystals and that crystallizing agents have an influence on the crystal structures. Amino acid mutation is a minor factor for the separation because no obvious point mutation making a specific cluster group was observed for the five kinds of proteins. In the classification of Hb and Mb, the species of protein source such as humans, rabbits, and mice is another significant factor. When the difference in amino sequence is large among species, the species of protein source is the primary factor causing cluster separation in the classification of crystal structures.



1. INTRODUCTION

Protein structure is related to biological function and thus provides essential information for structure-based drug design.[1,2] Proteins fold into respective unique structures, in which properties of the folding structures are divided into four distinct levels: primary structure of the amino sequence of polypeptide chains, secondary structure of an α-helix or β-sheet, tertiary structure of the complicated folding of peptide chains, and quaternary structure of a combination of two or more chains. A change at any structural levels will have an effect on the biological function.

X-ray crystal analysis is one of the most widely accepted techniques for determining protein structures. The number of structures deposited in Protein Data Bank (PDB) has been increasing exponentially in the past two decades, and the number of newly deposited structures in 2014 was about 8000. This huge accumulation of protein structures is due to the progress in protein crystallization[3] and X-ray sources[4] and also the utility of software for model building. Owing to the progress in crystallographic study, information on many protein crystal structures has become available. Even for a single kind of protein, variation in the amino sequence or protein conformation causes a difference in biological function. For example, the D30N mutation in HIV-1 protease is one of the well-known primary mutations for drug resistance.[5,6] HIV-1 protease with many amino mutations exhibits severe resistance to many inhibitors.[7] The change in conformation of the ras protein works as a switch for signal transduction.[8] Shape variation of a protein binding pocket influences the molecular recognition between the protein and a ligand.[9] Perhaps due to the variation of protein structure, some crystal structures are, however, sometimes incompatible with several biochemical and functional findings.[10] The diverse information on protein crystal structures is too multifarious to systematically understand the biological function and its alteration due to conformational change. If the conformational change or the structural variation of protein is limited, classification of the protein structure will be helpful to deduce the intrinsic characteristic of proteins.

Some different methods for the classification of diversity in protein structure have been proposed.[11−13] In our previous study, about 500 crystallographic structures of HIV-1 PRs were classified into six groups using cluster analysis. To our surprise, the cluster groups are distinguished not by the amino acid mutation but by the difference in space groups of crystals. A major factor for the separation of crystal structures is the space group, and the space group depends on the agents used in protein crystallization. A space group represents the positional and directional arrangement of molecules in a unit cell of the crystal and the arrangement is characterized by geometrical symmetry.[14] According to detailed analysis on the cluster

Received: January 29, 2015
Published: July 31, 2015

DOI: 10.1021/acs.jcim.5b00052
J. Chem. Inf. Model. 2015, 55, 1673−1685

classification of HIV-1 PRs, for example, crystal structures containing the L90 M mutation, which is known as a resistant mutation of Saquinavir and Nelfinavir,[15] were scattered all over the groups. Amino acid mutation is a minor factor for distinguishing the whole structure of HIV-1 PRs,[16] while the mutation is related to the difference in the side-chain and has an influence on the enzymatic specificity of a ligand. In this study, we examined four other kinds of proteins to determine whether the above-described finding is justified for proteins other than HIV-1 PR. X-ray crystal structures of proteins registered in PDB were used to performed cluster analysis. For reliable cluster analysis, one kind of protein to be examined should have a large number of crystal structures. Hence, hemoglobin (Hb), myoglobin (Mb), hen egg-white lysozyme (HEWL), and human serum albumin (HSA) were selected for the present study. In addition to the above four proteins, human immunodeficiency virus type 1 protease (HIV-1 PR) in the class B subtype was also examined by cluster analysis.

2. METHODS

2.1. Preparation of Data Set. The X-ray crystal structures of Hb, Mb, HEWL, HSA, and HIV-1 PR were taken from PDB by using the queries "hemoglobin", "myoglobin", "hen egg white lysozyme", "human serum albumin", and "human immunodeficiency virus type 1 protease". The total numbers of crystal structures downloaded were 642, 361, 473, 106, and 697, respectively. Since some crystal structures were not directly relevant to the intended proteins, they were eliminated from the data set. Crystal structures that have missing residues not at the N-terminal or C-terminal sides were also eliminated. Structures obtained by a technique other than X-ray diffraction were also eliminated. Consequently, 611, 355, 468, 89, and 618 structures were left as candidates for the initial data set.

2.2. Data Filtration. Hb is a tetramer composed of two pairs of two subunits designated α and β (Figures 1a and 1b). The α-subunit contains 141 amino acid residues, and the β-subunit consist of 146 residues (WT of human: 1XXT). Such crystal structures that have chemical modification were also excluded. By checking every structure, 146 crystal structures were selected for the data set of Hb.

Mb is a monomeric heme protein comprised of 152 amino acid residues (WT of sperm whale: 1DUK) (Figure 1c). Then, 284 crystal structures were selected.

HEWL is a monomer composed of 129 amino residues (WT: 1UC0) (Figure 1d). To examine the difference by diffraction methods, the structures obtained from powder diffraction were also included. The number of crystal structures for HEWL in the data set was 336.

HSA is a monomer protein, and the residues from His3 to Gly584 appear in the wild-type crystal structure of 1GNJ (Figure 1e). Then, crystal structures that have missing residues at the N-terminal and C-terminal sides were included. To align the sequence, the coordinates of the residues before Lys4 and after Leu583 were deleted. In several crystal structures, one unit cell contains two molecules. Hence, only chain A was used in the analysis. Consequently, 63 crystal structures were selected for HSA.

HIV-1 protease is a homodimer, in which each monomer consists of 99 amino residues (Figure 1f). The data set was restricted to proteins derived from the class B subtype (WT: 4DJP), and the selected number of HIV-1 PRs was 488.

2.3. Cluster Analysis. The coordinates of substrates, inhibitors, ions, water molecules, and all other heteroatoms were removed from all of the data. First, the main-chain atoms are extracted from every structure. The extracted coordinates are fitted to that of the wild type. Then, the average coordinates of the main-chain atoms are derived from the 146 structures in the case of Hb. All of the structures are fitted to the average one to calculate the root-mean-square deviations (RMSDs). Then, a 146 × 146 matrix of RMSD values were generated from 146 crystal structures. Euclidean distance was calculated in the RMSDs matrix. On the basis of the RMSDs, the structures are classified into groups by performing cluster analysis with the nearest neighboring method using the "hclust" function of R software.[17,18] Finally, amino mutation, space group of the crystal, species of protein sources, Matthews coefficient, resolution X-ray diffraction, and crystallization conditions were surveyed for every structure of the respective clusters. In the nearest neighboring method, 146 structures initially provide 146 clusters with each cluster being composed of a single member. One pair of clusters that have the nearest distance is merged, and the number of clusters is decreased by one. By repeating the search of the nearest two clusters and merging them, all of the structures are finally connected as a tree called a dendrogram. The x-axis of the dendrogram is the label number for the crystal structures. The y-axis is the distance for the least dissimilarity among the individual crystal

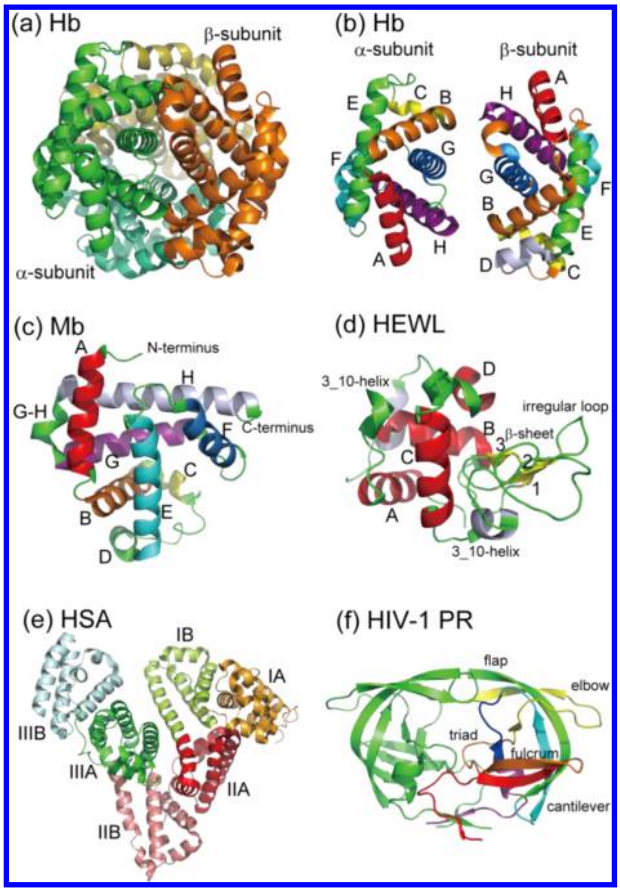


Figure 1. (a) Crystal structure of Hb (PDB: 1XXT). Two types of subunits are indicated by different colors. (b) Schematic representation of the secondary structures in two subunits of Hb. The helical segments of two subunits are represented by colors. (c) Crystal structure of Mb (PDB: 1DUK). Colors represent the helical regions. (d) Crystal structure of HEWL (PDB: 1UC0). (e) Crystal structure of HSA (PDB: 1GNJ). The respective domains are shown by colors. (f) Crystal structure of HIV-1 PR (PDB: 4DJP).

DOI: 10.1021/acs.jcim.5b00052
J. Chem. Inf. Model. 2015, 55, 1673−1685

structures based on RMSD matrix. To classify the crystal structures from the dendrogram, the criteria of distance for separation or the number of groups should be determined in advance. In the present study, the number of groups was determined in the respective cluster analysis and the number was set to five in case of Hb. As for Mb, HEWL, HSA, and HIV-1 PR, the same procedure was performed for cluster analysis.

3. RESULTS

3.1. Clustering of the Protein Crystal Structures of Hb.

A dendrogram deduced from cluster analysis is shown in Figure 2a. From the shape of the tree and the branches, the protein

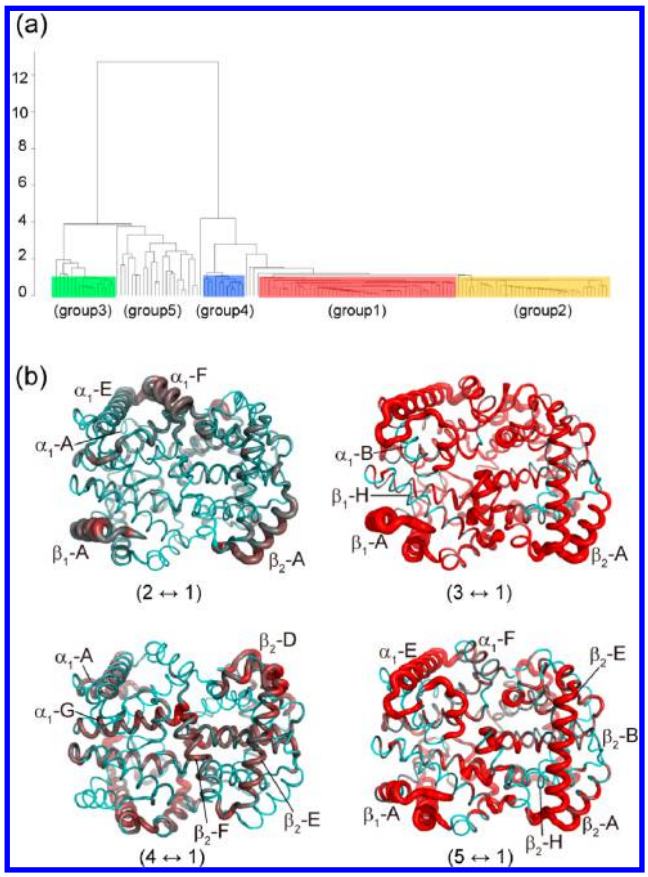


Figure 2. (a) Dendrogram for cluster analysis of crystal structures of Hb. A total of 146 structures were grouped into five clusters. The height of the tree indicates the root-mean-square distance of the main-chain atoms among the crystal structures. (b) Comparison of the average structures of cluster groups. Deviation of the main-chain atoms in the average structures of group 2, group 3, group 4, and group 5, measured from that of group 1. The deviation increases as the color changes from cyan to red.

crystal structures of Hb are separated into five clusters. The clusters are labeled from group 1 to group 4 that are from major to minor in number of members. The structures not assigned to groups 1−4 are categorized into group 5. A list of cluster members of the respective groups is present in List S1 of the Supporting Information. The numbers in the respective groups are 52, 40, 16, 11, and 27. About 36% of the structures belong to the largest cluster, group 1. Group 2 accounts for 27% of the structures. The smallest cluster, group 4, contains less than 8% of the structures.

The average structure obtained for each cluster group is superimposed on that of group 1 for comparison (Figure 2b). Hb is composed of four polypeptide subunits, two α and two β-chains. Each α-chain contains seven helical and seven nonhelical segments, while each β-chain contains eight helical and six nonhelical segments (Figure 1b).[19] Protein structures exhibit C2 symmetry due to the orientation of the $\alpha_1\beta_1$ dimer relative to $\alpha_2\beta_2$.[20] A structural comparison between groups 1 and 2 shows large deviations at helical segment F in the α-chain and at helical segment A in the β-chain (Figure 2b (2↔1)). The amplitude of structural deviation between groups 1 and 3 is obviously larger than that between groups 1 and 2 and that between groups 1 and 4. In the comparison of groups 1 and 3, the deviation at helical segment B in the α-chain and helical segment H in the β-chain is small as shown by the cyan color in Figure 2b (3↔1). A comparison between groups 1 and 4 in Figure 2b (4↔1) shows a large structural difference at segments A and G of the α-chain and segments D, E, and F in the β-chain. The deviations are also distributed symmetrically to the other dimer. In the comparison of groups 1 and 5, the deviation is partially small at helical segments A, F, and H in α-chains and B, F, and H in β-chains, while the other parts show large differences (Figure 2b (5↔1)).

To find a critical factor to characterize the respective clusters, the properties of the cluster members were examined in terms of crystallization condition, species of protein source, and amino acid mutation. It was found that the primary factor to distinguish the crystal structures of Hb was the space group of the crystal as shown in Table 1. This is in agreement with the results of our previous study for HIV-1 PR.[16] The major space group in the crystals of Hb is P 21 21 2 and the second one is P 1 21 1. The members of these two space groups reach 74.7% of all of the crystal structures. As seen in Table 1, most of the crystal structures bearing the P 21 21 2 space group belong to group 1. Group 4 also contains crystal structures with the P 21 21 2 space group. Most of the crystal structures with the P 1 21 1 space group are in group 2. Group 3 mainly contains crystal structures with P 21 21 21 and P 32 2 1 space groups. Since group 5 consists of structures that are not assigned to the other groups, it is natural that the cluster members of group 5 are broadly distributed over the space groups.

3.2. Clustering of the Protein Crystal Structures of Mb.

A dendrogram obtained from cluster analysis of Mb is shown in Figure 3a. Judging from the shape of the tree, 284 structures of Mb are separated into six clusters (Figure 3). As was the case for the classification of Hb, the clusters are labeled from group 1 to group 5 from major to minor. The structures of group 6 are not assigned to any other groups. The numbers of structures in the groups are 129, 57, 46, 19, 7, and 26, respectively. A list of cluster members of the respective groups is shown in List S2. The largest cluster, group 1, contains about 45% of the structures. The smallest cluster, group 5, accounts for about 2% of the clusters.

Mb is a monomeric heme protein comprised of eight right-handed α-helices that are connected by short nonhelical segments (Figure 1c).[21,22] The average structures of the respective groups were obtained and superimposed with that of group 1. A comparison between groups 1 and 3 (Figure 3b (3↔1)) indicates a structural difference at helical segment C and at the C-terminus. The deviation between groups 1 and 3 is obviously larger than that between groups 1 and 2 (Figure 3b (2↔1)). Furthermore, there is large amplitude in deviation between helical segments G and H. A comparison of groups 1

DOI: 10.1021/acs.jcim.5b00052
J. Chem. Inf. Model. 2015, 55, 1673−1685

Table 1. Number of Crystal Structures Classified by Space Group for Hemoglobin (Hb)

space group	group1	group2	group3	group4	group5	total
P 21 21 2	50(0.96)[b]	0(0.00)	0(0.00)	11(1.00)	2(0.07)	63
P 1 21 1	2(0.04)	36(0.90)	1(0.06)	0(0.00)	7(0.26)	46
P 21 21 21	0(0.00)	4(0.10)	7(0.44)	0(0.00)	13(0.48)	24
P 32 2 1	0(0.00)	0(0.00)	6(0.38)	0(0.00)	0(0.00)	6
others[a]	0(0.00)	0(0.00)	2(0.13)	0(0.00)	5(0.19)	7
total	52	40	16	11	27	146

[a]Others is the sum of those for C 1 2 1, P 41 21 2, P 1, and P 61 2 2 space groups. [b]The value in parentheses represents the ratio relative to the total number of crystal structures in each group.

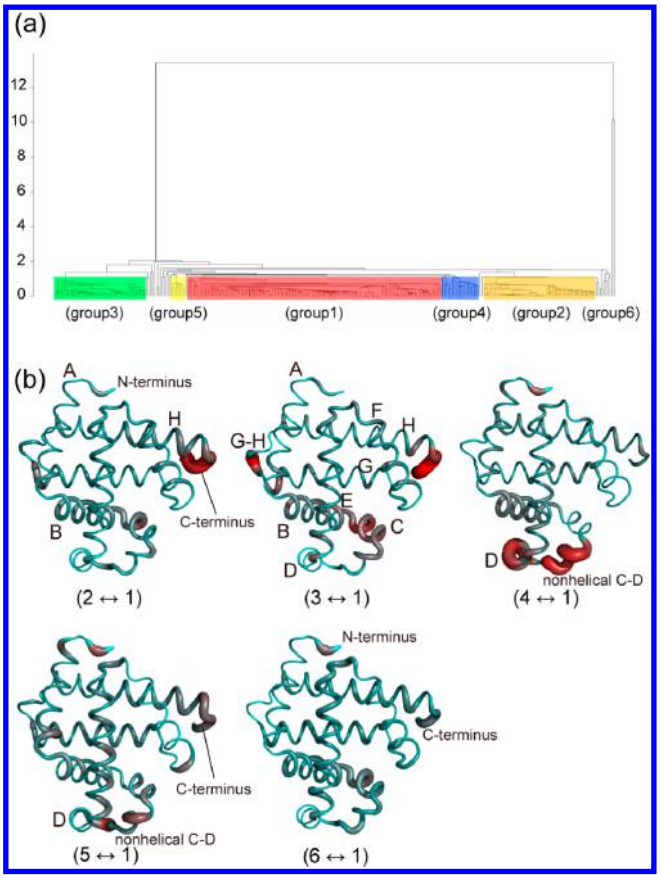


Figure 3. (a) Dendrogram for cluster analysis of crystal structures of Mb. A total of 284 structures were grouped into six clusters. The height of the tree indicates the root-mean-square distance of the main-chain atoms among the crystal structures. (b) Comparison of the average structures of cluster groups. Deviation of the main-chain atoms in the average structures of group 2, group 3, group 4, group 5, and group 6, measured from that of group 1. The deviation increases as the color changes from cyan to red.

and 4 (Figure 3b (4↔1)) shows that there is a large difference in nonhelical segments C−D and helical segment D. A comparison of groups 1 and 5 (Figure 3b (5↔1)) shows that there are remarkable deviations at helical segments C−D and the C-terminus.

The members of the respective clusters are classified in terms of space group (Table 2). The major space group for the crystal of Mb is P 6. The second major one is P 1 21 1. The members of these two space groups reach 85% of the crystal structures. Most of the crystal structures bearing the P 6 space group are in group 1 as shown in Table 2. Most of the crystal structures bearing the P 1 21 1 space group are restricted to groups 2 and 3. Group 4 consists of all of the crystal structures with the P 21 21 21 space group. All the structures bearing the P 21 21 2 space group are restricted to group 5. Group 6 is composed of crystal structures with miscellaneous space groups.

3.3. Clustering of the Protein Crystal Structures of HEWL. A dendrogram obtained from cluster analysis of HEWL is shown in Figure 4a. The 336 crystal structures of HEWL are separated into five clusters. The numbers of members in the groups are 276, 25, 12, 11, and 12, respectively, and the cluster members are shown in List S3. About 82% of the structures are in the largest cluster, group 1. The smallest cluster, group 4, accounts for about 3% of the structures.

HEWL consists of an α helical region of four α-helices and one 3_{10}-helix, one triple-stranded antiparallel β-sheet, and an irregular loop with two disulfide bridges (Figure 1d).[23,24] A comparison between groups 1 and 3 shows structure differences at the irregular loop and helical segment C (Figure 4b (3↔1)). The deviation between groups 1 and 3 is larger than that between groups 1 and 2 (Figure 4b (2↔1)). A comparison of groups 1 and 4 (Figure 4b (4↔1)) indicates that deviations are broadly distributed over helical segments A, C, and D, the β-sheet C segment, and the irregular loop. A comparison of groups 1 and 5 shows that there is a notable deviation at helical segment D and the irregular loop (Figure 4b (5↔1)).

The members of the respective clusters are classified in terms of space group (Table 3). The major space group for the

Table 2. Number of Crystal Structures Classified by Space Group for Myoglobin (Mb)

space group	group1	group2	group3	group4	group5	group6	total
P 21 21 21	6(0.05)[b]	0(0.00)	3(0.07)	19(1.00)	0(0.00)	7(0.27)	35
P 1 21 1	0(0.00)	57(1.00)	43(0.93)	0(0.00)	2(0.29)	14(0.54)	116
P 6	123(0.95)	0(0.00)	0(0.00)	0(0.00)	0(0.00)	1(0.04)	124
P 21 21 2	0(0.00)	0(0.00)	0(0.00)	0(0.00)	5(0.71)	0(0.00)	5
others[a]	0(0.00)	0(0.00)	0(0.00)	0(0.00)	0(0.00)	4(0.15)	4
total	129	57	46	19	7	26	284

[a]Others is the sum of those for P 41, I 21, A 2, and P 61 2 2 space groups. [b]The value in parentheses represents the ratio relative to the total number of crystal structures in each group. For example, group 1 contains 129 structures, and 95% of the 129 structures bear a P 6 space group.

DOI: 10.1021/acs.jcim.5b00052
J. Chem. Inf. Model. 2015, 55, 1673−1685

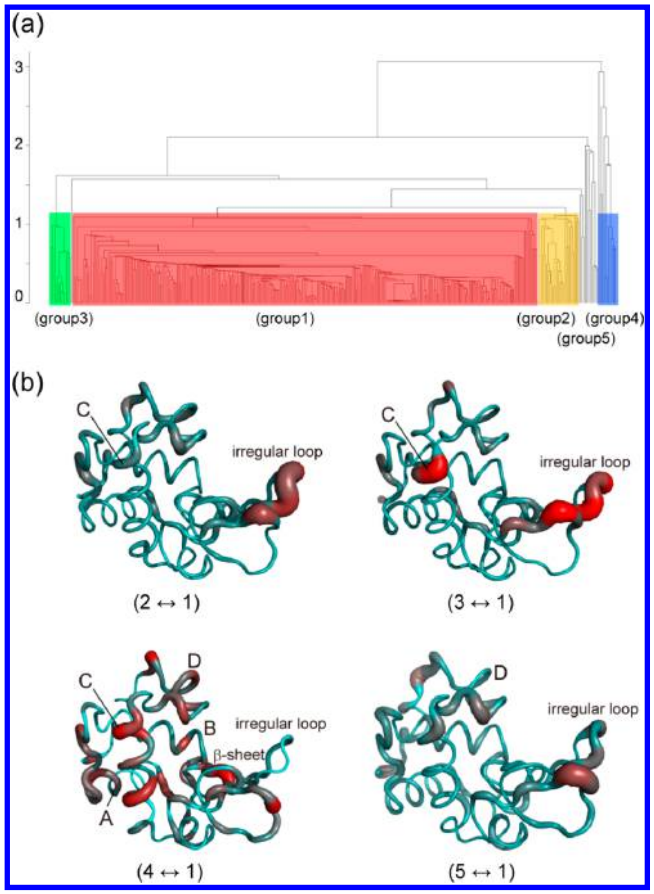


Figure 4. (a) Dendrogram for cluster analysis of crystal structures of HEWL. A total of 336 structures were grouped into five clusters. (b) Comparison of the average structures of cluster groups. See also the caption of Figure 2.

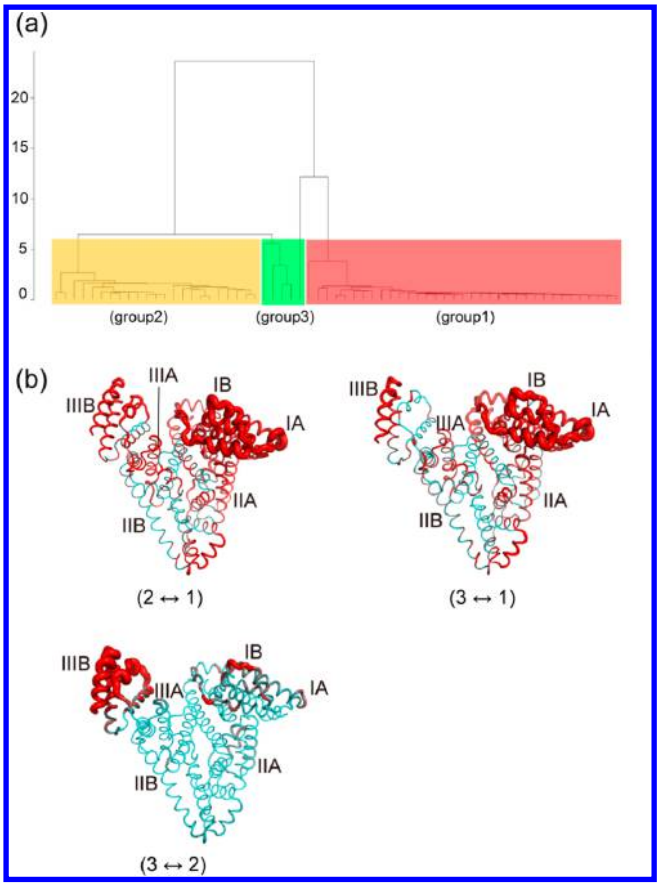


Figure 5. Dendrogram for cluster analysis of crystal structures of HSA. A total of 63 structures were grouped into three clusters. (b) Comparison of the average structures of cluster groups. Deviations of the main-chain atoms in the average structures of group 1, group 2, and group 3 are compared.

crystals of HEWL is P 43 21 2, which is in about 88% of the structures. The second major space group is P 21 21 21, which is in about 5% of the structures. In Table 3, group 1 consists only of crystal structures with P 43 21 2. The remaining crystal structures with P 43 21 2 are distributed in group 4 and group 5. All of the crystal structures bearing the P 21 21 21 space group are in group 2. All structures bearing P 1 are in group 3. In addition to structures bearing P 43 21 2, group 5 contains crystal structures bearing P 1 21 1.

3.4. Clustering of the Protein Crystal Structures of HSA. A dendrogram obtained from cluster analysis of HSA is shown in Figure 5a, in which the 63 crystal structures are separated into three clusters. The structures in group 3 are those not assigned to groups 1 and 2. The three cluster groups contain 35, 23, and 5 members as shown in List S4. The number of structures in groups 1 and 2 accounts for 87% of the total number of structures.

HSA consists of three homologous domains labeled I, II, and III. Each domain is further classified into two subdomains, A and B (Figure 1e).[25] A comparison of the average structures indicates that the distribution of deviations between groups 1 and 2 (Figure 5b (2↔1)) is almost the same to that between groups 1 and 3 (Figure 5b (3↔1)). The structure deviations are prominent at domain I, domain IIA, and domain IIIB. In a comparison of groups 2 and 3 (Figure 5b (3↔2)), the deviation is obvious at domain IIIB.

The members of the respective clusters are classified in terms of space group (Table 4). All of the crystal structures bearing a C 1 2 1 space group are in group 1, and they account for 94%

Table 3. Number of Crystal Structures Classified by Space Group for Hen Egg White Lysozyme (HEWL)

space group	group1	group2	group3	group4	group5	total
P 43 21 2	276(1.00)[b]	3(0.12)	0(0.00)	10(0.91)	5(0.42)	294
P 21 21 21	0(0.00)	18(0.72)	0(0.00)	0(0.00)	0(0.00)	18
P 1 21 1	0(0.00)	1(0.04)	0(0.00)	0(0.00)	7(0.58)	8
P 1	0(0.00)	1(0.04)	11(0.92)	0(0.00)	0(0.00)	12
others[a]	0(0.00)	2(0.08)	1(0.08)	1(0.09)	0(0.00)	4
total	276	25	12	11	12	336

[a]Others is the sum of those for C 1 2 1, P 61 2 2, and A 1 space groups. [b]The value in parentheses represents the ratio relative to the total number of crystal structures in each group.

DOI: 10.1021/acs.jcim.5b00052
J. Chem. Inf. Model. 2015, 55, 1673−1685

Table 4. Number of Crystal Structures Classified by Space Group for Human Serum Albumin (HSA)

space group	group1	group2	group3	total
C 1 2 1	33(0.94)[b]	0(0.00)	0(0.00)	33
P 1	1(0.03)	21(0.91)	0(0.00)	22
others[a]	1(0.03)	2(0.09)	5(1.00)	8
total	35	23	5	63

[a]Others is the sum of those for P 1 21 1, P 41 21 2, and P 21 21 2 space groups. [b]The value in parentheses represents the ratio relative to the total number of crystal structures in each group.

of the members in group 1. Most of members with a P 1 space group are in group 2. Since group 3 consists of structures not assigned to other groups, group 3 contains other miscellaneous space groups.

3.5. Clustering of the Protein Crystal Structures of HIV-1 PR. A dendrogram obtained from cluster analysis of HIV-1 PR is shown in Figure 6a. The 488 crystal structures of HIV-1 PR are separated into six clusters. Group 6 consists of structures that were not assigned to groups 1−5, and a list of cluster members is provided in List S5. The numbers of group members are 171, 77, 63, 44, 17, and 116, respectively. The largest cluster is group 1, and it contains 35% of the structures. The smallest cluster, group 5, contains less than 3% of the structures.

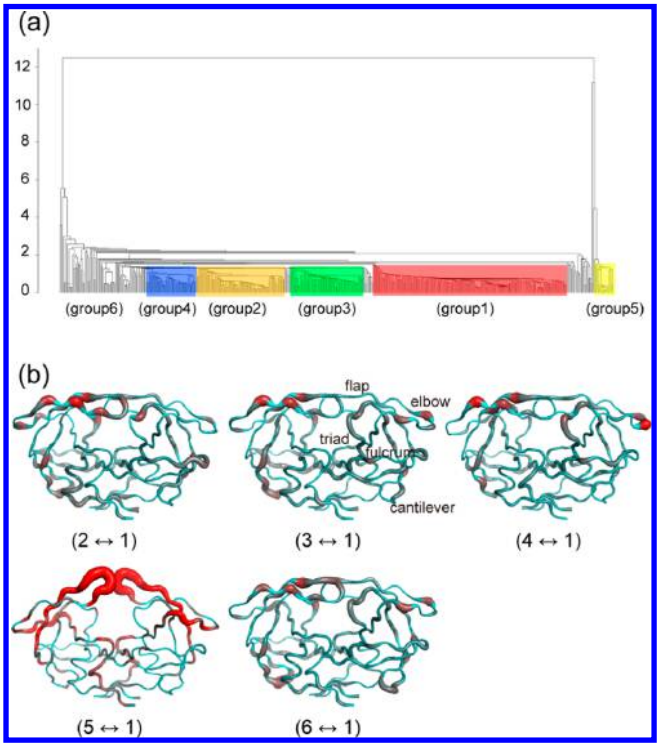


Figure 6. (a) Dendrogram for cluster analysis of crystal structures of HIV-1 PR. A total of 488 structures were grouped into six clusters. (b) Comparison of the average structures of cluster groups. See also the caption of Figure 2.

HIV-1 PR is composed of functional regions named flap, cantilever, elbow, catalytic triad, and fulcrum (Figure 1f).[16] A comparison of average structures in Figure 6b (5↔1) indicates that the flap and elbow regions have large structural differences among clusters. Furthermore, the flap is semi-opened in group 5, and the shape is quite different from those of other groups.

The members of the respective clusters are classified in terms of space group (Table 5). It is notable that most of the crystal structures with a P 21 21 2 space group are in group 1. Most of the members of groups 2 and 4 have a P 21 21 21 space group. Most of the crystal structures with a P 61 space group belong to group 3. Most of the crystal structures with a P 41 space group belong to group 5. Since group 6 is a mixed group, this group includes crystals with all kinds of space groups.

3.6. Molecular Geometry in the Protein Crystals. In order to examine the influence of crystal packing on the structural deviations in Figures 2−6, the molecular geometry in protein crystals were examined. The crystal packing arrangement for every cluster group of Mb is shown in Figure 7 for example. In group 1, 95% crystal structures bear the P 6 space group, and they are in the hexagonal form. Several segments of the protein are near to the neighboring molecules, such as N-terminus, nonhelical segment C−D, helical segment E, helical segment F, nonhelical segment G−H, and C-terminus. A magnified view at the protein−protein contact in the crystal packing with the P 6 space group in group 1 is shown in Figure 8a. The closest distance at the molecular contact is about 2.8 Å. The distance is in the similar range to hydrogen binding. All the crystal structures in group 2 bears the P 1 21 1 space group, and they are in the monoclinic form. Some segments of the protein such as helical segment C, nonhelical segment C−D, nonhelical segment D−E, nonhelical segment E−F, helical segment H, and C-terminus are close to the neighboring molecules. The comparison of the average structures between groups 1 and 2 indicates a large deviation at helical segment C and C-terminus (Figure 3b). The helical segment C is located at the contact only for group 2, while C-terminus is at the contact both in groups 1 and 2. Hence, the difference in the average structure is not completely compatible with the molecular contact. Most of the crystal structures in group 3 (93%) and half of the crystal structures in group 6 (54%) have the P 1 21 1 space group that is the popular space group in group 2. The nonhelical segment

Table 5. Number of Crystal Structures Classified by Space Group for Human Immunodeficiency Virus Type 1 (HIV-1) Protease

space group	group1	group2	group3	group4	group5	group6	total
P 21 21 2	170(0.99)[b]	1(0.01)	0(0.00)	0(0.00)	0(0.00)	17(0.15)	188
P 21 21 21	1(0.01)	71(0.92)	0(0.00)	44(1.00)	1(0.06)	53(0.46)	170
P 61	0(0.00)	0(0.00)	58(0.92)	0(0.00)	0(0.00)	19(0.16)	77
P 41	0(0.00)	0(0.00)	0(0.00)	0(0.00)	16(0.94)	2(0.02)	18
P 1 21 1	0(0.00)	3(0.04)	5(0.08)	0(0.00)	0(0.00)	8(0.07)	16
others[a]	0(0.00)	2(0.03)	0(0.00)	0(0.00)	0(0.00)	17(0.15)	19
total	171	77	63	44	17	116	488

[a]Others is the sum of those for C 1 2 1, I 41 2 2, I 2 2 2, P 1, P 1 1 21, P 43, P 41 21 2, and P 43 21 2 space groups. [b]The value in parentheses represents the ratio relative to the total number of crystal structures in each group.

DOI: 10.1021/acs.jcim.5b00052
J. Chem. Inf. Model. 2015, 55, 1673−1685

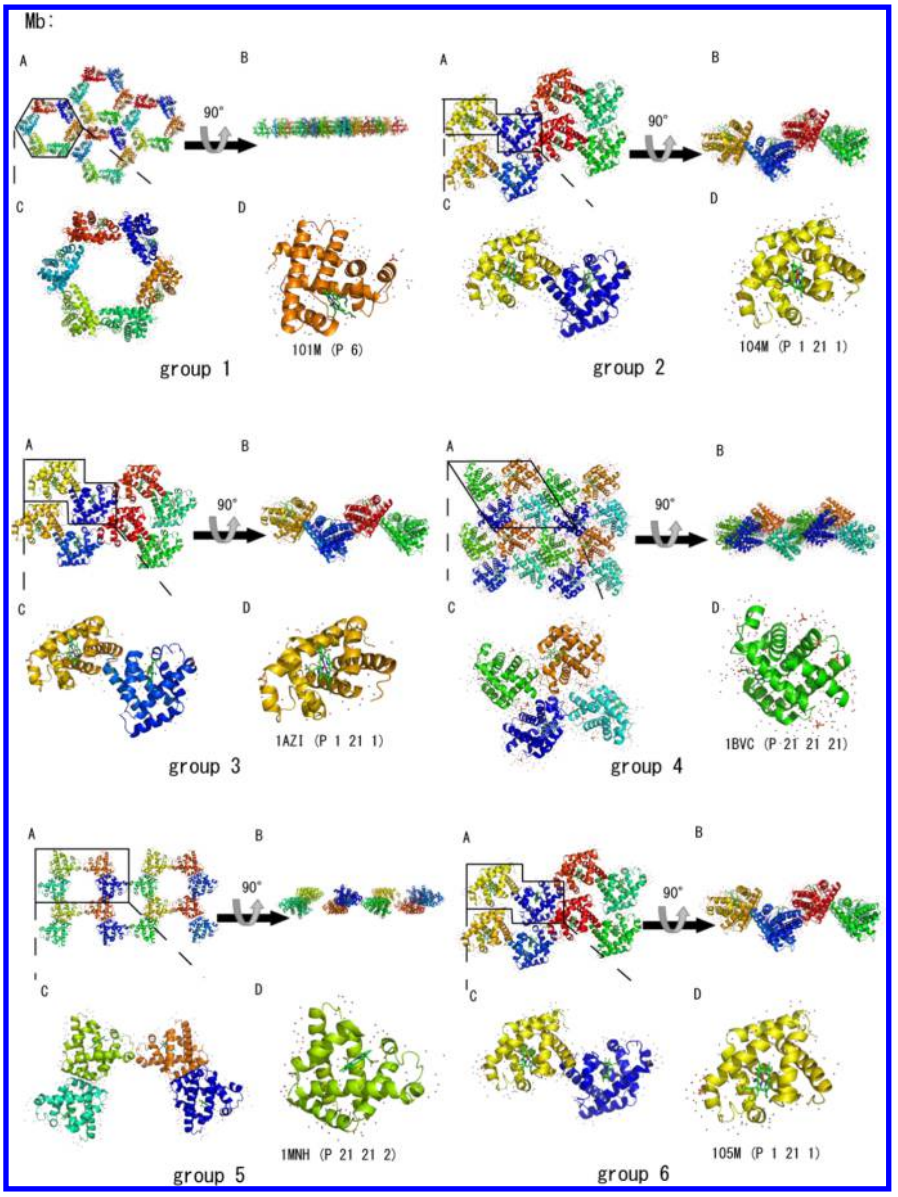


Figure 7. Molecular geometry in the crystal packing for every cluster group of Mb. One crystal structure that belongs to the most major space group is selected as a representative of each cluster group. (A) Arrangement of proteins in a crystal viewed from one crystal axis. (B) Arrangement of proteins in the crystal viewed from another direction. (C) Magnified view of the asymmetric unit. (D) Protein structure in the selected crystal with its PDB code and space group.

G−H of group 3 is largely deviated from that of group 1, while the deviation of this segment is small in groups 2 and 6. The deviation of the C-terminus of group 6 is not so large as those of groups 2 and 3. Although the crystal packing arrangements are common among groups 2, 3, and 6, the comparison of these average structures in Figure 3b show that the structural deviation is not always observed at the same regions. In group 4, all the crystal structures bear the P 21 21 21 space group, and they are in the orthorhombic form. There are many segments making contact to the neighboring molecules such as N-terminus, helical segment B, nonhelical segment C−D, nonhelical segment E−F, helical segment G, and nonhelical segment G−H. Large structural deviations are observed in the N-terminus, helical segment B, helical segment D, and nonhelical segment C−D (Figure 3b). A comparison between the crystal packing arrangement in Figure 7 and the structural deviation in Figure 3b suggests that not all segments adjacent to the neighboring molecules show large deviations. In group 5, most of the crystal structures bear P 21 21 2, and they are in the orthorhombic form. Several segments such as helical segment A, nonhelical segment A−B, nonhelical segment C−D, helical segment D, helical segment H, and C-terminus are close to the neighboring molecules while noticeable deviations are observed at N-terminus, helical segment A, nonhelical segment C−D, and C-terminus. A magnified view at the contact in the crystal packing with the P 21 21 2 is shown in Figure 8b. This figure also show that the closest protein−protein distance is in the range of hydrogen binding.

The packing arrangements for Hb, HEWL, HSA, and HIV-1 PR are shown in Figures S1−S4. The large deviations are sometimes seen in the segments that are close to the neighboring molecules. A close examination with these figures suggests that the packing arrangement varies depending on the space group and then has an influence on the difference in

DOI: 10.1021/acs.jcim.5b00052
J. Chem. Inf. Model. 2015, 55, 1673−1685

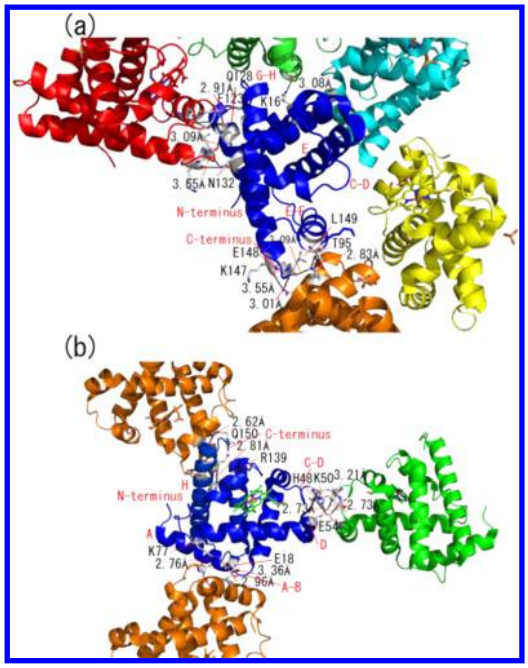


Figure 8. Molecular contact in the crystal packing for Mb. (a) A crystal structure with the P 6 space group in group 1 (PDB: 101M). (b) A crystal structure with the P 21 21 2 space group in group 5 (PDB: 1MNH). The protein−protein contact areas are shown in white.

protein structures among cluster groups, but the contact region is not always correlated with the structural deviations.

4. DISCUSSION

To make clear the amplitude of the structural difference, RMSD values among the cluster groups were measured by superimposing of the average structures of each cluster group on that of group 1. The RMSD values for Hb were calculated as shown in Table S7. The RMSD value between groups 1 and 3 (4.323 Å) is larger than that between groups 1 and 2 (2.296 Å) and that between groups 1 and 4 (2.410 Å). These results are consistent with Figure 2b. Although the RMSD between groups 1 and 2 (2.296 Å) is close to that between groups 1 and 4 (2.410 Å), the structural deviations are observed at the different regions in Figure 2b (2↔1) and Figure 2b (4↔1). There are large deviations at helical segment F in the α-chain and at helical segment A in the β-chain between groups 1 and 2, while large structural differences are observed at segments A and G of the α-chain and segments D, E, and F in the β-chain between groups 1 and 4. RMSD measurement for Mb in Table S8 indicates that the RMSD ranges from 2.28 to2.44 Å in case of the comparison between groups 1 and other groups. The RMSD between groups 2 and 5 is small (0.583 Å). The major space group is common between groups 2 and 3 (P 1 21 1). The crystal structures between groups 2 and 3 bearing the same space group show a large RMSD value (0.828 Å) compared to that between groups 2 and 4 (0.613 Å) and that between groups 2 and 5 (0.583 Å). The RMSD value between groups 1 and 2 is equal to that between groups 1 and 5. However, the structural deviations were observed at the different regions as shown in Figure 3b. RMSD values for HEWL in Table S9 indicates that the difference between groups 1 and 4 is larger than others. Although the comparison of the average structures between groups 1 and 3 exhibits a large deviation at the irregular loop, the RMSD value for the whole part between groups 1 and 3 is less than that between groups 1 and 4. RMSD values for HSA in Tables S10 indicates that group 1 and group 2 exhibit a large difference. This difference is due to the change of the positions of domain IA and IB as shown in Figure 5b. The RMSD values among the groups 2, 3, and 4 for HIV-1 PR in Tables S11 are small and below 0.55, while those measured from groups 1 and 5 is large. This means that groups 2, 3, and 4 can be regarded as one big cluster. Hence, the reason why the crystal structures with the same space group are separated in different cluster groups is not fully explained only from RMSD values.

The cluster analysis in the present study demonstrated that the structural diversity of protein crystals closely depends on the space group of the crystal. It is interesting to examine the relationship between amino acid mutation and structural classification. Amino acid mutations in the cluster members of the respective groups were surveyed for HIV-1 PR as shown in Table 6. The survey mainly focused on the primary resistant mutations for protease inhibitors. The Q7K mutation, which is used in experiments to increase expression efficiency, is distributed across all cluster groups. The D30N mutation, which is known as a resistant mutation for Nelfinavir, is seen in all of the groups except for group 2. Some clusters have no structure with G48V and I50V mutations, because of the limited number of crystal structures with G48V or I50V mutations. V82A,F,T and I84V mutations appear in all cluster

Table 6. Number of Crystal Structures Bearing Amino Mutations for Cluster Groups for HIV-1 PR

mutation	group1	group2	group3	group4	group5	group6	total
Q7K	96(0.56)[a]	75(0.97)	5(0.08)	43(0.98)	1(0.06)	67(0.58)	287
D25N	2(0.01)	7(0.09)	0(0.00)	7(0.16)	16(0.94)	14(0.12)	46
D30N	4(0.02)	0(0.00)	4(0.06)	2(0.05)	1(0.06)	6(0.05)	17
S37N,E	9(0.05)	74(0.96)	7(0.11)	23(0.52)	17(1.00)	62(0.53)	192
G48V	1(0.01)	2(0.03)	1(0.02)	0(0.00)	0(0.00)	3(0.03)	7
I50V	8(0.05)	1(0.01)	0(0.00)	1(0.02)	0(0.00)	4(0.03)	14
I54V,L	4(0.02)	2(0.03)	1(0.02)	0(0.00)	17(1.00)	23(0.20)	47
L63P	7(0.04)	73(0.95)	0(0.00)	17(0.39)	17(1.00)	42(0.36)	156
C67A	65(0.38)	1(0.01)	5(0.08)	27(0.61)	1(0.06)	25(0.22)	124
V82A,F,T	16(0.09)	6(0.08)	11(0.17)	12(0.27)	16(0.94)	28(0.24)	89
I84V	20(0.12)	3(0.04)	9(0.14)	7(0.16)	17(1.00)	18(0.16)	74
L90M	4(0.02)	0(0.00)	2(0.03)	5(0.11)	17(1.00)	20(0.17)	48
C95A	65(0.38)	1(0.01)	10(0.16)	27(0.61)	1(0.06)	27(0.23)	131

[a]The value in parentheses represents the ratio relative to the total number of crystal structures in each group.

DOI: 10.1021/acs.jcim.5b00052
J. Chem. Inf. Model. 2015, 55, 1673−1685

groups. The L90 M mutation, which is known as a resistant mutation for Saquinavir and Nelfinavir, is seen in all of the groups except for group 2. Therefore, no clear relationship between the mutation and cluster groups was observed in HIV-1 PR. Amino acid mutations in crystal structures for Hb, Mb, HEWL, and HSA were also surveyed as shown in Tables S1–S4. Because of the limited number of mutations in crystals, it cannot be definitely concluded that amino acid mutations have no essential influence on the crystal structure.

It will be informative to examine what experimental factor is involved in the space group. The primary chemical agents used in protein crystallization of Hb were surveyed as shown in Table 7. The most dominant agent was selected from the description in the PDB file, where the crystal structures without description on crystallization agent were omitted from the data set. Polyethylene glycol is a major agent for crystallization of Hb. About 63% of the crystals are obtained from this chemical. Ammonium sulfate and ammonium phosphate account for 28% of the total. In groups 1 and 4, all of the protein crystals bearing the P 21 21 2 space group were obtained by using polyethylene glycol. Most of the protein crystals in groups 2 and 3 were grown by using one of the above three chemical agents. Various chemical agents were used for group 5. If polyethylene glycol is used for crystallization, crystal structures bearing the P 21 21 2 space group will be obtained in a high rate. The crystal structures with the P 1 21 1 space group will be generated in a higher rate than other structures if ammonium sulfate/phosphate is used for the agents of protein crystallization. Of course, the other experimental factor like buffer condition or protein concentration will influence on the crystallization, and a single factor is not enough to determine the space group of crystals. Crystal agent is, however, an obvious factor than others. Therefore, Table 7 suggests that chemical agents are related to the cluster groups and then correlate with the space group.

Table 7. Number of Structures Classified by the Primary Agent for Protein Crystallization of Hb

agent	group1	group2	group3	group4	group5	total
ammonium sulfate	0	12	0	0	0	12
polyethylene glycol[a]	51	4	10	11	16	92
ammonium sulfate/phosphate	0	21	4	0	3	28
others[b]	0	2	1	0	7	10
data not shown	1	1	1	0	1	4

[a]Polyethylene glycol is the sum of PEG, PEG1000, PEG1450, PEG1500, PEG3350, PEG4000, PEG6000, and PEG8000. [b]Others include ammonium phosphate and sodium/potassium phosphate.

The primary chemical agents used in protein crystallization of Mb are shown in Table 8. Ammonium sulfate is a major agent for crystallization of Mb and distributed from groups 1 to 6 except group 4. The crystal bearing the P 6 and P 1 21 1 space groups were grown by ammonium sulfate. Polyethylene glycol is a major agent in group 4 for the structures with the P 21 21 21 space group. Ammonium phosphate is also utilized in groups 3 and 5 for the structures with P 21 21 2. Therefore, Table 8 also indicates that agents for protein crystallization correlate with the space group. The primary chemical agents used in protein crystallization of HIV-1 PR are shown in Table 9. Ammonium sulfate and sodium chloride are the two major agents for crystallization of HIV-1 PR. About 89% of the crystals were obtained from these two chemicals. Most of the crystal structures of groups 2, 3, and 4 are obtained by using ammonium sulfate. A variety of agents were used in group 6. In groups 1 and 5, most of the protein crystallizations were obtained by using sodium chloride. That is, a large number of crystal structures bearing the P 21 21 2 and P 41 space groups were obtained from sodium chloride and most of crystal structures bearing P 21 21 21 and P 61 were obtained from ammonium sulfate. Therefore, the space group is sensitive to the crystal agent. Table 9 again suggests that the crystallization agent influences on the space group. As for HSA and HEWL, a single chemical agent is used for protein crystallization in almost all structures as shown in Tables S5 and S6. Hence, no definitive suggestion can be deduced from these two.

In our previous study for HIV-1 PR,[16] in addition to the space group and chemical agents for crystallization, the pH condition, resolution of X-ray diffraction, temperature factor, and inhibitor molecules were surveyed for every cluster groups to find the influence on the separation of crystal structures. No obvious factor was identified except for the space group and the chemical agent. In the present study, the resolution of X-ray diffraction was also surveyed for Hb, Mb, HEWL, HSA, and HIV-1 PR as shown in Tables S17–S21. Although the resolutions are different among the cluster groups, no definitive relationship was observed between the resolution and the separation of cluster groups.

Matthews coefficient V_M is the crystal volume per unit of protein molecular weight, which indicates the fractional volume of solvent in crystal.[26] The distribution of V_M was reported to range from 2 to 3 Å^3/Dalton based on the analysis with 15,641 crystallographic PDB data.[27] The relationship between the cluster groups and Matthews coefficient V_M in protein crystallization of Hb was surveyed as shown in Table S12 and Figure S5. Table S12 shows that the average Matthews coefficients ranges from 2.23 to 2.55 and the standard deviation is less than 0.26. The cluster groups 1, 2, and 4, in which most of the crystal structures have the same one space group, show a low standard deviation. Because group 5 is the gathering of the crystal structures not belonging to the other groups and group 3 consists of the crystal structures bearing more than four kinds of space groups, large standard deviations are observed in Matthews coefficient. Since the most crystal structures in group

Table 8. Number of Structures Classified by the Primary Agent for Protein Crystallization of Mb

agent	group1	group2	group3	group4	group5	group6	total
ammonium sulfate	118	47	44	0	2	20	231
polyethylene glycol[a]	5	0	0	10	0	2	17
ammonium phosphate	0	0	2	0	5	0	7
sodium/potassium phosphate	0	0	0	3	0	1	4
data not shown	6	10	0	6	0	3	25

[a]Polyethylene glycol is the sum of PEG1000, PEG1550, PEG3550, PEG4000, PEG8000, and PEG10000.

DOI: 10.1021/acs.jcim.5b00052
J. Chem. Inf. Model. 2015, 55, 1673–1685

Table 9. Number of Structures Classified by the Primary Agent for Protein Crystallization of HIV-1 PR

agent	group1	group2	group3	group4	group5	group6	total
ammonium sulfate	14	65	20	41	0	40	180
polyethylene glycol[a]	2	0	4	0	0	5	11
sodium chloride	112	1	5	1	17	26	162
others[b]	15	0	1	0	0	14	30
data not shown	28	11	33	2	0	31	105

[a]Polyethylene glycol is the sum of PEG8000, PEG4000, and PEG3350. [b]Others include potassium chloride, sodium potassium tartrate, potassium triocyanate, sodium iodide, potassium iodide, and sodium bromide.

Table 10. Number of Crystal Structures Classified by Protein Source for Mb

species	group1	group2	group3	group4	group5	group6	total
sperm whale	129	57	0	19	0	16	221
horse	0	0	46	0	6	5	57
pig	0	0	0	0	1	1	2
asiatic elephant	0	0	0	0	0	1	1
loggerhead turtle	0	0	0	0	0	2	2
harbor seal	0	0	0	0	0	1	1
total	129	57	46	19	7	26	284

1 and group 4 have the same common space group P 21 21 2, the averages of Matthews coefficient are close to each other. The average Matthews coefficient of group 2 is obviously different from other groups. The solvent content are also summarized in Table S12. The average solvent content ranges from 0.44 to 0.52 and the standard deviation is less than 0.05. Most of the crystal structures in group 1 and group 4 have the same common space group. Hence, the average solvent content is almost identical to each other, and the deviation is small. Similar to the standard deviation of Matthews coefficient, the standard deviation of solvent content for group 3 is the largest among all the groups. Since Matthews coefficient is related to the space group, it is natural that the coefficient shows the difference among cluster groups.

In the case of Mb, the average of Matthews coefficient ranges from 1.73 to 3.06, and the standard deviation ranges from 0.04 to 0.40 (Table S13 and Figure S6). Group 5 contains the crystal structures with two kinds of space groups (P 1 21 1 and P 21 21 2), and the standard deviation is large (±0.40). Group 5 contains five crystal structures with the P 21 21 2 space group, and the average Matthews coefficient only for the five structures is 3.04 ± 0.04. Group 1 contains 123 crystal structures with the P 6 space group, and the average Matthews coefficient only for these 123 crystal structures is 3.11 ± 0.14. These results mean that single space group, P 1 21 1, shows a low standard deviation. Most of crystal structures in groups 2 and 3 have the same space group, and the average Matthews coefficient is close to each other. The results of the solvent content are shown in Table S13. The averaged solvent content ranges from 0.31 to 0.60. The solvent contents of the crystal structures in groups 2 and 3 are also closed to each other. Since the crystal structures in group 2 have only one kind of space group, the standard deviation is very small.

The results of Matthews coefficient for HEWL are shown in Table S14 and Figure S7. In group 4, 10 crystal structures were obtained from the powder diffraction method, and Matthews coefficient is not applied for these 10 structures. Hence, there is only one crystal structure bearing the P 61 2 2 space group in group 4, and the standard deviation is 0.0. In group 2, crystal structures contains many different space groups. Then the standard deviation is larger than other groups. The Matthews coefficient of group 3 is different from those of other groups because most of the crystal structures have the P 1 space group.

In the case of HSA, the standard deviation of Matthews coefficient of group 3 is the largest among the three cluster groups (Table S15 and Figure S8). This is because group 3 is composed of the mixture of crystal structures with different space groups. In the case of HIV-1 PR, the average of Matthews coefficient ranges from 2.09 to 2.67 (Table S16 and Figure S9). The standard deviation of group 6 is the largest among all the cluster groups because group 6 is the mixture of crystal structures with different space groups. Since most of the crystal structures in groups 2 and 4 have the same common space group, P 21 21 21, their Matthews coefficients are close to each other. Since group 2 contains the structures bearing other space groups, the standard deviation of group 2 is larger than that of group 4. For the same reason, the standard deviation of the Matthews coefficient of group 3 is slight higher than those of groups 2 and 4. It is found from the above analysis that there is a relationship between space group and Matthews coefficient. This relationship is reflected to the separation of cluster groups.

The species of the protein source of Mb were surveyed as shown in Table 10. The number of crystal structures from the sperm whale accounts for 78% of all crystal structures. Those from the horse and other species account for 20% and 2%, respectively. All of the crystal structures of groups 1, 2, and 4 are from the sperm whale. All of the crystal structures of groups 3 and 5 are from the horse except for one crystal structure from the pig. Because group 6 is a mixed group, there are a variety of sources. Because of the large majority of crystal structures from the sperm whale, only Mb proteins of the sperm whale were examined again by cluster analysis (Figure 9). The space group is responsible for the classification of crystal structures as shown in Table 11 except for group iv, which is composed of the crystal structures not assigned to groups i-iii. In Table 11, most of crystal structures bearing a P 6 space group belong to group i, which is the same as group 1 in Table 2. There are 57 crystal structures bearing a P 1 21 1 space group in group ii, which is identical to group 2 in Table 2. Furthermore, 19 crystal structures bearing a P 21 21 21 space group are in group iii, which is identical to group 4 in Table 2. An important finding is that the crystal structures bearing the same space group P 1 21

DOI: 10.1021/acs.jcim.5b00052
J. Chem. Inf. Model. 2015, 55, 1673−1685

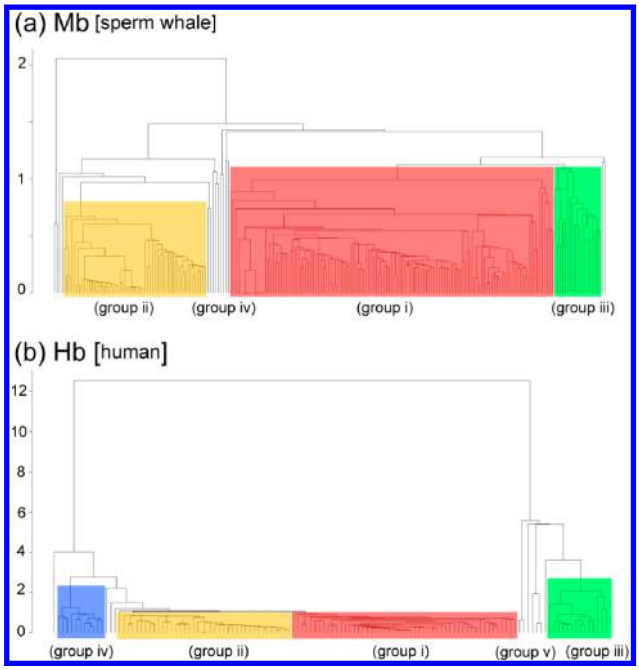


Figure 9. (a) Dendrogram for cluster analysis of crystal structures of Mb proteins from the sperm whale. A total of 221 structures were grouped into four clusters. (b) Dendrogram for cluster analysis of crystal structures of human Hb. A total of 128 structures were grouped into five clusters.

Table 11. Number of Crystal Structures Classified by Space Group for Mb from the Sperm Whale

space group	group i	group ii	group iii	group iv	total
P 21 21 21	6(0.05)[a]	0(0.00)	19(1.00)	4(0.25)	29
P 1 21 1	0(0.00)	57(1.00)	0(0.00)	9(0.56)	66
P 6	123(0.95)	0(0.00)	0(0.00)	1(0.06)	124
P 41	0(0.00)	0(0.00)	0(0.00)	1(0.06)	1
P 61 2 2	0(0.00)	0(0.00)	0(0.00)	1(0.06)	1
total	129	57	19	16	221

[a]The value in parentheses represents the ratio relative to the total number of crystal structures in each group.

1 are separated into two groups in Table 2, while one of the groups disappears in Table 11. One group is from the sperm whale, while the other is from the horse. These results suggest that the protein source is a critical factor to influencing cluster separation. Rashin et al. obtained a similar result with a large set of 291 Mb structures by principal component analysis (PCA).[13] As shown in Table 11, cluster analysis only with Mb from the sperm whale resulted in the same classification with all of the crystal structures. Therefore, the species of protein source is the primary factor to distinguish the crystal structures, and the space group of the crystals is the second factor. This dominancy of the protein source is natural because the amino sequence is different among species.

The species of the protein source of the crystal structures for Hb were surveyed as shown in Table 12. The crystal structures with the largest number are from humans. The structures of this source account for 88% and are distributed in all of the groups. The structures from other sources account for 12% and are only distributed in groups 3 and 5. Group 5 contains structures from other sources. In order to eliminate the difference in protein sources, only human Hb was examined. As

Table 12. Number of Crystal Structures Classified by Protein Source for Hb

species	group1	group2	group3	group4	group5	total
human	52	40	14	11	11	128
rabbit	0	0	2	0	0	2
goose	0	0	0	0	1	1
dromedary	0	0	0	0	1	1
mouse	0	0	0	0	1	1
meleagris gallopavo	0	0	0	0	2	2
pig	0	0	0	0	2	2
maned wolf	0	0	0	0	1	1
mammoth	0	0	0	0	1	1
dog	0	0	0	0	2	2
japanese quail	0	0	0	0	1	1
ostrich	0	0	0	0	1	1
horse	0	0	0	0	3	3
total	52	40	16	11	27	146

shown in Table 13, cluster analysis only with Hb from humans resulted in almost the same classification with all of the crystal structures in Table 1. The major space group for Hb is P 21 21 2, and the second major one is P 1 21 1. In spite of the diverse examination of crystallization condition or crystal characteristic, the reason why the different cluster groups in 1 and 4 have same space group P 21 21 2 is not clear.

For HEWL, 13 crystal structures were obtained by the powder diffraction method (1JA2, 1JA4, 1JA6, 1JA7, 1SF4, 1SF6, 1SF7, 1SFB, 1SFG, 2A6U, 2HS7, 2HS9, and 2HSO). Of the 13 structures, 10 crystals bear a P 43 21 2 space group, and all of them are classified into group 4. The other three crystal structures bear a P 43 21 2 space group, and they are in the mixed group 5. These results indicates a different diffraction method leads to a difference in the protein structures solved.

In Table 4, the crystal structures of HSA are classified into three groups. The crystal structure bearing a C 1 2 1 space group includes one molecule in a unit cell, and most of them are classified into group 1. The crystal structure bearing a P 1 space group includes two molecules in a unit cell. In cluster analysis, only the coordinates of chain A were extracted. These two types of space group are clearly separated except for one crystal structure. The reason for the separation is that the interaction of two molecules results in a difference in the protein conformation.

In Hb and Mb, the most important part is the heme-binding site. The site is inside of the enzymes and the segments showing large deviations in Figures 2b and 3b were apart from the heme-binding site. HEWL is a small protein, and the active site is exposed to solvent. The active site of HEWL is shown in Figures S10, comparing the structural deviations between groups 1 and 2 of Figure 4b. In Figure S10c, the structural difference is observed outside the active site. The structural deviations in Figure 4b (4↔1), however, contain the residues at the active site. HSA has a function of reservoir of other proteins. Large structural deviations were observed at the outer parts of IA, IIB, and IIIB in Figure 5b. Those areas are not responsible for the reservoir function. In HIV-1 PR, the active site is at the center of the protein. Only the flap region, which showed large structural deviation in Figure 6b (5↔1), is related to the enzymatic activity. Therefore, in the case of Hb or HSA, that has a large molecular weight, the active site hardly coincides with the region in which large structural deviations were observed in comparison among cluster groups. In the case

DOI: 10.1021/acs.jcim.5b00052
J. Chem. Inf. Model. 2015, 55, 1673−1685

Table 13. Number of Crystal Structures Classified by Space Group for Human Hb

space group	group i	group ii	group iii	group iv	group v	total
P 21 21 2	50(0.96)[b]	0(0.00)	0(0.00)	11(1.00)	2(0.18)	63
P 1 21 1	2(0.04)	36(0.90)	1(0.07)	0(0.00)	2(0.18)	41
P 21 21 21	0(0.00)	4(0.10)	7(0.50)	0(0.00)	3(0.27)	14
P 32 2 1	0(0.00)	0(0.00)	6(0.43)	0(0.00)	0(0.00)	6
others[a]	0(0.00)	0(0.00)	0(0.00)	0(0.00)	4(0.36)	4
total	52	40	14	11	11	128

[a]Others is the sum of those for C 1 2 1, P 41 21 2, P 1, and P 61 2 2 space groups. [b]The value in parentheses represents the ratio relative to the total number of crystal structures in each group.

of HEWL or HIV-1 PR, which has a relatively small molecular weight, the structural deviations are sometime observed near the active site.

5. CONCLUSION

Cluster analysis was applied to the crystal structures of Hb, Mb, HEWL, HSA, and HIV-1 PR, which have abundant entries in the crystallographic data in the PDB site. The downloaded data were filtered in view of missing residues and neat conformation in the secondary structure to prepare a data set that is appropriate for cluster analysis. The structures of Hb (n = 146) were classified into five groups by the nearest neighboring method. The structures of Mb (n = 284) were classified into six groups. Those of HEWL (n = 336) were clustered into five groups. Those of HSA (n = 63) were clustered into three groups. Those of HIV-1 PR (n = 488) were classified into six groups. Information on amino acid mutation, space group of protein crystals, species of protein source, and crystallization condition were surveyed for every member of the respective clusters. A major factor to distinguish the cluster groups is the space group of crystals. Chemical agents used in protein crystallization have a critical influence on the structural difference for all of the proteins. From the examination of Hb and Mb, the species of protein source was found to be a more crucial factor for classification. From the examination of HEWL, the difference in diffraction method was found to be another crucial factor. From the examination of HIV-1 PR, it was found that mutations for drug resistance had little influence on the separation of whole-body crystal structures.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jcim.5b00052.

Member lists for the respective groups in the cluster analysis. Surveys on the amino mutations introduced in proteins and the chemical agents used for crystallization. (PDF)

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: hoshino@chiba-u.jp.

Author Contributions

All authors have given approval to the final version of the manuscript.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was supported by a grant for Scientific Research C from the Japan Society for the Promotion of Science.

■ REFERENCES

(1) Hoffman, I. D. Protein Crystallization for Structure-Based Drug Design. *Methods Mol. Biol.* **2012**, *841*, 67−91.

(2) Deschamps, J. R. The Role of Crystallography in Drug Design. *AAPS J.* **2005**, *7*, E813−E819.

(3) Russo Krauss, I.; Merlino, A.; Vergara, A.; Sica, F. An Overview of Biological Macromolecule Crystallization. *Int. J. Mol. Sci.* **2013**, *14*, 11643−11691.

(4) Warren, B. E. *X-ray Diffraction*; Courier Corporation; 1969; Chapter 8, pp 95−103.

(5) Ode, H.; Neya, S.; Hata, M.; Sugiura, W.; Hoshino, T. Computational Simulations of HIV-1 Proteases–Multi-Drug Resistance due to Nonactive Site Mutation L90M. *J. Am. Chem. Soc.* **2006**, *128*, 7887−7895.

(6) Ko, G. M.; Reddy, A. S.; Kumar, S.; Bailey, B. A.; Garg, R. Computational Analysis of HIV-1 Protease Protein Binding Pockets. *J. Chem. Inf. Model.* **2010**, *50*, 1759−1771.

(7) Agniswamy, J.; Shen, C. H.; Aniana, A.; Sayer, J. M.; Louis, J. M.; Weber, I. T. HIV-1 Protease with 20 Mutations Exhibits Extreme Resistance to Clinical Inhibitors through Coordinated Structural Rearrangements. *Biochemistry* **2012**, *51*, 2819−2828.

(8) Milburn, M. V.; Tong, L.; deVos, A. M.; Brünger, A.; Yamaizumi, Z.; Nishimura, S.; Kim, S. H. Molecular Switch for Signal Transduction: Structural Differences Between Active and Inactive Forms of Protooncogenic Ras Proteins. *Science* **1990**, *247*, 939−945.

(9) Kahraman, A.; Morris, R. J.; Laskowski, R. A.; Thornton, J. M. Shape Variation in Protein Binding Pockets and Their Ligands. *J. Mol. Biol.* **2007**, *368*, 283−301.

(10) Castelli, M.; Clementi, N.; Sautto, G. A.; Pfaff, J.; Kahle, K. M.; Barnes, T.; Doranz, B. J.; Dal Peraro, M.; Clementi, M.; Burioni, R.; Mancini, N. HCV E2 Core Structures and mAbs: Something is Still Missing. *Drug Discovery Today* **2014**, *19*, 1964−1970.

(11) Nanni, L.; Brahnam, S.; Lumini, A. Prediction of Protein Structure Classes by Incorporating Different Protein Descriptors into General Chou's Pseudo Amino Acid Composition. *J. Theor. Biol.* **2014**, *360*, 109−116.

(12) Rashin, A. A.; Rashin, A. H.; Jernigan, R. L. Protein Flexibility: Coordinate Uncertainties and Interpretation of Structural Differences. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2009**, *65*, 1140−1161.

(13) Rashin, A. A.; Domagalski, M. J.; Zimmermann, M. T.; Minor, W.; Chruszcz, M.; Jernigan, R. L. Factors Correlating with Significant Differences Between X-ray Structures of Myoglobin. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2014**, *70*, 481−491.

(14) Hestenes, D.; Holt, J. W. Crystallographic Space Groups in Geometric Algebra. *J. Math. Phys.* **2007**, *48*, 1−25.

(15) Ode, H.; Ota, M.; Neya, S.; Hata, M.; Sugiura, W.; Hoshino, T. Resistant Mechanism against Nelfinavir of Human Immunodeficiency Virus Type 1 Proteases. *J. Phys. Chem. B* **2005**, *109*, 565−574.

(16) Qi, F.; Fudo, S.; Neya, S.; Hoshino, T. A Cluster Analysis on the Structural Diversity of Protein Crystals, Exemplified by Human

DOI: 10.1021/acs.jcim.5b00052
J. Chem. Inf. Model. 2015, 55, 1673−1685

Immunodeficiency Virus Type 1 Protease. *Chem. Pharm. Bull.* **2014**, *62*, 568−577.
(17) Chambers, J. *Software for Data Analysis: Programming with R*; Springer Science & Business Media: 2008; Chapter 4, pp 79−110.
(18) Gentleman, R. *R Programming for Bioinformatics*; CRC. Press: 2008; Chapter 2, pp 5−66.
(19) Katz, D. S.; White, S. P.; Huang, W.; Kumar, R.; Christianson, D. W. Structure Determination of Aquomet Porcine Hemoglobin at 2.8 A Resolution. *J. Mol. Biol.* **1994**, *244*, 541−553.
(20) Lukin, J. A.; Kontaxis, G.; Simplaceanu, V.; Yuan, Y.; Bax, A.; Ho, C. Quaternary Structure of Hemoglobin in Solution. *Proc. Natl. Acad. Sci. U. S. A.* **2003**, *100*, 517−520.
(21) Kendrew, J. C.; Bodo, G.; Dintzis, H. M.; Parrish, R. G.; Wyckoff, H.; Phillips, D. C. A Three-Dimensional Model of the Myoglobin Molecule Obtained by X-ray Analysis. *Nature* **1958**, *181*, 662−666.
(22) Kendrew, J. C.; Dickerson, R. E.; Strandberg, B. E.; Hart, R. G.; Davies, D. R.; Phillips, D. C.; Shore, V. C. Structure of Myoglobin: A Three-Dimensional Fourier Synthesis at 2 A. Resolution. *Nature* **1960**, *185*, 422−427.
(23) Williams, M. A.; Thornton, J. M.; Goodfellow, J. M. Modelling Protein Unfolding: Hen Egg-White Lysozyme. *Protein Eng., Des. Sel.* **1997**, *10*, 895−903.
(24) Matsuo, K.; Watanabe, H.; Tate, S.; Tachibana, H.; Gekko, K. Comprehensive Secondary-Structure Analysis of Disulfide Variants of Lysozyme by Synchrotron-Radiation Vacuum-Ultraviolet Circular Dichroism. *Proteins: Struct., Funct., Genet.* **2009**, *77*, 191−201.
(25) Sugio, S.; Kashima, A.; Mochizuki, S.; Noda, M.; Kobayashi, K. Crystal Structure of Human Serum Albumin at 2.5 A Resolution. *Protein Eng., Des. Sel.* **1999**, *12*, 439−446.
(26) Matthews, B. W. Solvent Content of Protein Crystals. *J. Mol. Biol.* **1968**, *33*, 491−497.
(27) Kantardjieff, K. A.; Rupp, B. Matthews Coefficient Probabilities: Improved Estimates for Unit Cell Contents of Proteins, DNA, and Protein-Nucleic Acid Complex Crystals. *Protein Sci.* **2003**, *12*, 1865−1871.

DOI: 10.1021/acs.jcim.5b00052
J. Chem. Inf. Model. 2015, 55, 1673−1685